Research Article

International Journal of Advanced Computer Research, Vol 7(31)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.731007

Research on visualization methods of online education data based on IDL and hadoop

Yu Lasheng, Wu Xu* and Yang Yu
School of Information Science and Engineering, Central South University, China

Received: 03-March-2017; Revised: 28-May-2017; Accepted: 31-May-2017

Abstract
The research and development of Big Data, Cloud Computing, IoT and other new technologies, provide a strong technical support for vigorously promoting construction of online education. By using the Hadoop distributed file system (HDFS), MapReduce, Sqoop, HBase and other function modules of Hadoop, it can be easy to do normalized processing of massive education data which obtained from different online educational platforms. At the same time, education data can be quickly converted into graphic images, such as line drawing, contour map and grid surface map and so on, by using interactive data language (IDL) to program. These images provide a more scientific and intuitive method of researching on educational data. Experiments show that the visualization of massive education data will play an important role in the process of helping government to carry out scientific, educational decision-making, teacher to launch effective teaching activities and student to improve the efficiency of personalized learning.

Keywords
Big data, IDL, Visualization, Hadoop.

1.Introduction
At present, with the rapid development of scientific research, big data is not limited to economic society, but also to change people's work and lifestyle at an alarming rate along, such as smart city, smart campus, and recommend system are the product of big data. Also, because of big data, there are unprecedented and profound changes having taken place in the field of education. In the future, students can access to all the high quality education resources from education cloud, which based entirely on the individual needs of themselves, because the campus, online education platform, teaching content and other educational resources are extremely rich and all open to everyone [1-3].

Creating educational environments, setting educational classrooms, converting educational time and space, reforming teaching approach, collecting educational data and making educational decisions, all these can rely on past experience and subjective analysis of schools and teachers teaching to determine arrangements, while in the era of big data it will become more scientific and reasonable [4-6].

Previous data processing capacity is limited by the bottleneck of education and learning, data mining analysis, but it will regain vitality with the help of big data technologies now [7]. However, different teaching, learning and data management systems led to the exchange of information sharing very difficult, as information technology applied in education was lagging behind in the past. If a teacher wants to learn about a particular student's learning status, she needs to query several databases. However, in a different database, data format may be quite different, which makes it very difficult to analyze learning situation. To a certain extent, it restricted the improvement of the quality of teaching, learning and management. To introduce big data technology into smart education, we must first solve the issue that a lot of education data format is not uniform and non-standard, which is the basis of data mining and analysis of massive education data. The use of large-scale data processing, open-source framework Hadoop, can effectively achieve standardization of education data, while the use of interactive data language (IDL) can help to visualize and analyze the standardized and normalized education data. All of these can provide effective suggestions for the government to implement decisions, schools to improve the standard of teaching and students to improve learning skills. This paper will excavate valuable information hidden

*Author for correspondence

in students' studying scores as the goal, by using interactive data language (IDL) to explore the methods of education data visualization based on big data environments.

2.Related work

Data visualization techniques can be simply understood as a scientific technique that can used to represent data by converting data into kinds of graphics and images which can make it easier for people to discover the links between data and the law hidden in the data than directly observing on data. Chernoff-faces technique proposed by Herman Chernoff can represent a set of high-dimensional data as a series of cartoon images that map a point to 18 faces of a face for finding the rules of the data. Inselberg et al. [8] introduced a non-projected parallel coordinate system which reduces the dimension of the data by mapping hyper-surfaces to planar graphics. Swayne et al. [9] provided Xgobi system which combines the use of advanced visualization tools and technologies to facilitate human-computer interaction, allowing users to flexibly manipulate the same data with different views, while also providing a number of online help tools integrated into GIS, XGvis and other systems for application. Huang et al. [10] proposed a new high-dimensional data space filling method SFMDVis, to maximize the use of display areas, in order to eliminate the effects of edge crossover and improve the linear correlation between visual perception and different variables. Molina-Solana et al. [11] combined fuzzy logic with large-scale simulation visualization to enhance the processing of image perception and help users to quickly analyze the interest nodes. In this paper, we try to build an online education data visualization framework based on IDL and Hadoop.

3.A new frame of data visualization

IDL is a kind of fourth-generation computer language, was developed by a wholly owned subsidiary of Kodak's RSI in America and planning to invest in the market, which is a matrix-oriented and used for data visualization and application [12]. IDL has been widely used in aerospace, earth science, and military at present. IDL can quickly and easily converted data into images, which will help to promote future analysis and understanding. Math library functions built-in IDL can greatly reduce the workload of image processing algorithms. Programs written in IDL can run without modifications on other platforms that can run IDL, for it can be easily connected to C, C++, and also supports the database operations compatible with the ODBC interface standard. Hadoop developed by the Apache which is a distributed system infrastructure, which leverages the cluster of high-speed computing and storage [13 14]. Users in a distributed system don't need to know the specific circumstances underlying, which can be used for distributed application development. An online education data visualization framework based on IDL and Hadoop is shown in *Figure 1*.

