International Journal of u-and e-Service, Science and Technology
Vol.7, No.5 (2014), pp.329-340
http://dx.doi.org/10.14257/ijunnesst.2014.7.5.29

The Research of Data Mining Based on Application Data Pool

Minjie Bian[1], Jue Gao[1], JiePin Xu[2] and Honghao Gao[3]

[1]*School of Computer Engineering and Science Shanghai University*
[2]*Shanghai Shang Da Hai Run Information System Co., Ltd*
[3]*Computing Center, Shanghai University*
200444 Shanghai, P.R. China
bianmj0302@shec.edu.cn

Abstract

Today, people use various kinds of information technology applications to deal with applications in our daily life, which generates lots of information. However,most of the informationis just stored in many Distributed Heterogeneous Databases (DHDs) as log records, instead of being used abundantly and effectively. In this paper, we mainly discusses about how to use these data in useful ways by Data Mining (DM). Relative to thetraditional Data Mining based on Data Warehouse (WD),we propose a definition named Application Data Pool (ADP)replacing WD in this paper. And we design a KnowledgeDiscovery in Databases (KDD) model with ADP to use these data more efficiently. At last we use an application in Shanghai University ID Card designed with the ADP to prove the effectiveness and feasibility in KDD.

Keywords: *KDD, DM, DW, DHDs, ADP*

1. Introduction

With the development of the network and computer technology, we deal with various kinds of applications in our daily life, such as OA, daily shopping, ID authentication and so on. These applications generate lots of data which describes people's daily behaviors. Most of these data are just stored in different kinds of databases instead of being taken advantage well. In this paper, we design a new way, based on ADP, to do data mining in which we can analyze the data more effective and more feasibility so that the decision analysis can be made more accurately.

The traditional KDD [1, 2] model can be used to find useful information from various kinds of databases. The KDD includes Data Preprocessing, Data Mining, Data Assessment and Knowledge Representation. The traditional KDD model is shown in figrue1. The data are generated from various applications and stored in DHDs. Preprocessing these data creates the DW. The DM is based on the DW and sometimes some themes involved. After the process of DM, the regular data can be obtained. Some decision analysis can be executed based on these data and knowledge can be refined by data assessment. Finally, the knowledge is showed to the user by knowledge representation, such as a report or a chart.

