An Effective User Behavior Modeling Approach for Data Services in the Field of Materials Engineering

Xin Cheng
School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China
Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083
Email: chengxin0613@gmail.com

Changjun Hu, Yang Li, Wei Lin
School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China
Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083
Email: {hu.cj.mail, mailbox.liyang, xueshulinwei}@gmail.com

***Abstract*—New challenges about data services have arisen, especially consider the impact of user behavior. For dealing with the problem of distributed heterogeneous data sharing and satisfying the demands of data service, the complex association and dynamic changes should be tracked timely. Thereby it has become increasingly important to build a data services framework based on the user behavior analysis. In this paper, we propose an effective user behavior modeling approach based on the Open Cloud Service Architecture (OCSA) to manage the domain scientific data. Firstly, we build the data services framework in the cloud environment, and describe the related theories and conceptual model about this framework. Based on this, we elaborate the definition and classification of user behavior. Secondly, we construct the User Behavior Model (UBM) by tracking the user operation behavior respectively from the time dimension and the space dimension. Finally, we designed and realized a behavior-oriented Materials Scientific Data Sharing Service Platform. Through the description and comparison of application case, the validity of user behavior modeling method is verified.**

***Index Terms*—data services, user behavior modeling, cloud computing, materials engineering**

I. INTRODUCTION

Along with the rapid growth of massive data and the continuous development of diverse service demands, researchers are facing more new challenges about the data processing and the service diversification. The "big data" management issues [1][2] are particularly prominent in the aspect of domain scientific data service. Especially in the field of materials engineering, scientists and the engineering staffs find that it is difficult to manage, analysis [3], deal with and reuse the materials scientific data effectively. The scientific data of materials engineering are mainly constituted by the complicated experimental data and the process product data. Dealing with the large-scale scientific data sets could not catch the speed of data generation. The traditional database management mode and service framework cannot meet the growing needs of data services. Therefore, it is important and urgent to construct a new architecture of data services to support the large-scale scientific data sharing and applications.

The data services of materials engineering mainly faced with the general problems of big data management, such as the features of massive, distributed, heterogeneous, and so on. In addition, it also faced with the special problems for reusing and sharing the materials scientific data. These special issues could be described as the following three aspects.

Complex association: there are complicated relationships between different data, and the associate degrees are uncertain, which leads to the complex and uncertain data operation behavior [4]. For example, the hardness properties data of alloy steel are closely associated with the density properties data of iron element. Therefore, the operation behavior about the basic elements properties data would affect the access about alloy data.

Real-time dynamic changes: the data values, data types, data relationships are always dynamically changing. It demands for the high efficiency real-time access of data services. For example, with the changes about temperature conditions of steel materials, the tensile properties are also changed in different time.

Conditional correlation: the different time, locations, environments, and even social factors might affect the

Manuscript received January 23, 2013; revised April 27, 2013; accepted May 12, 2013.
Corresponding Author: Changjun Hu, hu.cj.mail@gmail.com

doi:10.4304/jcp.8.11.2972-2979

data, relationships, and user operation behavior. For example, the aerodynamics research about aerospace composites needs considering the influence factors of the gravitational acceleration parameter values which changed in different space locations (latitude, altitude).

It could be seen that the traditional modes of data services have been unable to deal with these new issues and challenges, thus we need to build a new data services architecture, which could meet the new data features and service demands. In this paper, we propose a new data services framework based on the cloud environment, and research the user behavior modeling approach to support the data services of materials engineering. Then it leads the domain scientists to manage the scientific domain "big data" efficiently, and to reuse the scientific data by deeply mining the data relationships. Thereby realize the humanized and individualized e-Science applications.

II. Research Status

Recently, the related research works about data services have become increasingly prominent in the field of information science research. For the different requirements of data services, researchers conduct the related research from different aspects. It mainly includes that studying the MapReduce technology of cloud computing for the high-performance task assignments and the efficient storage; constructing the cloud platforms for the distributed data processing; proposing the concept of dataspace [5] for the big data management; researching the ontology technology for the semantic representation of heterogeneous data sources; and so on. Although some of these researches are just started in recent years, but their prospects are very broad and valuable. Especially the dataspace technology has the considerable room to develop and in-depth study.

