Big Data Collection and Analysis Framework Research for Public Digital Culture Sharing Service

Guigang Zhang, Jian Wang, Weixin Huang, Haixia Su, Zhi Lv, Qi Yao, Shufeng Ye
Institute of Automation, Chinese Academy of Sciences
Beijing, China
guigang.zhang@ia.ac.cn

Abstract*—the big data collection and analysis of public digital culture sharing service is researched in this paper. Big data includes three types of data, namely: ancillary service data, public digital culture sharing service platform operation data and user data. The aim is to build a data analysis platform for the three classes of data. Through the analysis of the three types of data collected, the use of resources and the operation of the platform can be mastered for providing better service for resource organization and scheduling of platform. Through the analysis of three types of the collected data, it can realize all kinds of statistics and analysis services in multidimensional. This paper presents a personalized recommender system of public digital cultural resources.

Keywords-Public Culture; big data; Cloud Computing; Framework.

I. INTRODUCTION

In this paper, we need to build a national public digital culture sharing service platform. In order to complete the study, mainly includes the following requirements.

(1)Data collection of the public digital culture sharing service

Demand for data collection mainly includes the following three aspects. They are:1)Ancillary services data collection.2)National public digital culture sharing service platform operation data collection. 3)the user data acquisition

(2)Building a Big Data analysis platform for the sharing service data

The platform can implement the analysis of the three kinds of data aforementioned. Therefore, building a big data analysis based on the public digital culture sharing service platform is particularly important. Through analysis, we can help the platform to achieve resource organization, deployment, scheduling and optimization, etc.

(3)Data analysis and research of the public digital culture sharing service

Through the analysis of the data of the public digital culture sharing service, it can grasp the public digital culture resources service condition, national public digital culture sharing service platform operation and further understand user behavior. It provides decision support for improving the platform resources organization and scheduling, platform operation efficiency, and better improves the quality of service for the users. Therefore, the implementation of data analysis and research of the public digital culture sharing service is an important requirement.

(4)Personalized recommender system research and development of public digital culture resources

Public digital culture sharing service platform has amounts of user behavior data.

II. RELATED WORK

A. Big Data

MapReduce programming framework for public cultural data processing include: Google MapReduce[1], Hadoop MapReduce, Pregel[2], Dremel[3], Hadoop++[4], CoHadoop[5], Haloop[6], Twister[7], Microsoft Dryad[8],Spark and HadoopDB. HadoopDB is a kind of data processing framework which is between MapReduce and PDBMS, and it takes into account the MapReduce capacity of processing massive data and the computing power that its relational database carried on.

B. Recommendation Technologies

Recommendation systems[9-16] are the most important application form of personalized services, which is defined as when the people provide the resources; it can aggregate and assign them to the appropriate recipient system. But the term recommendation system now has a broader connotation: Describing these systems that treat personalized recommendation as outputs, or using a personalized manner to guide the users choosing interesting or useful entries in a great optional space. These systems have strong ability in this environment where network information is far more than any individuals' research capability. Recommender system often consists of three parts. Behavior recording module is responsible for recording the behavior that can reflect user preferences, such as purchase, download, rating and so on. The function of model analysis module is to achieve an analysis of user behavior record. With different algorithms, it establishes model that describes the user's preference information. Finally, the recommended module, recommends the content that might be interested to the users by filtering from the target users.