ISSN: 2005-4246 IJUNESST

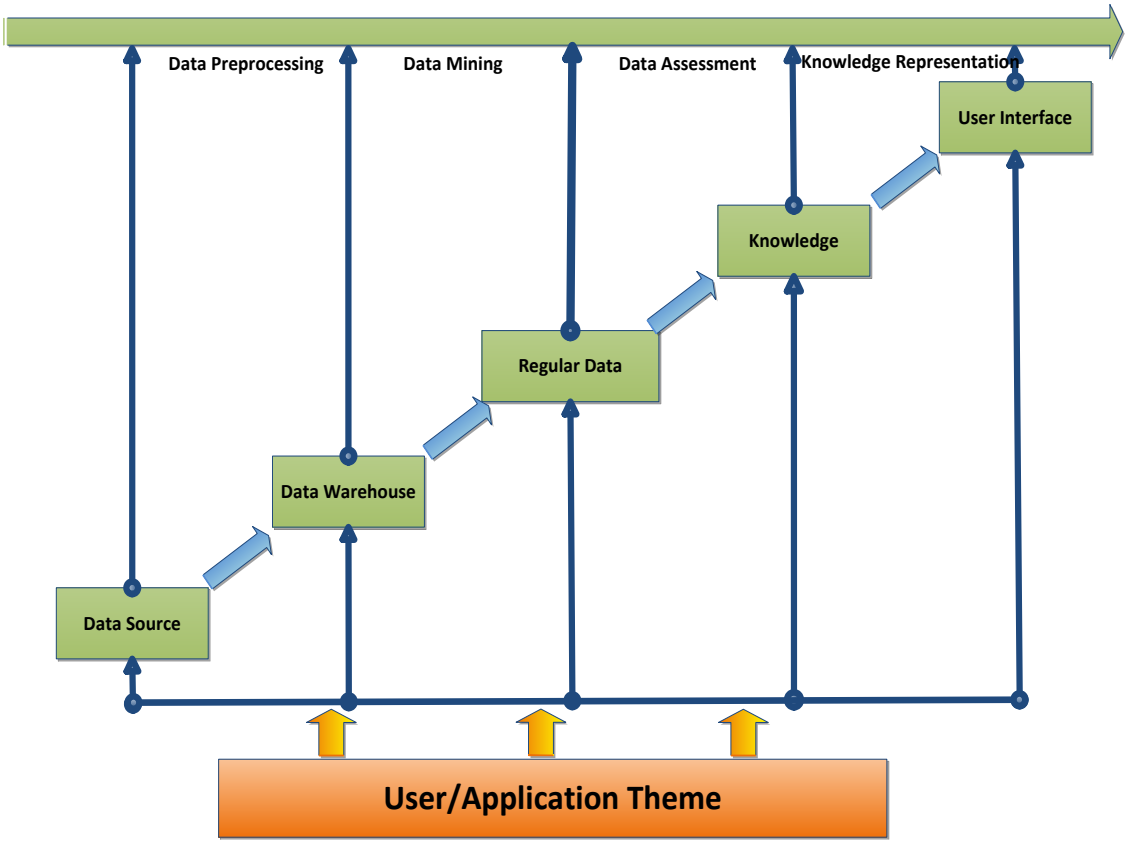

Figure 1. The Traditional KDD Model

As shown in Figure 1, the DM is based on the DW. In this paper, we will design another data container, Application Data Pool (ADP), to deal with the data for DM which will improve the feasibility and effectiveness of KDD. The KDD model will be changed as figure 2. Moredetail will be discussed in section 2.

DM is a process which can reveals implicit, previously unknown and potential information from a large amount data in the database [3]. The main methods of DM includeEstimation, Classification, Prediction, Affinity Grouping or Association Rules, Clustering etc. We choose classification to put data into several categories for our application in this paper which can help the managers make the right decision in daily work.

This paper will be organized as follow: In the first section, we describe the meaning of our research.In the second section, we will propose a new definition: Application Data Pool (ADP) which is the main research of this paper. In the third section, the KDD model based on ADP will discussed. Then we will give an application example for our model in the fourth section. In the fifth section, the conclusion will be made.