Around the related technology of cloud computing, researchers build some distributed platform frameworks in succession. R. L. Grossman et al propose an Open Science Data Cloud (OSDC) [6], which uses the Hadoop and MapReduce technology to build the cloud platform for supporting the analysis, processing and management of large-scale scientific data sets. L. Zhang et al introduce a Cloud Computing Open Architecture (CCOA) [7], which supports the service reuse and customize based on seven architectural principles. Thus they have realized the effective integration of Oriented Service Architecture (SOA) and the virtualization in the ecological environment of cloud computing. S. Loebman et al from University of Washington use Hadoop/Pig [8] to manage the large-scale scientific data sets which arising from the astrophysical simulation experiment. Thereby they have improved the RAM scalability, enhanced the I/O bandwidth services, and significantly verified the advantage in response speed respect of large-scale simulation data analysis queries. M. F. Husain et al propose a massive data semantic storage and retrieval architecture [9] based on the Hadoop technology to effectively improve the processing performance of storage and query. L. Youseff et al propose a cloud computing unified joint ontology construction method [10] to support the semantic integration by taking the ontology technology into the cloud computing applications. Thus they have achieved the efficiently retrieval service based on the ontology technology.

For trying to seek a new technology to deal with the new challenges of big data management, M. Franklin et al proposed the concept of dataspace [5] in 2005. Followed by this, researchers have designed a wide variety of dataspace architectures and models, and have realized several systems which meet the individual application needs. Such as iMeMex[11], Semex[12], PAYGO[13], UDI[14] etc. These dataspace prototype systems could satisfy the demands of data services in a certain extent, such as the needs of semantic integration, on-demand services, pay-as-you-go, and so on. Despite this, most technical programs are limited in the coarse-grained architecture research. These prototype systems rarely consider the dynamic evolution of data model, especially considering the analysis and modeling research about user behavior. Based on this, we have proposed the concept of Virtual DataSpace (VDS) [15]. Its technology method could make the physical data into virtualization processing, and then through the data association modeling to realize the efficient query service. In order to further develop this theory and model system, we should in-depth study the dynamic evolution model of VDS, especially combined with the research of user behavior modeling.

Along with the significant influence on data management caused by user behavior, the research about data operation behavior has become increasingly important. X. Hu et al propose an optimized ant colony clustering algorithm (OACA) [16] in dynamic pattern discovery. They have explored the structured formula to describe the users' browsing behavior patterns, and analyzed the adaptive features of them. C. Xu et al propose a novel user click behavior identification method [17] based on the hidden semi-markov model. And then verified the effectiveness of this method through applied to an educational website. F. Benevenuto et al present the first workloads analysis prototype [18] in online social networks based on the analysis of detailed clickstream data. It has revealed the key features of the social network workloads. These above methods have some certain referential significance for our research, but they have not constructed a complete set of behavioral analysis model. Meanwhile, most of them lack of the idea of dynamic modeling, they have no particular emphasis on the importance about the operating time, thus they did not depth discuss the user behavior model in the time dimension.

In summary, considering the effect of user behavior and the support of dynamic distributed cloud architecture environment, we propose a new cloud data services framework based on the user behavior modeling in VDS. Through describing the user behavior modeling approach, we could track the data operation process, obtain the trend of dynamic changes about data and its relationships, and then effectively achieve the individualized and real-time data services in the field of materials engineering.

III. DATA SERVICES FRAMEWORK BASED ON THE CLOUD ENVIRONMENT

A. Open Cloud Service Architecture (OCSA)

I. Elsayed et al proposed a dataspace management system architecture [19] which combined the dataspace concept with the grid technology. As the grid protocol architecture, Open Grid Services Architecture (OGSA) is proposed by the basis of traditional grid "five layers hourglass structure" and Web Service technology. It could effectively support the heterogeneous distributed data management based on the theory analysis and the case description about the dataspace technology within OGSA. However, OGSA is unable to solve the "big data" problems which have the feature of dynamic real-time demand. Based on this, we propose an Open Cloud Services Architecture (OCSA) to realize the efficient data services based on the cloud environment [20]. OCSA could be defined as follows.

Definition 1: Open Cloud Services Architecture is a six-tuples, OCSA= (DR, CSE, VDS, RS, BS, AIS). DR is the abbreviation of "Data Resources", which denotes the underlying physical data resource sets in OCSA. CSE is the abbreviation of "Cloud computing Supporting Environment", which denotes the virtualization support environment that built by the related technology of Hadoop. VDS is the abbreviation of "Virtual DataSpace", which is mainly responsible for the data management and on-demand services by using the semantic mapping and dynamic evolution mechanism. RS is the abbreviation of "Requirement Space", which supports the user demand modeling by combined with the workflow technology to support the dynamic combination and management of services. BS is the abbreviation of "Behavior Space", which optimizes the data services by the behavior analysis and modeling in order to support the proactive and user-friendly applications. AIS is the abbreviation of "Application Instance Sets", which includes the personalized and customized service instances that provided by the OCSA for various types of users. The abstract framework of OCSA is shown in Fig.1.