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.37

IEEE computer society

III. BIG DATA PLATFORM FRAMEWORK OF PUBLIC DIGITAL CULTURAL SHARING SERVICES

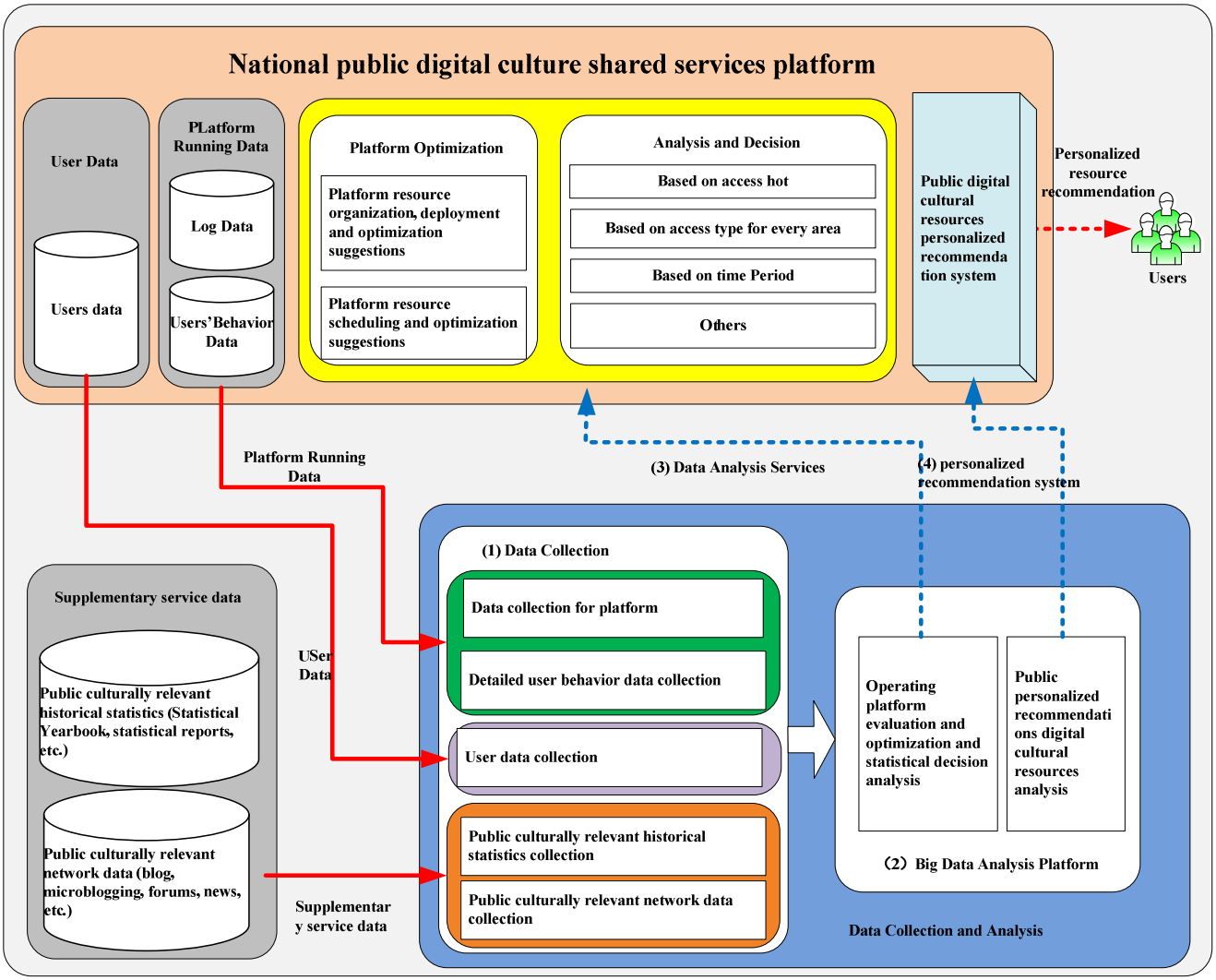

Fig.1. The big data platform framework of public digital cultural sharing services

Firstly, data of ancillary service, national public digital culture sharing service platform operation data and users' data is acquired. These acquired data will be stored into a well-built big data analysis platform, which is used to store, analyze, and handle data. Then the big data analysis platform will estimate and optimize the sharing data, and make application analysis and decision. In the end, a recommender system is researched and developed to accomplish characterized recommendation based on users' interest by analyze their behavior.

IV. THE DATA ACQUISITION FOR PUBLIC DIGITAL CULTURAL SHARED SERVICES.

Figure 2 shows the data acquisition technology roadmap for public digital cultural shared services. The part of the research is mainly to realize the collection of three types of data as shown below: The acquisition of auxiliary service data, collection of user's data with offline experiences and collection of running data in public sharing service platform of the nation's digital culture platform.(1)Auxiliary service data acquisition concretely including public cultural collection related to historical statistical data and the data collection of all kinds of websites related to public culture. (2) User data acquisition. Application system mainly for the user's offline experience , such as brush painting experience with various data collected in the user device, especially the behavioral data. After collecting the user's data, the data will be stored in the database, for later analysis.(3) Running data collection on the platform. It mainly gathering public digital culture running data sharing service on the nation's platform.