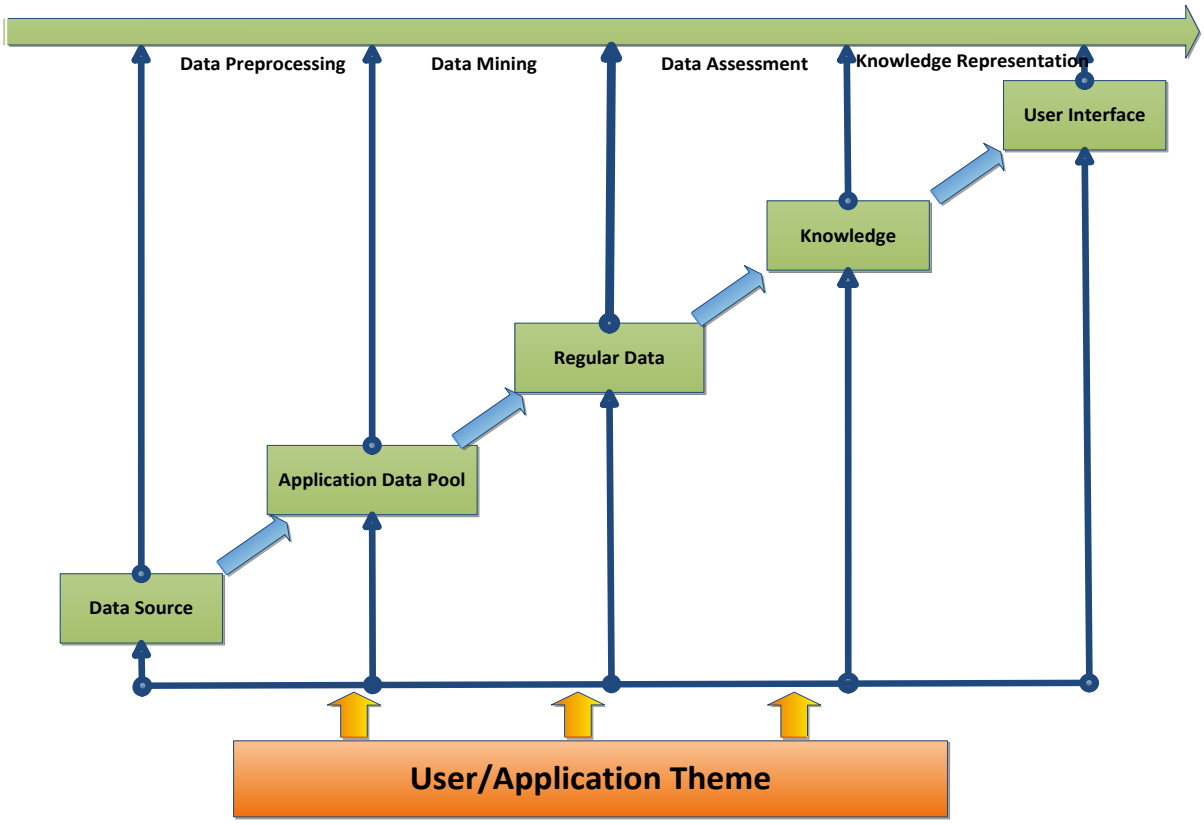

Figure 2. KDD with ADP

2. Application Data Pool

With the development of the network and computer technology, many kinds of applications are applied in our daily life and these applications generate lots of data in databases. For some reasons, different application use DHDs (MySQL, SQL Server, Oracle, etc.) to store their own data, so the first problem is how to collect data from DHDs.

The tradition KDD use DW as data source to support DM.However, the DW cannot guarantee the efficient of the data. Because, the data will be increased or updated every moment in actual applications, and there may be some data redundant or out of date.These will increase the computing pressure and decrease the precision of DM.

In order to solve the problem, we propose a new definition in this paper, Application Data Pool (ADP). ADP is a data container which imports new data from DHDs where applications store their data, and removes the dataout of date or redundancy as well. ADP can automatically change its capacity and the speed of data importingor dataremoving according to the changing of the data sources or the purpose of DM. The ADP will keep the balance between the data importing and data removing finally. ADP can give a better support for DM in KDD. The structure of ADP is shown in Figure 3. ADP has a data importing controller, a data removing controller and a capacity balance controller.

2.1. Data Importing Controller

Data Importing Controller decides what data source to be imported to the ADP and control the flow of the data. The Data Importing Controller will do the Data Preprocessing, such as data integration, data cleaning. The Data Importing Controller will decide the implementation plan of the data importing, such as the importing frequency and the importing quantity, according to the user theme.

2.2. Data Removing Controller

Data Removing Controller decides the implementation plan of the data removing, such as redundant data or the data out of date, to control the flow of the data. If there is more theme to be analyzed or the applications changed, the ADP capacity can be changed by adjust the data removing controller.

2.3. Capacity Balance Controller

Capacity Balance Controller is designed to monitor the ADP capacity dynamically. According to the variability of the application and themes, the ADP capacity should be changed so that it can support the decision analysis. The Capacity Balance Controller adjust the ADP capacity by adjust the Data Importing Controller or Data Removing Controller dynamically.