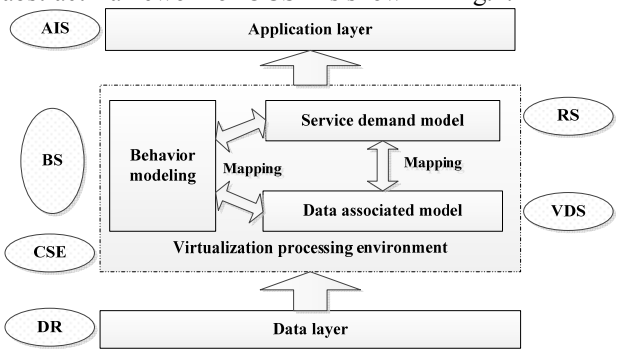

Figure 1. The abstract framework of OCSA.

From this framework, we can see that OCSA is a data services framework based on the cloud environment. The differences of the data services framework in the cloud environment from that in the grid environment and centralized architecture are described in Table Ⅰ.

TABLE Ⅰ.
THE CHARACTERISTICS COMPARISON OF DATA SERVICES FRAMEWORK

Angle	Centralized architecture	OGSA	OCSA
Core	Function	Service	Data and service
Processing	Centralized	Parallel	Virtualization
Reuse	Module	Service	Data and service
Storage	Remote server	Some locals	Everywhere
Allocation	By function	By task	On demand
Expansibility	Manual	Semi-automatic	Automatic
Data feature	Stability	Migratory	Associated evolution

From the comparison, we can see that OCSA have the significant features and advantages in four aspects: it emphasizes the data virtualization; data resources are allocated on demand; both the data and service could be reuse; it supports the associated evolution and the automatic expansion.

Therefore, OCSA could effectively get through the passageway between the underlying data resources and the upper layer application services. Based on the supporting of virtualization processing environment, we could find the association mapping among the data relationship model in VDS, the service demand model in RS, and the user behavior model in BS. Thereby, we could achieve the optimized and individualized data services based on the supporting of user behavior modeling.

B. Key Concepts and Theories about OCSA

From the abstract framework of OCSA, we could find that the related theoretical and technical issues are mainly focused on the following four areas.

(1) Data acquisition based on the cloud environment

Definition 2: The underlying physical Data Resources (DR) of OCSA is defined as a tuple, DR= (NDR, WDR), where NDS is the Node Data Resources, WDS is the Web Data Resources.

We need adopt different technical schemes to obtain the data from different sources. For the Node Data Resources (NDR), considering the huge interconnected barriers and data interoperability difficulty among distributed nodes, we adopt Web Service technology to build the interoperable interfaces for realizing the cross-domain access and sharing of node data. For the Web Data Resources (WDR), we use the Web Spider technology to intelligently crawl the relevant network information which scattered in the website, bbs, micro-blog, etc. It has the same parameter type "Link URL" between NDR and WDR; therefore, we could obtain the distributed massive data by combining them to the same data capture mechanism in order to solve the big data distribution problem.

In the meantime, we need adopt the Hadoop virtualization processing technology to build the Cloud computing Supporting Environment (CSE). The core technologies of Hadoop are HDFS and MapReduce. In which HDFS is the Hadoop Distributed File System which supports the large file operation for realizing the large-scale computing, storage and access. MapReduce is a distributed parallel programming model which provides the Map function to decompose tasks and uses the

Reduce function to aggregate the distributed processing results into the end results. The CSE based on the Hadoop technology could support the high performance computing and efficient store access for the big data management in OCSA. Thereby, it could effectively satisfy the virtualized data processing requirements of "physical distribution, logical unification".

(2) Conceptual model of Virtual DataSpace (VDS)

Virtual DataSpace (VDS) is the sets of data, services and its relationships which are related with the subjects and based on the supporting of virtualization. Compared with the traditional data management mode, VDS has the more significant features and advantages, such as the "data first" mode, more emphasizes the associated mapping and dynamic evolution, and more highlights the importance of service, etc. From the macro level, the whole VDS could be understood as the sets of all the data and their relationships. VDS could be defined as follow.

Definition 3: Virtual DataSpace is a tuple, VDS= (DS, DRS), where DS is the Data Sets, and RS is the Data Relationship Sets.

Virtual DataSpace is equivalent to the sum of all the sub virtual dataspaces; i.e. VDS=$\sum$S_VDS$_n$, n=1,2,…,N, where S_VDS$_n$ is the sub virtual dataspace, N is the total number of sub VDS. The data sets of S_VDS$_n$ which related the specific subject are the subset of DS, and the data relationship sets of S_VDS$_n$ which also related the specific subject are the subset of DRS.

Based on this, we could construct the data management mode by using the VDS technology. It adopts the methods of similarity calculation and weights allocation to build the mapping and evolution model about the data and their relationships. Thereby, it could get through the conceptual model from the underlying Data Resources (DR) to the Data Sets (DS) and the Data Relationship Sets (DRS) in VDS. And then, it could further build the global semantic view for mapping the data and relationships to the upper layer services.