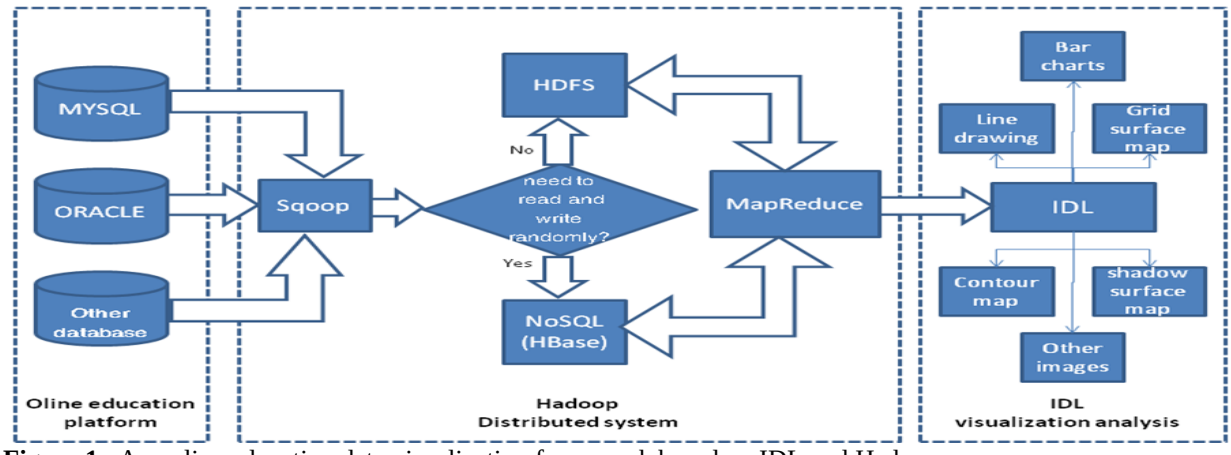

Figure 1 An online education data visualization framework based on IDL and Hadoop

3.1Standardized processing using Hadoop

Although education data formats varied widely, overall they are basically in the form of a database system for storage in a conventional system [15]. There are a number of ways to normalize education data. The traditional method to read a file would cost a long time, for it reads files in sequential order, not simultaneously. If using Hadoop clusters, it can process huge data sets in a parallel way, which can greatly shorten the file access time. Integrated in

Hadoop, HDFS distributed file system and parallel processing architectures MapReduce, and HBase, Sqoop and other subprojects can make it very easy to deal with a variety of educational data [16-18]. In order to facilitate understanding, we draw a data normalization processing flow chart based on Hadoop, the main idea of which is described in *Figure 2*. It is assumed that education data are stored in the Oracle. First, import the data from Oracle into HDFS or HBase by using Sqoop[16]. Then, using MapReduce to program, by which data are purposefully chosen to generate standardized and normalized data files prepared for visualization analysis.

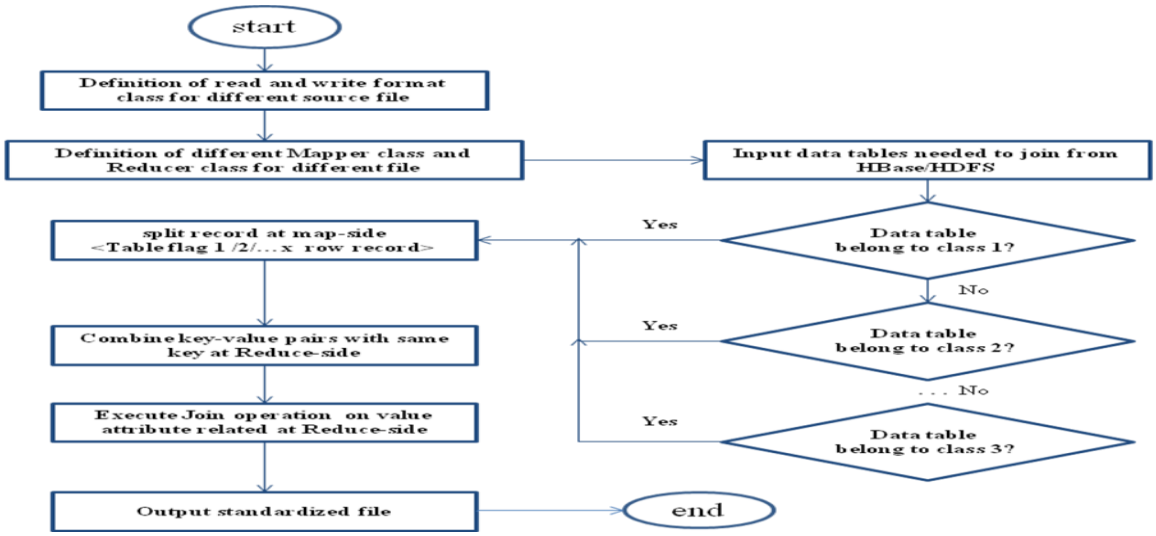

Figure 2 Process of normalizing data using Hadoop

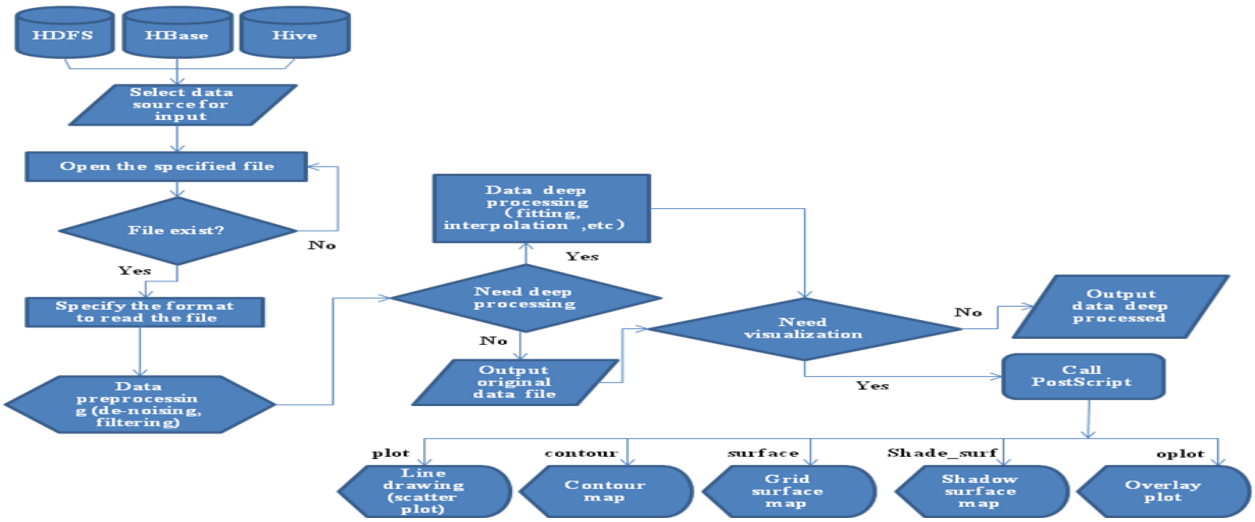

Figure 3 Process of visualization programming using IDL

3.2Visualization programming of data using IDL

One kind of application of the most common features of IDL is drawing. IDL can quickly use data to calculate the results and then to draw intuitive graphical images, especially for easily presenting the data trends or discovering the rules of data. While writing IDL programs, by using plot, contour, surface, shade_surf and other commands with Gsview (a drawing software), it will be easy to obtain line drawings, contour map, mesh surface plot, graphics shadow surface, as well as error bars, histograms, bar charts and so on respectively. With these intuitive graphics, it will be helpful and more accurate for the visualization analyzing of

educational data. The process of visualization programming by using IDL is shown in *Figure 3*.