2.4. Structure of ADP

As the Figure 3, we import data from DHDs to ADP, and remove the data which is out of date or redundancy. At the same time ADP can automatically change the capacity and the speed of data importing and removing, like a water pool where we pour waterfrom the top of the pool and leak water from the bottom. When the DHDs are changed, ADP can find the best balance point quickly automatically.

The design of ADPas follow:

Step1: Import the data from DHDs at a certain time interval. Design the Data Importing Controller according to the data sources of the applications and the themes.

Step2: Preprocess the data. The data preprocessing is to clean the redundant data or integrate data.

Step3: Use the capacity balance controller to monitor the ADP andjudge whether to increase or decrease the capacity of ADP or not by the change of the application and themes.

Step4: If its need to increase or decrease the capacity, the capacity balance controller will adjust the data importing controller or data removing controller and keep onmonitoring them to make ADP achieve the targetcapacity.

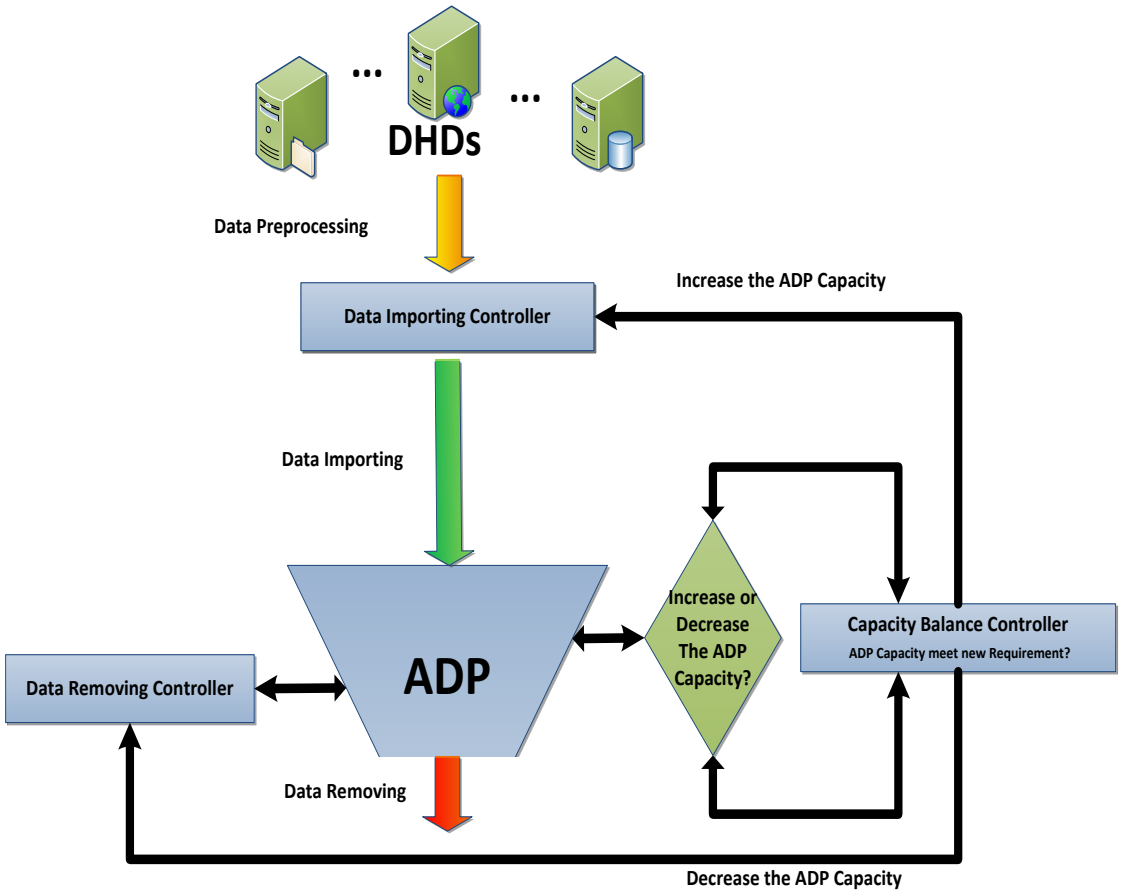

Figure 3. Structure of ADP

3. KDD Model with ADP

In thesecond section, we designADPas the data container to support DM. In this section, we will propose anapplication model using KDD modelbased on ADP. The application model is showed in Figure 4.