(3) Conceptual model of Requirement Space (RS)

Definition 4: Requirement Space is a four-tuples, RS= (U_i, RQ_j, SS_{ij}, SRS_{ij}), i=1,2,…,I, j=1,2,…,J, where I is the number of users, J is the number of requirements, U_i is the User of RS, RQ_j is the Requirement included in RS, SS_{ij} is the Service Sets which is needed by users, SRS_{ij} is the Service Relationship Sets.

Firstly, adopt the OWL-S technology to describe, register, publish and obtain the services in SS_{ij} and the service relationships in SRS_{ij} for supporting the semantic association mapping of services. Then, find the regularity of synergistic combination from various services, and depth mining the associated coordination mechanism by using the workflow technology. Thus it could efficiently achieve the service semantic collaboration which is demand-oriented. Finally, establish the mapping from VDS to RS based on the global semantic view.

(4) Conceptual model of Behavior Space (BS)

In order to provide the more real-time and initiative data services, we propose the Behavior Space (BS) as a new concept and method in OCSA.

Definition 5: Behavior Space is a four-tuples, BS= (U_i, BH_k, DS_{ik}, SS_{ik}), i=1,2,…,I, k=1,2,…,K, where I is the number of users, K is the number of user operator behaviors, U_i is the User of BS, BH_k is the Behavior included in BS, DS_{ik} is the data sets which are influenced by the behavior BH_k of user U_i, SS_{ik} is the service sets which are influenced by the behavior BH_k of user U_i.

Behavior Space could adopt the methods of statistics analysis and probability comparison to track the user behavior and discover the regularity of user habits. Through the conceptual element DS_{ik}, it could establish the contacts between BS and VDS. And through the conceptual elements U_i and SS_{ik}, it could establish the contacts between BS and RS. Then, based on this, blend the VDS, RS and BS together, it could support the improvement of data service, and provide the active service in OCSA. The final conceptual level framework of OCSA is shown in Fig. 2.

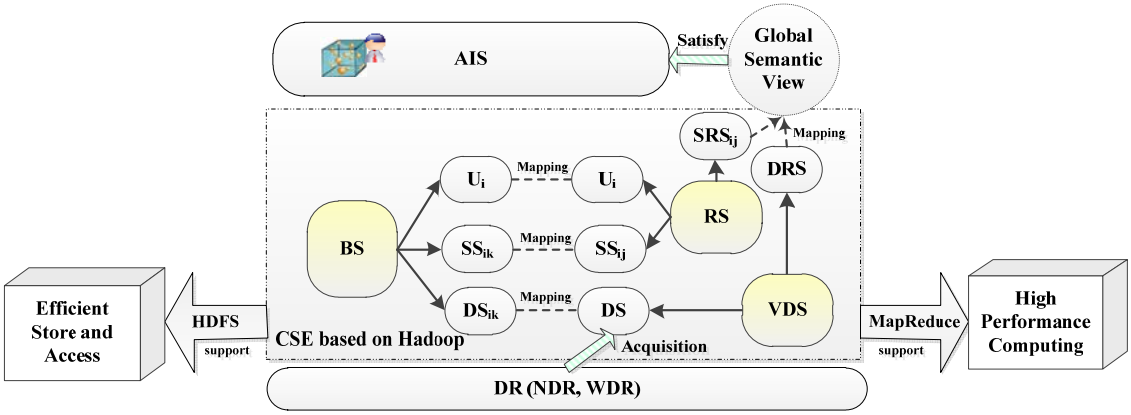

Figure 2. The conceptual level framework of OCSA

IV User Behavior Modeling In OCSA

From the above concepts and theories about OCSA, we can see that this data services framework based on the cloud environment could well solve the special problems, such as the distributed heterogeneous data resources, the complex association, the conditional correlation, etc. It could support the reuse and sharing of the domain scientific data. Nevertheless, there also exists some issues which need to be improved, especially the aspects

of dynamic evolution and real-time service. This requires the consolidated in-depth study from all the different research emphases. It mainly includes three aspects: the data evolution analysis in VDS, the user requirements modeling in RS, and the user behavior modeling in BS. This paper mainly focuses on the method research of the user behavior modeling; the more depth discussion about the data evolution and the requirements modeling would be elaborated in detail in our other papers.

A. Definition of User Behavior

Considering the conceptual model of behavior space, and combining the detailed description of BS in Definition 5, the analysis of user behavior mainly concentrated in the two concept tuples: U_i and BH_k. Around the two tuples, deeply analyze and research the user behavior in detail, we could get the more refined conceptual model about user behavior.

Definition 6: User Behavior (UB) refers to all of the user access operations and the regularity about user action habits which hidden in the superficial user access operations. User Behavior is a tuple, UB= (US, BHS), where US is the User Sets, UBH is the Behavior Sets.