4.A case for visualization analysis

The approach of visualization analyzing will be described in the case discussed below. We have obtained lots of score tables from some different online education platforms, and the data were generated by students belong to Class 1 to Class 4 from Grade 1 to Grade 6, who come from the same primary school. Perhaps these data are initially distinguished according to the different subjects and types of storage, such as Chinese scores table, Math scores table, English scores table, which have been shown in *Table 1*, having different data type.

It is difficult to observe and do analyzing, as the data is scattered. According to the methods introduced in section 2, we can divide tasks into 3 steps: first, import score tables from Oracle to HDFS or HBase using Sqoop; second, combine the related score tables of students who from the same Class and Grade into a same file using MapReduce programming; at the end, draw images by using IDL to program, and do visualization analysis of the students' score with the help of the images drawn by IDL. Experimental environment includes:

1) The operating system was Win7(32bit);
2) Pseudo distribution system of Hadoop built on Ubantu16.04 which installed on VMware10, containing Hadoop-1.0.4, HBase-0.90.4, Sqoop-1.2.0;
3) The database was Oracle11g;
4) Program codes realized by JDK1.6 and IDL8.3.

4.1Generating normalized files

4.1.1Import data to Hadoop

Taking score tables are stored in the Oracle database as an example. Before importing a score table from Oracle, Sqoop uses JDBC to retrieve all the columns and the corresponding data types of the score tables that will be imported, and then map all of them into java classes. In the process of MapReduce, Sqoop will use the corresponding java class to save the required field value. The Sqoop integrated code generator will use this information to create a corresponding class for keeping records taken from the score table in time. If you don't want to import the entire table every time, you can add key word where in the query clause to limit the import record range.

The command of importing data from Oracle to HDFS is:*sqoop import --connect jdbc : oracle : thin : @host:1521:orcl --username*** --password *** --table $oralceTableName --target-dir $hdfsPath --columns $columns --fields-terminated-by '\001' --m X*

The import process is finished after executing this command. Importing data from Oracle to HBase is similar to HDFS, but you must specify a primary key value. The command of importing data from oracle to the HBase as follows:

sqoop import --connect jdbc:oracle:thin:@localhost: 1521 : XX --username XX --password XX --table XXX --hbase-table XX(newname in HBase) –column--family XX --hbase-row-key id --split-by id --columns XX, XX, XX --hbase-bulkload --m X

4.1.2Join data using MapReduce

To program MapRedcue codes of doing standardized processing of the score tables in HDFS, that is the main purpose of realizing data joins on demand. There are two kinds of methods to join the data together in Hadoop [16]: map-side join and reduce-side join. Map-side join can perform full joins in the mapper function, which can drastically reduce the amount of data transferred to the reducer, but need to define job input rules. Reduce-side join is relatively simple, but the system overhead is too much, for all the data have to go through the shuffle before reach reducer. To ensure that the data is not lost and improve the efficiency of the Join operation, map-side join is recommended. The MapReduce code of Join operation is written relied on multiple input class provided by hadoop-1.0.4, including the definition of file read and write format class(FlagDataType), the definition of different file types mapper way class(MultiMapper1, MultiMapper2 and so on), the definition of Reducer way class(MultiReducer) and set parameters of the operation control four parts. The code of the operation control part is shown in *Table 1*.

If there are more files you can use multiple MultiInputs classes. The final standardized file format is shown in *Table 2*. It's easy to see, with the help of Hadoop, the students score are well to be normalized and stored. By the standardization process, all subjects score of the students in the same class from same grade are statistics into a file, and the data types are unified: Basic information is placed in the file header of the first five lines, and specific scores followed stored in columns, which would provide a great convenience for IDL to do visualization analyzing.

Table 1 The control paramerters setting for join

Class of parameters setting

public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
conf.set("delimiter", "\t");
Job job = new Job(conf, "duoyuanjoin2");
job.setJarByClass(duoyuanjoin2.class);
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(FlagDataType.class);
job.setReducerClass(MultiReducer.class); //set Reducer class
job.setOutputKeyClass(NullWritable.class);
job.setOutputValueClass(Text.class);
//MultiMapper1 for file1, MultiMapper2 for file2
MultipleInputs.addInputPath(job, new Path(args[0]), TextInputFormat.class,MultiMapper1.class);
MultipleInputs.addInputPath(job, new Path(args[1]), TextInputFormat.class,MultiMapper2.class);
Path outPath = new Path(args[2]); //definition output path
FileSystem fs = FileSystem.get(conf);
if(fs.exists(outPath)) {
fs.delete(outPath, true); //if file exist,delete first}
FileOutputFormat.setOutputPath(job, outPath);
System.exit(job.waitForCompletion(true)?0:1);
}