As Figure 4, we import the data from DHDs involved with the application at first. Because the capacity of each database from DHDsis often huge, we must importthe data in different efficient ways, such as we use Bulk Copy to import these data for SQL Server database because Bulk copy can import data more quickly and efficiently. Secondly, we need to preprocess the data, in this process we can delete the dirty data and keep the useful data so that our DM will be more precise. After this process, we got all the data which we need and put them into the ADP. Thirdly, we use the data in ADP to support DM. Use the statistics and analysis to make a report which can prove the effectiveness and feasibility of MD based on the ADP.

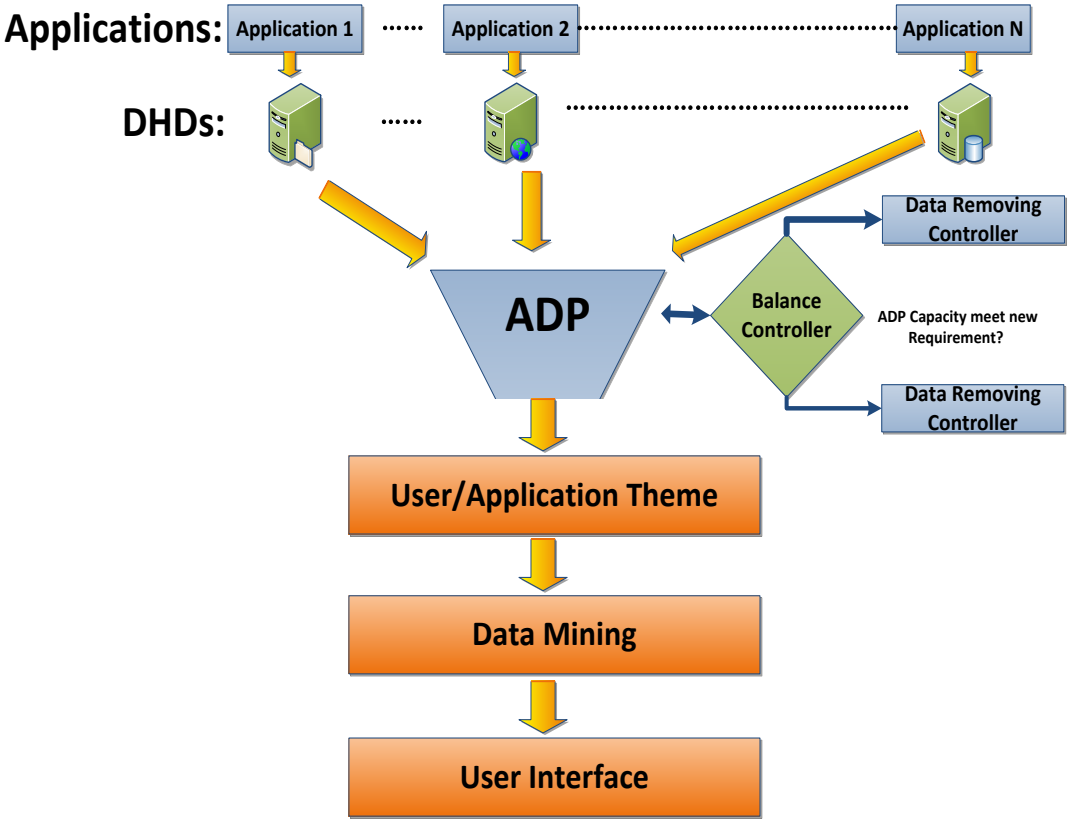

Figure 4 The Model of this Paper

4. Application Examples

Today, almost every university in Shanghai provide their student with campus cards. This kind of cards can be used as identify cards and consumption cards in the university, and it can be used to borrow books from the library as well. These applicationsgenerate thousands of record every day, and these data are stored in different database, such as the identify application may store the data in a SQL Server andthelibrary application may store the data in an Oracle database. So the databases are DHDs. In the Section 3, we proposed a new KDD model with ADP. In order to describe the model more detail, we will useStudent Behavior AnalysisSystem based on Campus Cards of Shanghai University (ShuSBAS)as an example. The main process of the ShuSBAS is shown in Figrue5.

We first import the data from all these databases in different way. In order to use these data more efficient, wepreprocessthese data toget rid of the dirty data, so the data in the ADP will be more pure.

We designed a capacity balance controller to monitor the ADP. The capacity balance controller in our application runs at 2:00 am very day. It will use the Data Importing Controller and Data Removing Controller to update the ADP. When the capacity balance find the ADP capacity should to be change, it will adjust the Data Importing Controller or the Data Removing Controller to change the ADP capacity.