It could be seen that User Sets means all the collection of users, i.e. $US=\sum U_i$, i=1,2,…,I, where I is the number of users. Similarly, Behavior Sets means all the collection of behaviors, i.e. $BHS=\sum BH_k$, k=1,2,…,K, where K is the number of operator behaviors. Therefore, the User Behavior Sets also means all the collection of operator behaviors which are generated by all the users; i.e. $UBS=\sum UB_{ik}=\sum UB_m$, m=1,2,…,M, where M is the number of user operator behaviors, and M=I*K.

Thereby, the user U_i and the behavior BH_k could be defined as follows.

Definition 7: User is a four-tuples, U= (u_sour, u_info, u_rela, u_inter).

The u_sour denotes the user source, which includes the related information of user, such as the region, IP address, source link, browser, operating system, etc.

The u_info denotes the user basic information, which includes the user judge, user name, password, permission, contact manner, etc. Where the user judge is a Boolean value, "1" represents the registered user, and "0" represents the ordinary visitor.

The u_rela denotes the relationships among users. It is a four-tuples, u_rela=(u_rela_name, u_rela_type, u_rela_content, u_rela_weight). In which the u_rela_name possesses the semantic feature, so that we could manage the semantic consistency of user relationships to solve the problem of semantic divergence. The u_rela_type denotes the logical relationship type in the permission level, which could be the father, son, neighbor, similar, opposition, etc. The u_rela_content is a N-tuples, u_rela_content=(u_1, u_2, …, u_N), it denotes that there exists the specific relationship between this user and those users. The u_rela_weight denotes the degree of association among the users in this relationship.

The u_inter describes the interested information by this user. It mainly includes the interested data, interested service, interested operation, and so on.

Definition 8: Behavior is a six-tuples, BH=(bh_name, bh_type, bh_time, bh_user, bh_obje, bh_rela).

The bh_name possesses the semantic feature, so that we could maintain the semantic consistency of behavior description.

The bh_type denotes the user operation type, which could be divided into many types by considering from different angles. For details, see below.

The bh_time denotes the user operation time, which contains two types of expression, one is the determined time point, and the other is the form of time range.

The bh_user denotes the user who led to this operation behavior. It could be associated with the user name in u_info in order to build the association mapping between user and operation behavior.

The bh_obje denotes the object which is operated by user. It could be the page link address, the specific data value, the query keywords, etc. Additionally, there exists the association degree between behavior and object, thus it could be denoted as "bh_obje_degree".

The bh_rela denotes the relationships among operation behaviors. It is a four-tuples, bh_rela=(bh_rela_name, bh_rela_type, bh_rela_content, bh_rela_weight). It is very similar to the definition of concept u_rela, therefore not be repeated here.

B. Classification of User Behavior

The user operation behavior could be classified into different types by considering from different angles. It could further extend more information by classifying the user behavior for the parameters "bh_type" in definition 8. The detailed description about the classification of user behavior is shown in Table Ⅱ.

TABLE Ⅱ.
CLASSIFICATION OF USER BEHAVIOR

Angle	Classification	Example
User-centric	Time distribution of user online, page distribution of users' access, habits comparison of registered users and non-registered users…	The number of online users is concentrated in the daytime.
Operation	Add, delete, modify, browse, query…	Add a line of data.
Page URL	Frequency of page access, page view, bounce rate, the previous page visited, the next page going to visit…	Today's page view is 200.
Time	Time on page, the last access time, visit situation in different time periods, access process in time order…	The last access time of a certain user is "2013-01-01".
Query related	Query number of keyword, associated words of keyword, clicked page after query…	The query number about the keyword of "Alloy" is 1000.
Data service	Data traffic (upload or download), normal rate of service…	100M downloads one day.
Behavior relation	Followed by different behaviors, inherited from the same behavior, similar behavior characteristics…	One behavior generally followed by another behavior.

C. User Behavior Tracking

Combining the above description about the definition and classification of user behavior, and considering the user, operation, behavior relationships, etc., we could track the changes of user behavior from the time dimension and the space dimension.

(1) User behavior tracking based on time dimension

Along the timeline of parameter "bh_time" in definition 8, conduct the probability statistics analysis for user operation behavior. When analyze the user behavior around the fixed parameter "bh_name" or "bh_type", i.e. statistics for the same operation behavior based on time dimension, we could get the time distribution situation about some particular operation behavior. When analyze the user behavior around the fixed parameter "bh_user", i.e. statistics for the same user based on time dimension, we could get the distribution situation of access time about some particular user. When analyze the user behavior around the fixed parameter "bh_obje", i.e. statistics for the same operator object on time dimension, we could get the time distribution situation about some particular operator object. For example, the frequency of access "ac1" by user "u1" for object "ob1" in the latest month is a very high number, it means that the user "u1" pay more attention to the object "ob1" recently.