Table 2 Standardized data file（Grade 3, Class c01）

Class:C01
Code:3023
Grade:3
Student information register table

ID	Name	Chinese	Math	English	SXPD	Science	Computer	Excellent-Rate（%）	Pass-Rate（%）	Age(year)
302301	Lily	77.0	92.0	69.0	86.0	84.0	92.0	33	100	8.0
302302	Tom	88.0	80.0	79.5	93.0	70.0	84.0	16	100	8.0
302303	Jim	71.0	68.0	72.0	85.0	89.0	82.0	0	100	8.1
302304	Lisa	57.0	88.5	90.0	88.0	72.0	79.0	16	82	8.1
302305	Bob	65.0	76.0	72.5	79.0	76.0	61.0	0	100	8.2
302306	Rose	70.0	87.0	61.0	72.0	61.0	83.0	33	100	8.2
302307	Green	85.0	50.0	82.0	84.0	78.0	70.0	0	82	8.2
302308	White	52.0	81.5	89.0	72.0	82.0	91.0	16	82	8.3
302309	John	52.0	83.0	70.0	89.5	81.0	83.0	0	82	8.4
302310	Mota	81.0	66.0	81.0	90.0	81.0	79.0	16	100	8.5
302311	David	72.5	82.0	83.0	82.0	80.0	72.0	0	100	8.7
302312	Ted	76.0	83.0	87.0	83.0	82.0	78.0	0	100	8.8
302313	Luis	82.0	75.0	90.0	75.0	77.0	82.0	16	100	8.9
302314	Mary	75.0	66.0	84.5	76.0	79.0	88.0	0	100	8.9
302315	Kaka	79.0	84.5	80.0	83.0	82.0	80.0	0	100	9.0
302316	Mill	86.0	84.0	73.0	81.0	83.0	85.0	0	100	9.0
302317	Bowe	82.0	82.0	79.0	81.0	84.0	80.0	16	100	9.1
302318	Billy	87.5	76.0	90.5	77.0	78.0	82.0	16	100	9.1
302319	Chesi	87.0	82.0	81.0	80.0	89.0	92.0	16	100	9.1
302320	Linda	89.0	97.0	94.0	91.0	89.0	93.0	66	100	9.2
302321	Nevil	86.0	89.5	88.0	90.0	87.0	91.0	33	100	9.2
302322	Peny	88.0	85.0	82.0	86.0	85.0	90.0	16	100	9.3
302323	James	87.0	92.0	89.0	86.0	91.0	95.0	50	100	9.4
302324	Mike	90.0	92.0	79.0	86.0	87.0	90.0	50	100	9.5
302325	Cart	87.5	82.0	85.0	87.0	89.0	90.0	16	100	9.6

4.2Visualization analysis

4.2.1Visual programming using IDL

It's easy to find all the education data file meet the conditions by calling file_search function built-in IDL. It is assumed that education data file are named in the form of ".dat" after normalization processed by Hadoop. First, using file_search function can quickly find all standardized registration documents of students, and count the number of files. Then, all subjects score of the students are read into the specified array with cycle approach, and the array data were averaged, summed, or data fitting processed. Finally, use the data processed to draw kinds of graphics for visual analysis by using commands like plot, oplot, contour, surface and so on. The main code is described as in *Table 3*.

Table 3 Function codes of visual programming

```
Function    Datavisual
Input  : Standardized score tables of all the students
         (Class 1-4, Grade1-4)
Output : Lines drawings, contour map, grid surface map,
         shadow surface map
Step1:reading data file
    for s=1,6,1 do begin
    for class=1,4,1 do begin
     openr,lun,file(s),/get_lun ;open file
readf,lun,XXX,  XXX,.....$format='xx'; read file format
                n1[i]=subject1
                n2[i]=subject2
                n3[i]=subject3
                …
          endfor
    endfor
Step2:doing related processing
    Z1=Function1(n1,n2,n3,...)  ;calling function1
    Z2=Function2(n1,n2,n3,...)  ;calling function2
    .....
Step3:drawing images
    set_plot,'PS'        ; calling PostCripts
    device,tupian.ps     ; create and open a ps file
     Plot,x,y            ; plot lines drawings
     Oplot,x,y           ; plot overlay lines drawings
     Contour,z,x,y       ; plot contour map
     Surface,z,x,y       ; plot grid surface map
     Shade_surf,z,x,y    ; plot shadow surface map
  device,/close_file     ;save and shut ps file
End
```

4.2.2Visualization analysis of students' score

It is convenient to convert online education data to line drawings, contour map, grid surface map and shadow surface map by using the methods and procedures discussed above. According to these images, we may find some useful rules hidden in the data, which can be hard to see from the data directly.

Line drawing of a separate course: Scatter plot, the line drawing is both drawn by plot function, while overlay line drawing is realized by oplot function. Taking Chinese score, for example, to analyze the change of score with grade and age through *Figure 4* to *Figure 7*, which are drawn from the score of 9 students randomly selected from every class and every grade. In all of the four figures, black, red, blue, Green respectively represents the class 1, class 2, class 3 and class 4 and Special symbols: ＋、※、－, △, ◇, □ respectively represents grade 1, grade 2, grade 3, grade 4, grade 5 and grade 6. In *Figure 4*, the score distribution with age can be seen, but hard to find valuable information contained in data. In *Figure 5*, it is also difficult to find rules in data, though it can see the vary trends between score and age. *Figure 6* and *Figure 7* were obtained by equal pitch translation on age and score of *Figure 5*. In *Figure 6*, which corresponding to the true age but score equally spaced, it can be more intuitive to observe the trend of change in each grade and each class. In *Figure 7*, which is more clearly than *Figure 6*, for score and age were both equally spaced, it can give us lots of useful information.