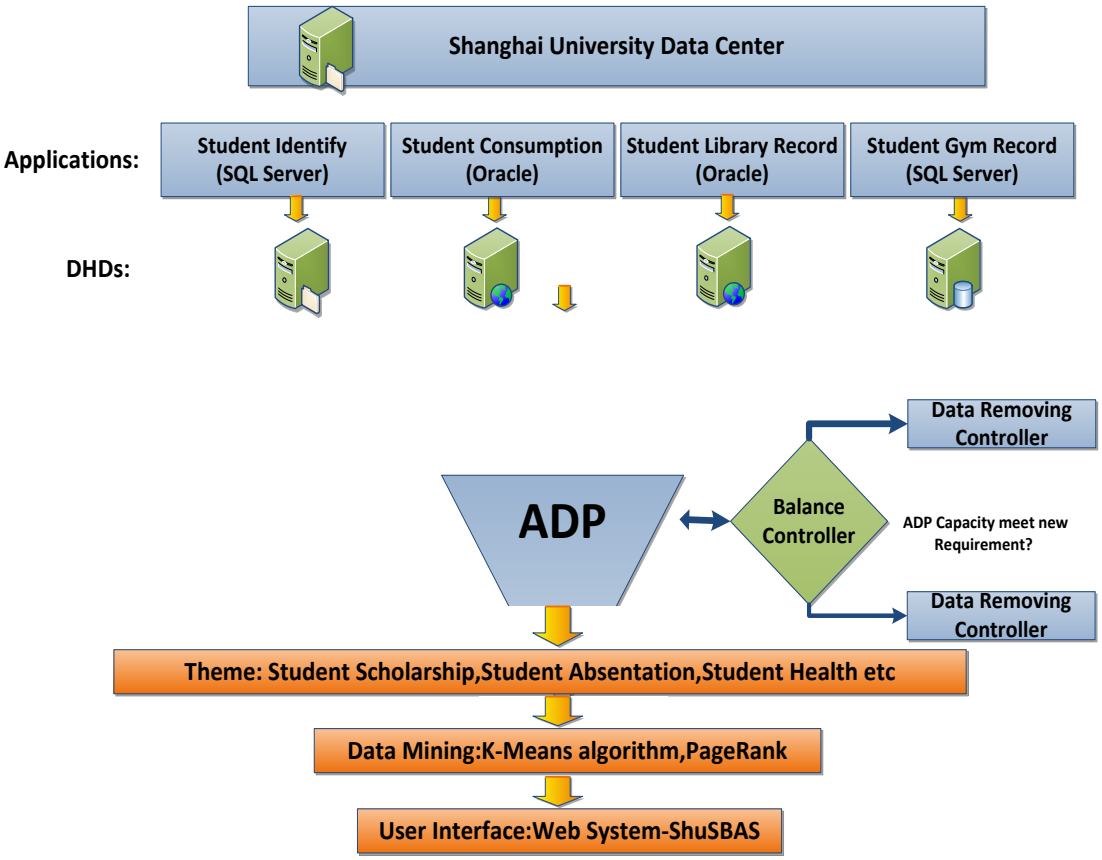

Figure 5. The Process of the Shanghai University Card Behavior System

We can begin to DM based on ADP, we use the K-Means algorithm [7] to find out the borrowing amount is proportional to the grade of the student in some way, and use the PageRank [8] we find out the borrowed amount is proportional to the value of the book.Figure 6-Figure8 is the user interface of ShuSBAS.

4.1. Finding out the Students who have Potential to get Scholarship

The student cardscontain lots of information, such as book borrowing record, consumption record and so on. According to the theme about the behavior records of students in school,we use K-means to deal with the ADP. The result shows that one cluster of students have potential to get scholarship. They spent lots of time in the library and borrow many books for study, show as Figure 6.

Figure 6. ShuSBAS--Find out the Students who have Potential to get Scholarship

4.2. Finding out the Students who may be Absent

In Shanghai University, students use their cards to entry permit in the campus and shopping by the cards and borrow books and so on. All these behavior will generate log records, which imply the location information of the students. So we can find out the students who absent for several days. Teacher will take attention to these students, which provides the security for students in campus. However, if we DM based on the traditional DW, we can't make it. For example, about 300000consumption log records are generated every single day. It's impossible for us to deal with these data in web application. Shanghai University has many other applications generating log records by student card as well.

In the paper, the ADP makes it possible to execute DM with the huge data, show as the follow Figure 7. The capacity balance controller keeps the ADP in appropriate capacity to execute DM, which makes the decision analysis moreeffectiveand feasible.

Figure 7. ShuSBAS--Find out the Students who may be Absent

4.3. Analysis of the StudentHealth

According to the change of themes, we can reset the ADP capacity to provide the DM with appropriate data. If we need know the students healthof Shanghai University, we should to change the implementation plan of the data importing. In the paper, we change the Data Importing Controller adding data from Shanghai University gym and hospital to the ADP. Then we reset the capacity balance controller to update the ADP so that we can do the statistical analysis about the students health, as shown in the Figure 8, which proves the KDD with ADP has more flexibility.