Assume the parameter "bh_name", "bh_type", "bh_user" and "bh_obje" are also not fixed, we could get the behavior sequence based on the time dimension. Conduct the probability statistics analysis around the behavior sequence, could amend and improve the parameter "bh_rela" gradually, i.e. dynamically maintain the relationships among the different user behaviors. For instance, there exists a fixed behavior sequence (A, B) sorted by time, if it has a very high probability of appearance, we could define the relationship between A and B as "guidance" and "inheritance", i.e. behavior A is the guidance of behavior B, meanwhile behavior B is the inheritance of behavior A. It is symmetrical between "guidance" and "inheritance". Generally, it has similarity between two inheritances which from the same guidance. For example, behavior C is also the inheritance of behavior A, thus there exists the higher degree of correlation between behavior B and behavior C. In this case, the parameter "bh_rela_weight" should be changed. In addition, if the time on page is long, or the last access time is very close, or the frequency of page access is a great number, thus the parameter "bh_rela_weight" should be increased.

In summary, we could track the user behavior based on time dimension to get more information such as the previous operation, the next operation, the similar operation, and so on. Thereby, through the user behavior tracking based on time dimension, we could recommend the next operation for users, and further improve the weight values of user behavior relationships along with the adoption or abstain by users.

(2) User behavior tracking based on space dimension

Around the parameter "bh_obje" in definition 8, the parameter "bh_obje_degree" could be changed by two manners.

One is the changes of parameter "bh_rela_weight". For example, when the weight of behavior relationship changed, it would likely provoke a new recommendation of operation behavior. Then it might lead to a new relationship between behavior and object. Thereby the association degree between behavior and object has been created.

The other is the probability statistics analysis for the user operation behavior. For example, user "u1" browses the page about "iron alloy material properties data" very frequently, then the association degree between behavior "browse" and object "iron alloy material properties data page" should be increased. Combining with the data association analysis in Virtual DataSpace (VDS), we could find that there exists a high association degree between "iron alloy" and "steel alloy". Then the page of "steel alloy material properties data" should be recommended for the user "u1". If user "u1" adopts this recommendation, a new operation behavior should be created, and meanwhile the data association degree between "iron alloy" and "steel alloy" in VDS should also be increased.

In summary, we could track the user behavior based on space dimension to get more information such as the page recommendation about data service, the changing of data association in VDS, and so on. It also could be regarded as the user behavior feedback for VDS.

D. User Behavior Model

From the above description and analysis, we could construct the User Behavior Model (UBM) for data services. The cloud environment of OCSA is very suitable for the modeling of user behavior. The UBM in the cloud environment mainly could support the following technical contributions.

- It is more convenient to realize the cross-domain tracking of user behavior.
- The behavior feedback could improve the data association and optimize the service recommendation.
- The data services combined with the UBM in OCSA would be more accurate and timely for users.

As shown in Fig. 3, construct the associations among UBM, RS and VDS based on the parameter mappings. Through the parameter mappings of "u_info", "bh_user", "U_i", "u_inter", "bh_obje" and "SS_{ij}", build associations between UBM and RS. Through the parameter mappings of "u_inter", "bh_obje" and "DS", build associations between UBM and VDS. Thereby, track the behavior operation by the supporting of parameter "bh_rela" from the time dimension, and in turn to support the change of parameter "bh_rela_weight". On the other hand, track the behavior object by the supporting of parameter "bh_obje" in UBM and data association in VDS from the space dimension, thus in turn to support the change of

parameter "bh_obje_degree" in UBM and parameter "data association degree" in VDS.

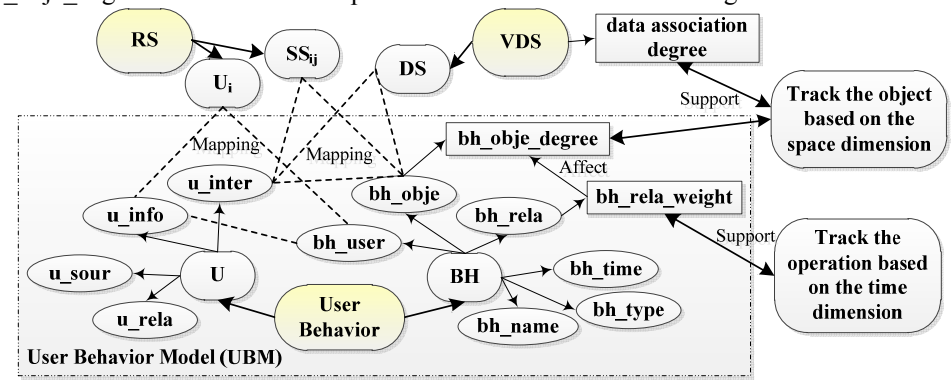

Figure 3. The user behavior model

In summary, we could get through the user demands, operation behaviors and data services by the supporting of parameter mapping and parameter improvement during the process of user behavior modeling. And then, we could construct the user behavior feedback mechanism to provide the individualized data services which are constantly improved.