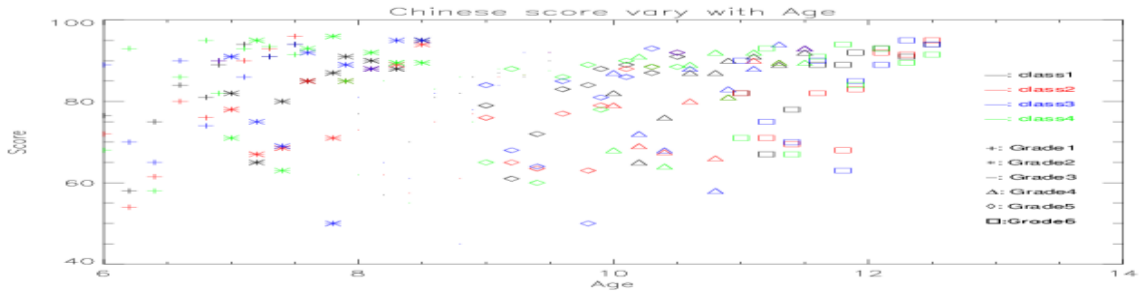

Figure 4 Overlay scatter plot

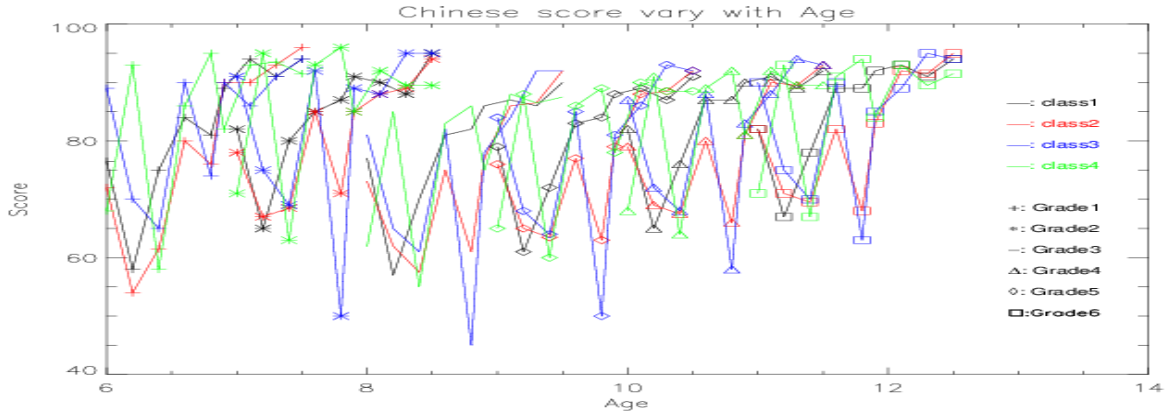

Figure 5 Overlay lines drawing

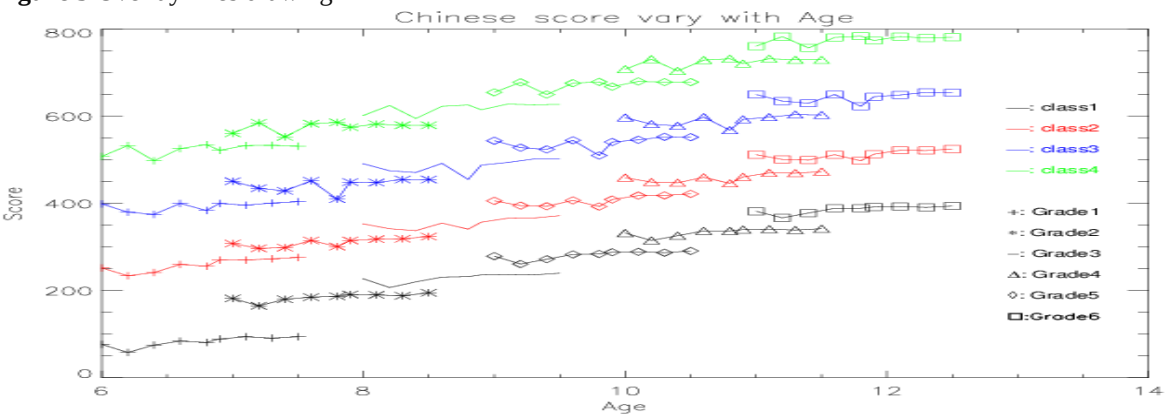

Figure 6 Overlay lines drawing (score spaced)

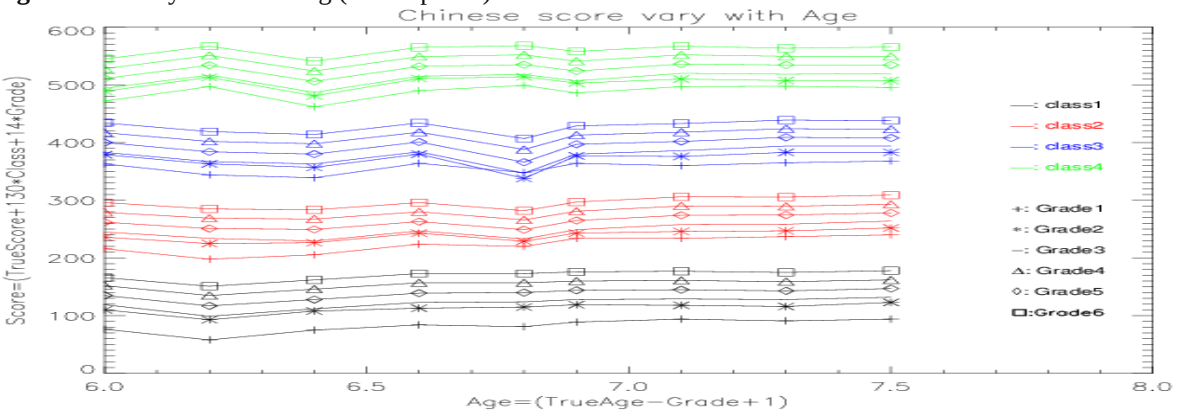

Figure 7 Overlay lines drawing (score, age spaced)

Analyzing the trend of the curve that represents a Chinese score of Class 1, which tagged with "＋" in *Figure 7*, we can see: students' scores fluctuate greatly with age before 6.9, and the change tends to be gentle after 6.9. This phenomenon can also be found if the similar process on the other subject, which can indicate that the 7-year old is more suitable for children beginning to study in school. Comparing the distance between the grades of the four classes in *Figure 7*, it can be seen that the distance between grade 2 and grade3 tagged by "※" and "－" is very close, that also means the score descend obviously from grade 2 to grade 3 , indicating that may have a relationship with the difficulty of the course raising in grade 3. At the same time, we can find that the distance between grade 2 and grade 3 increased with age increased, which can manifest the older students are more adaptable to environmental change and less effected on the difficulty of the course. Observing the distance

between classes in grade 4, grade 5 and grade 6, it also can see that as the growth of the age, the students' ability to adapt gradually strengthen, the age difference will gradually weaken.