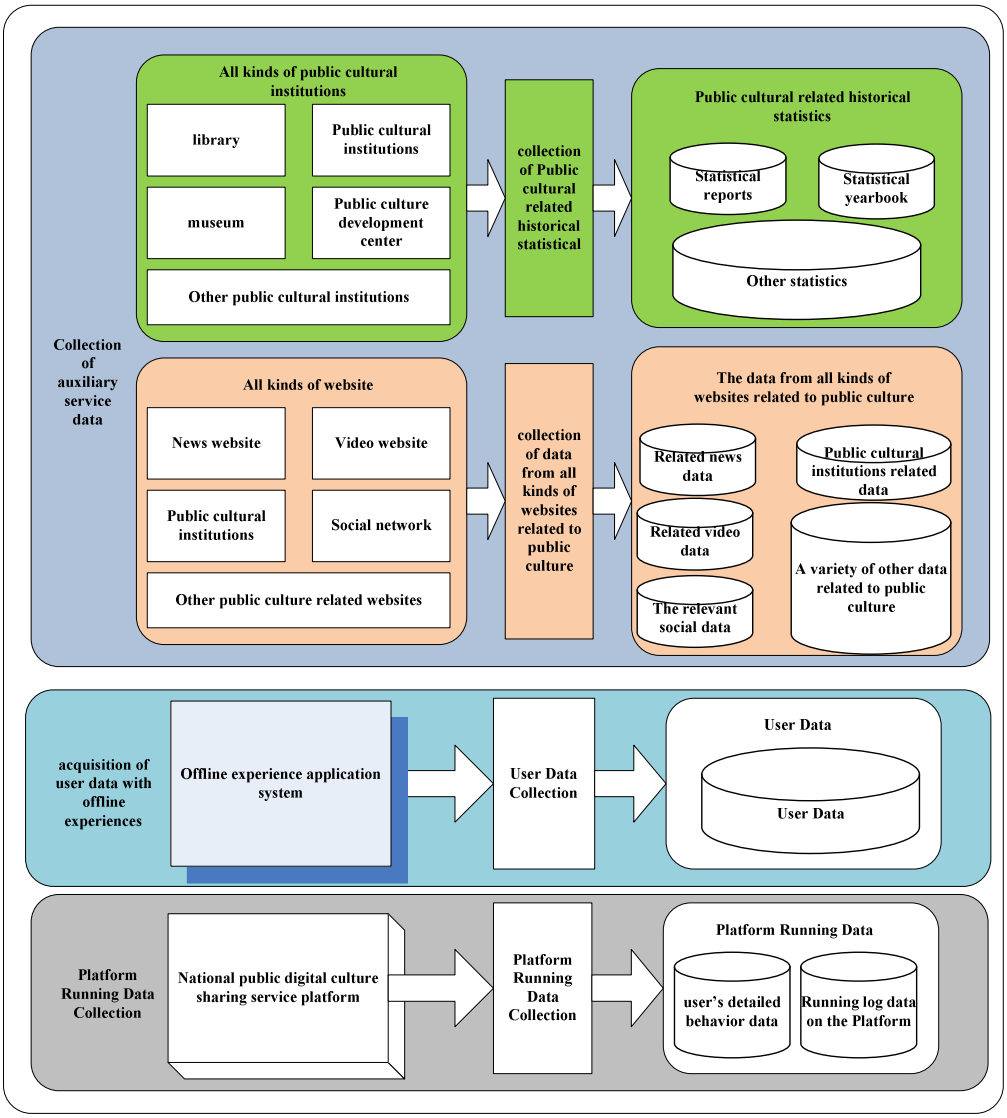

Figure 2. Data collection and technical route of public digital culture sharing service

V. THE BIG DATA ANALYTICS PLATFORM OF SHARING SERVICE DATA

(1) The big data analytics platform hardware layer of Data sharing service. It includes the storage facilities, computing facilities and network facilities, etc. By Purchasing the hardware to build the platform and virtualization software for the entire hardware virtualization management, and finally form a complete cloud Infrastructure environment for research.

(2) The data storage layer of the big data analytics platform with Data sharing service. The main services include storage of structured data service(using MySQL), unstructured data service (using HDFS and HBase of Hadoop system) and index data (using MySQL) to store. Making the Hadoop system to the front step of the hardware layer, directly ,data stored by the platform software level.

(3) The data management layer of the big data analytics with Data sharing service. Data management platform major resource management of public figures and cultural Data sharing service , metadata management of public figures and cultural Data sharing service and index management of public figures and cultural Data sharing service.

VI. RESEARCH OF DATA ANALYSIS AND SERVICE OF PUBLIC DIGITAL CULTURE

(1) data layer

Obtain public digital culture shared services data for various analyses from the shared services platform for big data analysis.

(2) Data analysis layer .

Including analysis for platform evaluation and optimiza tion and a variety of applications based on data of public di gital culture shared services. They mainly include the corre sponding methods and strategies for analysis for platform e valuation and optimization and a variety of applications.

(3) Applied decision layer.

The applied decision layer corresponds to data analysis layer respectively.

VII. PUBLIC DIGITAL CULTURE RESOURCE PERSONALIZED RECOMMENDATION SYSTEM

It consists of three subsystems, the user clustering subsystem, the personalized recommendation subsystem and the recommendation feedback subsystem.

(1)User clustering subsystem

Generating user's behavior model by using clustering technology to cluster sequences that mainly extract from path sequence of public digital cultural resources where users frequently access to.