V. Application Case In The Field Of Materials Engineering

Considering the features of data services in materials science, such as the complex association, the real-time dynamic changes, the conditional correlation, and so on, the user behavior analysis is very useful for the evolution and improvement of data services in the field of materials engineering. Therefore, we designed and realized the behavior-oriented Materials Scientific Data Sharing Service Platform based on the user behavior modeling.

This data sharing service platform has integrated the massive and heterogeneous data resources, such as ferrous materials, energy materials, organic polymer materials, biomedical materials, etc., which distributed in more than twenty research institutes that located in different regions. Currently, it has collected more than five hundred thousand data resource items, and the total visits have exceeded sixteen million.

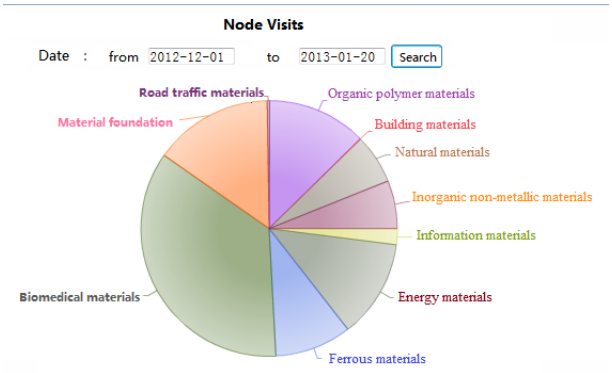

Figure 4. An example of node visits statistics

Fig. 4 illustrates an example of node visits statistics based on the user behavior modeling of materials domain. First delimit the time range as from "2012-12-01" to "2013-01-20". Then search and get the statistics situation of node visits. Thereby, through the analysis of this pie chart, we could acquire the related knowledge that the relevant information about biomedical materials is an access hotspot recently. Based on this understanding, we could recommend some specific data services about biomedical materials for the suitable users. Accordingly, it could realize the individualized data services for the materials scientific data sharing by the wide variety of user behavior analysis.

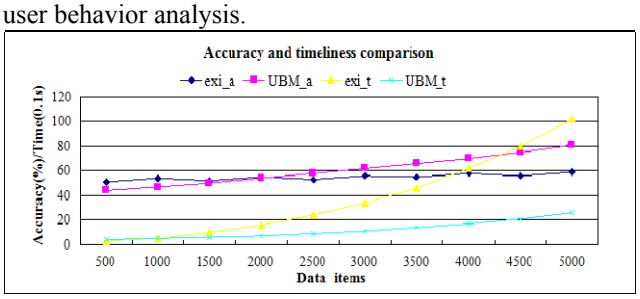

Figure 5. The accuracy and timeliness comparison

Fig. 5 compares the proposed data services mode with the model in the existing works from two aspects of accuracy and timeliness. Through the experimental analysis and comparison, it could be seen that when the data items has gradually increased, the accuracy only has a little change for the data services mode which is without considering the user behavior analysis; but the accuracy for the data services mode which is based on the user behavior modeling would significantly improve with the increasing of data items. Meanwhile, compared with the other models, UBM has the significant advantages for the data services recommendation in the aspect of timeliness. Thereby, it could verify and confirm the effectiveness of the user behavior modeling approach in the open cloud service environment.

VI. Conclusions and Future Work

In this paper, an effective user behavior modeling approach for data services in the cloud environment is proposed. We have realized the accurate and timely data services in the field of materials engineering by describing the related conceptual theories and analyzing the tracking process of user behavior. The future work could be summarized as the following two aspects. On

the one hand, further in-depth analyze the user behavior patterns for mining more valuable research ideas, and then improve the user behavior model. On the other hand, based on the open cloud service environment, design a set of analysis rules about the user behavior, and adopt the optimal algorithm to support the quantitative analysis about the user behavior association, thus achieve the more personalized and diversified data services.

ACKNOWLEDGEMENT

The work in this paper is supported by the R&D Infrastructure and Facility Development Program under Grant No. 2005DKA32800, the 2012 Ladder Plan Project of Beijing Key Laboratory of Knowledge Engineering for Materials Science under Grant No. Z121101002812005, the Key Science-Technology Plan of the National 'Twelfth Five-Year-Plan' of China under Grant No.2011BAK08B04, and the National Key Basic Research and Development Program (973 Program) under Grant No. 2013CB329606.