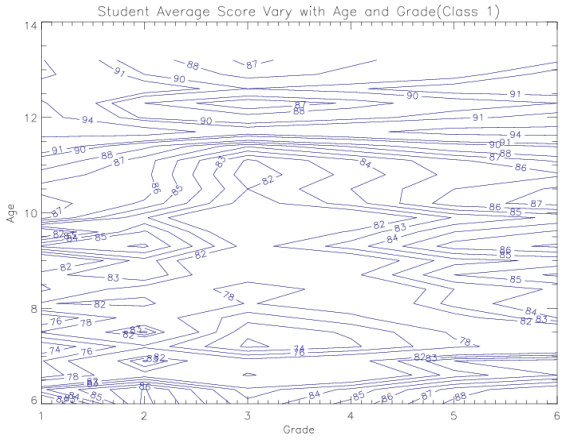

Figure 8 Contour map of average score (class 1)

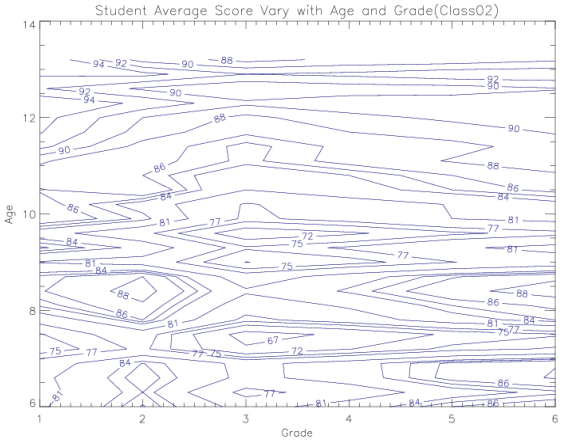

Figure 9 Contour map of average score (class 2)

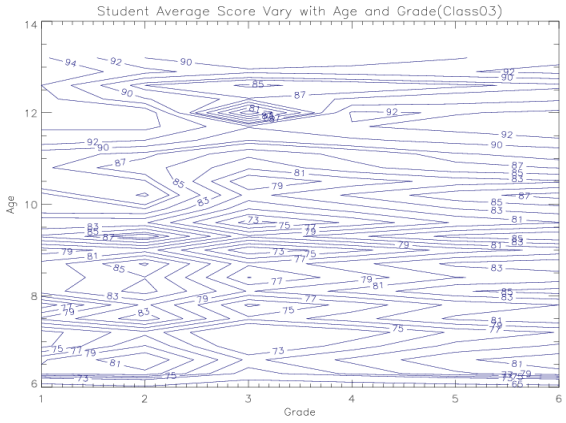

Figure 10 Contour map of average score (class 3)

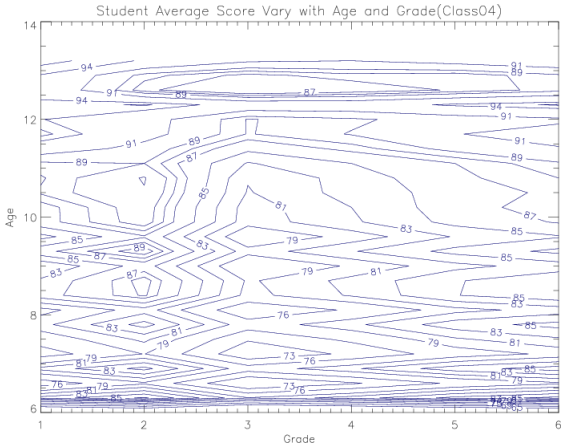

Figure 11 Contour map of average score (class 4)

1) Contour maps of average score

The contour maps are drawn by using the contour function according to the average scores with grade and age of the four classes, as shown in *Figure 8* to *Figure11*. By connecting the points of the same score of each age and grade, the relationship between the average score of the students and the age and grade is displayed in different levels. It can be seen from the contour maps of the four classes: In grade 1 and grade 2, the average score of students increases with age, indicating that students are gradually adapting to courses; In grade 3, the average score is falling to the bottom, almost no more than 90 points, indicating that as the difficulty of the course grows, students of all the ages become a certain degree of not adaptation, but the impact on the older is relatively small; in grade 4, grade 5 and grade 6, the average score gradually increased all the ages, and rate of increase with age is getting smaller and smaller. The results of analyzing show that the overall quality of students is growing with age. But as the difficulty of the course grows, it is increasingly difficult to obtain high scores.

2) Grid (shadow) surface maps of average score

Comparing to line drawing and contour maps, grid (shadow) surface maps are more intuitive. The grid surface maps (*Figure12, 14, 16, and 18*) and shadow surface maps (*Figure13, 15, 17, and 19*) are drawn by using surface function and shade_surf function. It can clearly see the change context of average scores with the growth of age and grade from grid surface maps of the four classes. Observing the changes of the top of the surface, we can find: in grade 1 and grade 2, the surface changes are very complex, indicating that students are in a stage of adaptation; in grade 2, there is a small raised, indicating that after the adjustment stage, students generally adapt to the

study life, and the average score is steadily rising; in grade 3, the deepest groove of the surface appears on the top, indicating the difficulty of courses increased, the average scores are generally low in all ages and grades; in grade 4, grade 5 and grade 6, the surface of all the classes tend to be gentle , in addition to class 1 with a deep groove, indicating that the difference between average score and age has gradually weakened. The polarization is more serious in class 1 that may be related to learning methods or teacher approach chosen by student themselves. These laws can also be seen from the shadow surface maps of the four classes. The colour depth of the shadow surface shows the different height and orientation of the surface, and it is more convenient to observe, especially when the data to be analyzed is not smooth enough, the shadow surface map analysis is more effective.