(2)Personalized recommendation subsystem

Getting personalized recommendations by using the users-public digital cultural resources collaborative filtering algorithm.

Users-public digital cultural resources collaborative filtering recommendation are the personalized recommendations that is based on the user's nearest neighbor information calculated from public digital cultural resources matrix .

(3)Recommending feedback subsystem

By dynamically adjusting the matching scores between the users and the public digital cultural resources, achieving the personalized public digital cultural resources recommendation system that is mostly based on users' behavior.

VIII. CONCLUSIONS

This paper discussed big data collection and analysis framework of public digital cultural services. Especially, we analyzed data collecting from ancillary service data, national public digital culture sharing service platform operation data and users' data. Big data analysis platform was established based on three types of data above. Through the analysis of the three types of data, we mastered how to use these resources and how the platform works, which could provide better services for resource organization and management of the platform. Moreover, with the three types of data, from the angle of multi-dimension, we realized various kinds of statistical analysis services. And a public digital cultural resource personalized recommender system was given. Future research in this study focused on the following points:

(1) Public digital cultural data resource scheduling optimization.

(2) Develop a public digital cultural resource personalized recommender system.

ACKNOWLEDGMENT

This research was supported by: (1) the Support Program of the National '12th Five-Year-Plan' of China under Grant No. 2015BAK25B04;(2) the Support Program of the National '12th Five-Year-Plan' of China under Grant No. 2015BAK25B03;(3) Special Project for Civil-Aircraft, MIIT and (4) 2014 annual central cultural industry development fund.

REFERENCES

[1] Map Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In: Brewer E, Chen P, eds. Proc. of the OSDI2004. California: USENIX Association, 2004. PP:137-150.

[2] Grzegorz Malewicz, Matthew H. Austern, Aart J.C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski.Pregel: A System for Large-Scale Graph Processing,Proceedings of the 2010 ACM SIGMOD International Conference on Management of data,Pages 135-146.

[3] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis. Dremel: Interactive Analysis of Web-Scale Datasets.Proc. of the 36th Int'l Conf on Very Large Data Bases (2010), pp. 330-339.

[4] Dittrich J, Quian'e-Ruiz JA, Jindal A, Kargin Y, Setty V, Schad J. Hadoop++: Making a yellow elephant run like a cheetah (without it even noticing). PVLDB2010, 2010,3(1-2):518−529.

[5] Mohamed Y. Eltabakh,Yuanyuan Tian,Fatma Özcan, Rainer Gemulla,Aljoscha Krettek,John McPherson. CoHadoop: flexible data placement and its exploitation in Hadoop.Proceedings of the VLDB Endowment VLDB2011 Endowment Hompage archive Volume 4 Issue 9, June 2011.Pages 575-585.

[6] Bu YY, Howe B, Balazinska M, Ernst MD. HaLoop: Efficient iterative data processing on large clusters. PVLDB2010, 2010,3(1-2): 285−296.

[7] Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, SeungHee Bae, Judy Qiu, Geoffrey Fox, Twister: A Runtime for Iterative MapReduce, The First International Workshop on MapReduce and its Applications (MAPREDUCE'10) .PP:110-119.

[8] Isard M, Budiu M, Yu Y, Birrell A, Fetterly D. Dryad: Distributed data-parallel programs from sequential building blocks. ACM SIGOPS Operating Systems Review, 2007,41(3),PP:59-72.

[9] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.IEEE TKDE, 17(6), pp. 734–749, 2005.

[10] Y. Ge, H. Xiong, A. Tuzhilin, etc. An energy-efficient mobile recommender system. In ACM SIGKDD'10, pp. 899–908, 2010.

[11] Y. Koren, R. Bell and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. In IEEE Computer, vol.42(8), pp. 30–37, 2009.

[12] G. D. Abowd, C. G. Atkeson, J. Hong, and et al. Cyber-guide: A mobile context-aware tour guide. Wireless Networks, 3(5), pp. 421–433, 1997.

[13] O. Averjanova, F. Ricci, and Q. N. Nguyen. Map-based interaction with a conversational mobile recommender system.In UBICOMM'08, pp. 212–218, 2008.

[14] K. Cheverst, N. Davies, and et al. Developing a context-aware electronic tourist guide: some issues and experiences. InACM SIGCHI, pp. 17–24, 2000.

[15] Dogan Gursoy, Ken W. McCleary.An Integrative Model of Tourists/ Information Search Behavior.Annals of Tourism Research, Volume 31, Issue 2, April 2004, Pages 353–373.

[16] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-Based Collaborative Filtering Recommendation Algorithms, In: Proceedings of the 10th international conference on World Wide Web, 2001, 285-295.