REFERENCES

[1] C. Lynch, "Big data: How do your data grow?," *Nature*, vol. 455, no. 9, pp.28-29, 2008.
[2] D. Howe, M. Costanzo, P. Fey, et al., "Big data: The future of biocuration," *Nature*, vol. 455, no. 9, pp.47-50, 2008.
[3] G. Wei, "Complex Learning System for Behavior Factor based Data Analysis," *Journal of Software*, vol. 8, no. 4, pp.1003-1010, 2013.
[4] Q. Chen, Y. Ou, H. Sun, "Design and Implement of Customer Communication Behavior Analysis System," *Journal of Software*, vol. 6, no. 8, pp.1484-1491, 2011.
[5] M. Franklin, A. Halevy, D. Maier, "From databases to dataspaces: a new abstraction for information management," *In Proceedings of the ACM SIGMOD Record*, pp.27-33, 2005.
[6] R. L. Grossman, Y. Gu, J. Mambretti, et al., "An overview of the Open Science Data Cloud," *In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pp.377-384, 2010.
[7] L. Zhang, Q. Zhou, "CCOA: Cloud Computing Open Architecture," *In Proceedings of the IEEE International Conference on Web Services*, pp.607-616, 2009.
[8] S. Loebman, D. Nunley, Y. C. Kwon, B. Howe, M. Balazinsk, J. P. Gardner, "Analyzing massive astrophysical datasets: Can pig/hadoop or a relational DBMS," *In Proceedings of the Workshop on Interfaces and Architecture for Scientific Data Storage (IASDS)*, pp.1-10, 2009.
[9] M. F. Husain, L. Khan, M. Kantarcioglu, B. Thuraisingham, "Data Intensive Query Processing for Large RDF Graphs Using Cloud Computing Tools," *In Proceedings of the IEEE 3rd International Conference on Cloud Computing (CLOUD)*, pp.1-10, 2010.
[10] L. Youseff, M. Butrico, D. D. Silva, "Toward a Unified Ontology of Cloud Computing," *In Proceedings of the Grid Computing Environments Workshop 2008 (GCE'08)*, pp.1-10, 2008.
[11] L. Blunschi, J. P. Dittrich, O. R. Girard, S. K. Karakashia, M. A. V. Salles, "A dataspace odyssey: The iMeMex personal dataspace management system," *In Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR)*, pp.114-119, 2007.
[12] X. L. Dong, A. Halevy, "A platform for personal information management and integration," *In Proceedings of VLDB 2005 PhD Workshop (CIDR'05)*, pp.26-30, 2005.
[13] J. Madhavan, S. R. Jeffery, S. Cohen, et al., "Web-scale Data Integration: You can only afford to Pay As You Go," *In Proceedings of CIDR*, pp.342-350, 2007.
[14] A. D. Sarma, X. Dong, A. Halevy, "Bootstrapping pay-as-you-go data integration systems," *In Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp.861-874, 2008.
[15] Z. Liu, C. Hu, Y. Li, "DSDC: a domain scientific data cloud based on virtual dataspaces," *In Proceedings of the 26th IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)*, pp.2176-2182, 2012.
[16] X. Hu, T. Mu, W. Dai, H. Hu, G. Dai, "Analysis of Browsing Behaviors with Ant Colony Clustering Algorithm," *Journal of Computers*, vol. 7, no. 12, pp.3096-3102, 2012.
[17] C. Xu, W. Shi, Q. Xiong, "A Novel User Click Behavior Identification Method Based on Hidden Semi-Markov Model," *AISS: Advances in Information Sciences and Service Sciences*, vol. 3, no. 11, pp.52-59, 2011.
[18] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, "Characterizing User Behavior in Online Social Networks," *In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (IMC'09)*, pp.49-62, 2009.
[19] I. Elsayed, P. Brezany, A. M. Tjoa, "Towards Realization of Dataspaces," *In Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)*, pp.266-272, 2006.
[20] L. Mao, Y. Yang, H. Xu, "Design and Optimization of Cloud-Oriented Workflow System," *Journal of Software*, vol. 8, no. 1, pp.251-258, 2013.

Xin Cheng was born in Henan, China, in 1983. She is a Ph. D candidate of School of Computer and Communication Engineering, University of Science and Technology Beijing, China. Her research interests include data engineering, knowledge engineering and semantic web.

Changjun Hu was born in Hebei, China, in 1963. He received the PhD degree in parallel computing from Peking University, where he is a Full Professor and Doctoral Advisor, the Associate Dean of School of Computer and Communication Engineering. His research interests include data engineering, software engineering, parallel processing and cloud computing.

Yang Li was born in Shandong, China, in 1983. She received the PhD degree in evolutionary computing from University of Science and Technology Beijing, where she is a lecturer in School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interests include evolutionary computing and data mining.

Wei Lin was born in Fujian, China, in 1990. She is a master candidate of School of Computer and Communication Engineering, University of Science and Technology Beijing, China. Her research interests include ontology construction and semantic mapping.