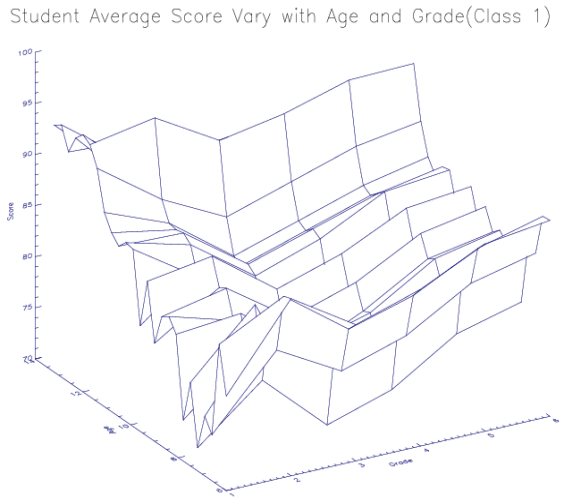

Figure12 Grid surface map of average score (class 1)

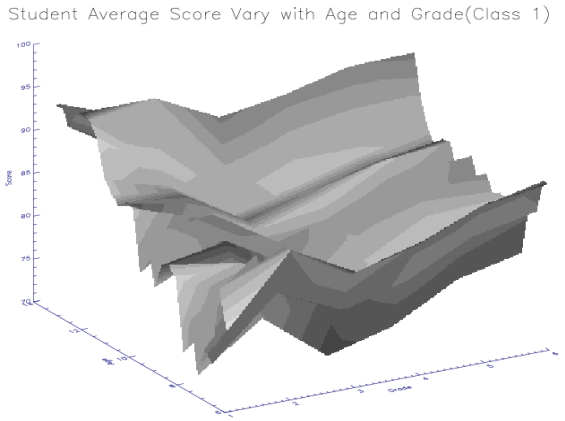

Figure13 Shadow surface map of average score (class 1)

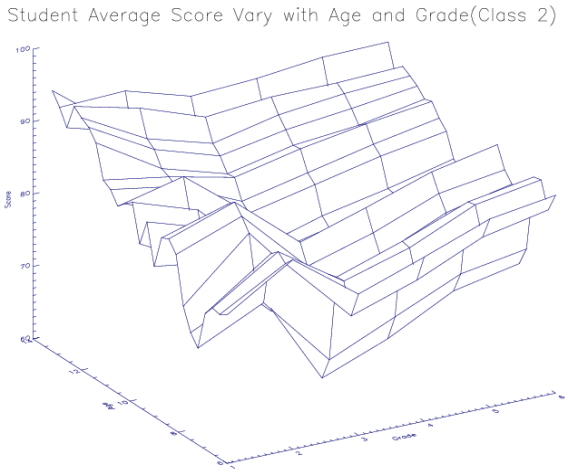

Figure14 Grid surface map of average score (class 2)

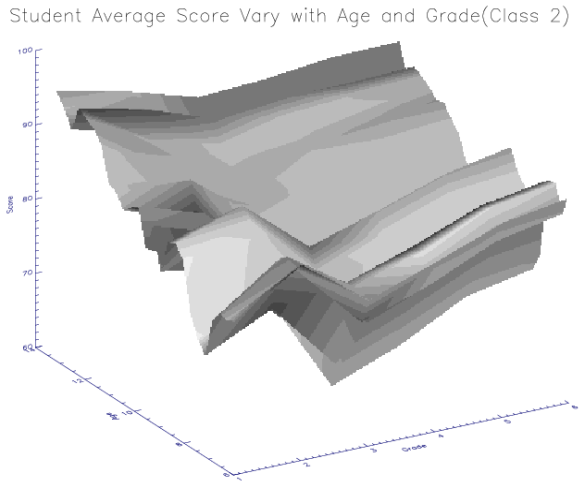

Figure15 Shadow surface map of average score (class 2)

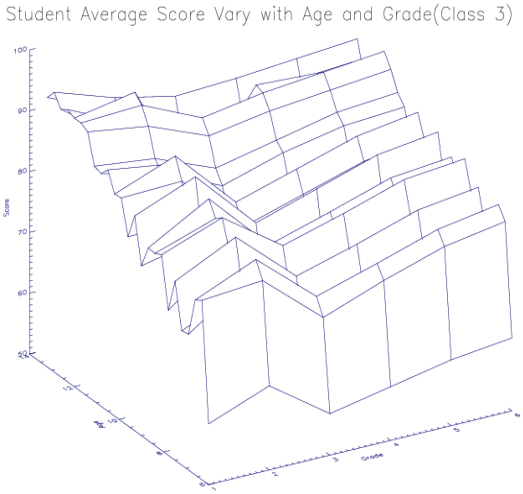

Figure16 Grid surface map of average score (class 3)

Figure17 Shadow surface map of average score (class 3)

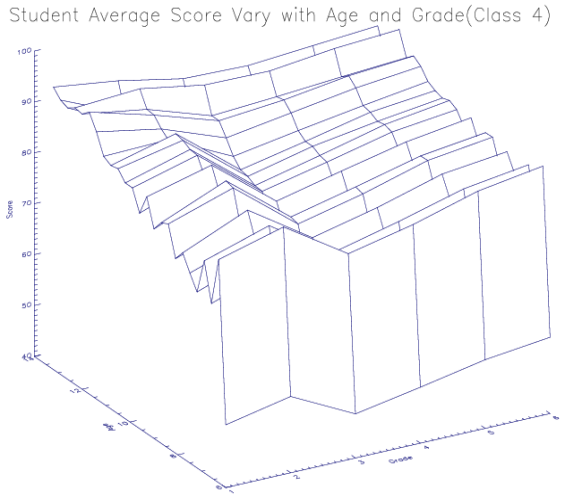

Figure18 Grid surface map of average score (class 4)