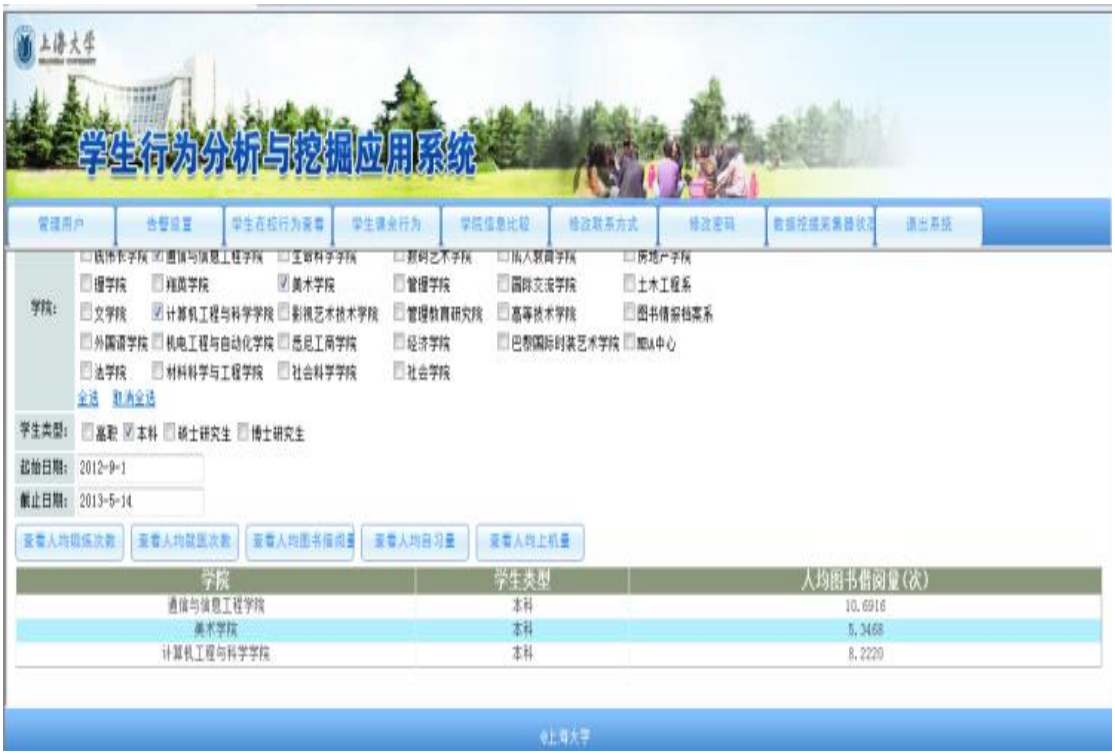

Figure 8. ShuSBAS--Find out the College whose Students may be Lack of Physical Exercise

5. Conclusions

In this paper, instead of using DW, we proposed the ADP to store the data which import from the DHDs where application stores their own data. Based on the ADP, we can DM on these data. The application model we proposed in this papercan be used to all kinds of applications without too much to change. At last, we took the ShuSBAS as an example and use different kinds of DM methods to get the conclusion what we want, which prove the effectiveness and feasibilityof DM based on ADP. The ADP is designed to make DM effective and the KDD model more flexible.

Acknowledgement

This work was supported by Young University Teachers Training Plan of Shanghai Munici pality under Grant No. ZZSD13008.We gratefully acknowledgeand thank those who provided comments and suggestions. The anonymous reviewers and the editor of this paper are also acknowledged for their constructive comments and suggestions. I'd like to express my sincere thanks to all those who have lent me hands in the course of my writing this paper. First of all, I'd like to take this opportunity to show my sincere gratitude to my supervisor, Prof.Xu HuaHu, who has given me so much useful advices on my writing, and has tried his best to improve my paper. Secondly, I'd like to express my gratitude to my colleague, Dr.Gao HongHao who offered me references and information on time. Without their help, it would be much harder for me to finish my study and this paper.

References

[1] M. Kantardzic, "Data mining: concepts, models, methods, and algorithms",**(2011)**.
[2] L. Agosta, "The Essential Guide to Data Warehousing", Prentice-Hall, Inc.**(2000)**.
[3] W.H.Inmon, "The data warehouse and data mining", Communications of the ACM, vol. 39, no. 11, **(1996)**.
[4] T.H.O.Bao,"Knowledge Discovery and Data Mining Techniques and Practice", http://www.netnam.vn /unescocourse/knowlegde/3-5.htm.**(2005)**.
[5] S. Guha, R. Rastogi and K. Shim, "CURE: an Efficient Clustering Algorithm for Large Databases",Information System Journal, vol. 26, no. 1, **(1998)**, pp.35-58.
[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques",Morgan Kaufmann Publishers, **(2000)**.
[7] A. K. Jain,"Data clustering: 50 years beyond K-means", Pattern Recognition Letters, vol. 31, no. 8,**(2010)**, pp. 651–666.
[8] L.Page, S.Brin, and R. Motwani, and T.Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report. Stanford InfoLab, **(1999)**.

Author

MinJie Bian is a PhD candidate in graduate students in School of Computer Engineering and Science Shanghai University. His main research is about computer vision. He works at the Information Technology Office of Shanghai University during the study of PhD.