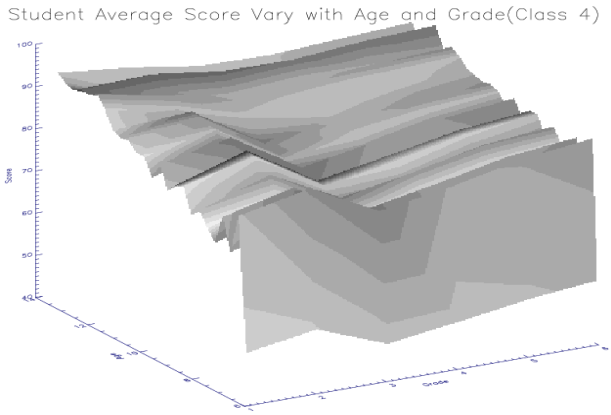

Figure19 Shadow surface map of average score (class 4)

5.Conclusion

Mining useful information from massive online education data is the inevitable trend of future smart campus construction. It can be very convenient to do effective visual analysis on online education data, by giving full play to the distributed cluster system, Hadoop's parallel processing function and IDL's powerful graphical image display function. Transforming the huge data sets that are messy into images which are more intuitive, such as line drawing, contour maps, grid surfaces, histograms, bar charts, and so on, allows us to dig more important information contained in educational data, provides a strong follow to improve learning methods, rich teaching approach, boost the quality of education. Although in this case the data from the experiment is just the tip of the iceberg and student's score is only a part of the sampling. We can also find some laws hidden behind the data by doing a visual analysis of the limited data sets. Especially in the Information era with high-speed network development, these data are all-round full-time, online education data not only reflects the students' personalized learning situation, but also reflects the effective use of educational platform, as well as the implementation of the national education system. Meanwhile, we should know that preparing standardized data file is the base of visualization. There is a long way for us to go on how to do more effectively data standardized processing and we need to focus on how to regulate a variety of education, teaching and learning software from the source in the next research process.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] http://moodle.sgu.edu.cn/moodle/pluginfile.php/10859/mod_resource/content/0/zt1/_2011-2020_.pdf. Accessed 20 February 2017.

[2] Michalik P, Stofa J, Zolotova I. Concept definition for Big Data architecture in the education system. In international symposium on IEEE applied machine intelligence and informatics 2014 (pp. 331-4). IEEE.

[3] Picciano AG. Big data and learning analytics in blended learning environments: benefits and concerns. International Journal of Interactive Multimedia and Artificial Intelligence. 2014; 2(7):35-43.

[4] Self RJ. Governance strategies for the cloud, big data, and other technologies in education. In international conference on utility and cloud computing 2014 (pp. 630-5). IEEE.

[5] Williamson B. Digital education governance: data visualization, predictive analytics, and 'real-time'policy instruments. Journal of Education Policy. 2016;31(2):123-41.
[6] Xiang F, Ming-hua Y, Xiao-ling M, Yong-h W. Learning analysis system architecture based on big data technologies. Journal of East China Normal University : Natural Science. 2014; 31(2):20-29.
[7] Chernoff H. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association. 1973; 68(342):361-8.
[8] Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In proceedings of the 1st conference on visualization 1990 (pp. 361-78). IEEE Computer Society Press.
[9] Swayne DF, Cook D, Buja A. XGobi: Interactive dynamic data visualization in the X Window System. Journal of Computational and Graphical Statistics. 1998; 7(1):113-30.
[10] Huang TH, Huang ML, Nguyen QV, Zhao L, Huang W, Chen J. A space-filling multidimensional visualization (SFMDVis) for exploratory data analysis. Information Sciences. 2015; 390: 32-53.
[11] Molina-Solana M, Birch D, Guo YK. Improving data exploration in graphs with fuzzy logic and large-scale visualisation. Applied Soft Computing. 2017; 53: 227–35.
[12] Dianwu Y. Visualization tools IDL entry and improve. Machinery Industry Press; 2003.
[13] Donglai F. Hadoop massive data processing: technical details and project combat. Posts and Telecom Press; 2016.
[14] Ding ZB, Yuan F, Dong HW. Application of data mining to analysis of university students' grades. Computer Engineering and Design. 2006; 27(4):590-2.
[15] Turkington G. Hadoop beginner's guide. Packt Publishing Ltd; 2013.
[16] Jain A, Bhatnagar V. Crime data analysis using pig with hadoop. Procedia Computer Science. 2016; 78:571-8.
[17] KV R S, Kavya N P. Trend analysis of E-Commerce data using Hadoop ecosystem[J]. International Journal of Computer Applications, 2016;147(6):1-5.

Yu Lasheng, Vice professor in Central South University of China, ACM and CCF member, ACM/ICPC golden medal coach. He received the B.Sc. degree in Computer Science, the Master degree and a Ph.D. degree in Control Theory and Control Engineering from Central South University. He is the editor of the Journal of Convergence Information Technology and Advances in Information Sciences and Service Sciences, etc., he is also the reviewer for the journals such as Future Generation Computer Systems, journal of Parallel and Distributed Computing, Artificial Intelligence Review, etc. He has published at least 70 papers on Agent technologies or Algorithms, and has published 3 books. He has organized and implemented many projects which have created great achievements in the society. His main research interests include agent technologies and applications, structure and algorithm and smart computing etc.

Wu Xu, received a Bachelor's degree in Electronic Information Engineering from Wuhan University, Wuhan, Hubei Province, China, in 2006. At present is a Master graduate of Central South University in China, majoring in computer science and technology. His research interests include machine learning, visulization and big data.
Email:187353898@qq.com

Yang Yu, received a Bachelor's degree in computer science and technology from Central South University ,Changsha, Hunan Province, China, in 2014. At present is a Master graduate of Central South University in China, majoring in computer science and technology. His research interests include machine learning, recommendation and algorithm design.