Big Data Analysis in a Social Learning Platform

Lan Huang[1,a], Yunfeng Wei[2,b], Alessio Zamboni[3,c], Jing Zhang[2,d] , Hao Xu[1,4*,e]

[1]School of Computer Science and Technology, Jilin University, Changchun 130012, China ;

[2]School of Software, Jilin University, Changchun 130012, China ;

[3]School of information engineering and computer science DISI, University of Trento, Via Sommarive 9 I-38123, Trento, Italy;

[a]huanglan@jlu.edu.cn, [b]zealer_zz@126.com, [c]hannialessio@gmail.com,[d]zhangjing_cc@126.com,[e] xuhao@jlu.edu.cn

*Corresponding Author

Keywords: Big Data Analysis; Social Learning Platform; Hadoop; MapReduce; R Programming.

Abstract. As the amount of networking information production keeps growing, big data analysis has been widely used in the field of education. In recent years, teacher-student interactions on the Web are increasing, this will generate a lot of data. The use of big data analysis methods to analyze these data has become a popular research lately. In this paper, we use Hadoop big data analytic platform to analyze the log data of Jilin University Website, which is used to analyze these data to find out during when students surf the Internet most so that teachers can better communicate with students during these periods. By doing so, these data has not only been put to good use, but can also make teacher-student interactions more effective.

Introduction

In the big data era of explosive data growth, big data analysis in the field of education has always been under great concern. The frequent teacher-student interactions on the web will generate a lot of data, through the analysis of which[1] some conclusions can provide teachers with decision-making methods, some can help teachers better understand the students, therefore, analysis of these data is very meaningful.

The examples revealed in this paper make use of Hadoop big data analysis techniques to analyze the log data on the website of Jilin University, where the function of Q&As between teachers and students[2] can effectively analyze when students browse the Website so that teachers can choose a better time to communicate with them. Using the R language to display the data can clearly show when students login in and provide more accurate information for teachers, which will undoubtedly make the communication between teachers and students more effective, saving time for both teachers and students.

The paper is organized as follows. In the next section, we will first introduce the use of related technologies in this paper, such as Hadoop big data analysis and one of its main computing framework MapReduce, both of which are now the focus of attention. In the third quarter we will mainly describe the specific analysis of this example, and the results of data analysis revealed by using the R language; in the fourth quarter, we will mainly describe what the conclusions based on the results are, as well as how schools and teachers can better arrange the curriculums on this basis; Finally, my future plans in the relevant fields.

Related technologies

Hadoop. Big Data analysis tool used here is Hadoop, whose development is based on Java language;and is used because its large data processing applications, good off-line data analysis, a distributed file system and a Map-Reduce computing framework structure; compared with the traditional method of analysis, its speed improves a lot. The computing nodes are divided into a name

node and several data nodes, the former is for management, control, and maintenance of data nodes, while the latter is used for storing data.

Mapreduce. MapReduce is a programming model for data processing, and its process can be divided into three parts: Map function, primarily used for decomposition of parallel tasks; Combine process, mainly used to improve the efficiency of Reduce; and Reduce functions, the post-treatment of results[3]. A variety of languages by MapReduce programs can run on Hadoop, primarily the development of Java-based applications, it has the advantage of processing large data sets.

Analysis process

Big Data analysis techniques are widely used in various industries, its processes of analysis are basically the same, firstly, obtaining of data；there are many methods for data acquisition, data can be obtained via the back-end server, you can also use Python to obtain data online; secondly, MapReduce programming, the compiling of Mapreduce program is mainly based on the Java language, and can also be compiled by other languages; thirdly, running of the written MapReduce programs on Hadoop big data analytics platform again, it is constituted by a number of nodes cluster; Finally, the demonstration of analyzed results via graphics or other ways, which can provide supporting data for further decisions or applications; the following is the specific process of analysis:

Step1: obtaining the source data from the backend server of the website, the source of data is for all the students who have access to it, the data form the backend server of the website are stored in the form of logs under a fixed route, when students click or browse a certain link on the website, there will be corresponding records on the backend server, thus a large number of log data will be created. Generally, log data has a fixed format, and will be better processed by using Hadoop cluster.

Step2: Programming Hadoop's MapReduce by using the Java language, first, write Map and Reduce functions, Map function can cut out the date and hour part in the source data and calculate through a simple program what day the certain date is, they will be saved in the Key -Value pairs, i.e. (DayofWeek, 1) and (HourofDay, 1), then carry out Combine process, the result is (DayofWeek, 1,1,1,1 ...) and (HourofDay, 1,1,1,1 ...), and finally the Reduce function, will result in a sum of two, resulting in a final data, i.e. (DayofWeek, m) and (HourofDay, n) (m, n is a positive integer). The running process of what hits per hour is shown in Figure 1.

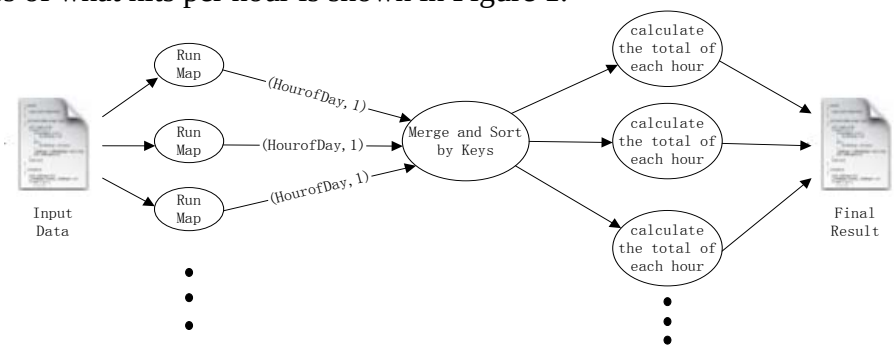

Fig.1 Process of analysis

Step3: Run this example on a Hadoop cluster, the operating environment is Hadoop cluster, the cluster has three virtual points, one of them is the name node, the other two are data nodes, the system is the Unbutu system of linux and the tool for running Java code is Eclipse. Get the results of analysis after a short time, the results will be stored in the document to get ready for the next data show.

Step4:According to the results obtained from step 3, using the R language in the form of a line chart for data display, R language cannot only provide the data analysis, but the data display also has a natural advantage, you can take advantage of the trend of different charts[4], and presentation of data dynamic, as used herein, is the R language line chart, which can be a good use of data to show the ups and downs, the trend, the highest and the lowest point.

Figure 2 shows the students' views of the website within a week, it can be clearly seen in the figure that views of the site are relatively stable from Monday to Friday, which were about 39,000

times, but on the weekend, the students' views of the site were relatively fewer, with less than 28 000 times on average,in Figure 2.

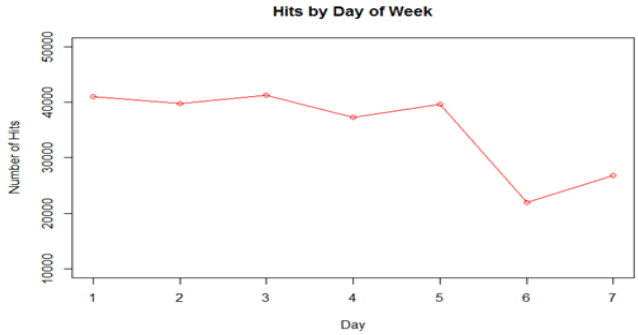

Fig.2 Views of every week

After analyzing views of every week, now let's take a look at the views of each day, which can be divided into two cases, one is views of every day from Monday to Friday, and the other is views of the weekends. First, take a look at the daily views from Monday to Friday, in Figure 3.

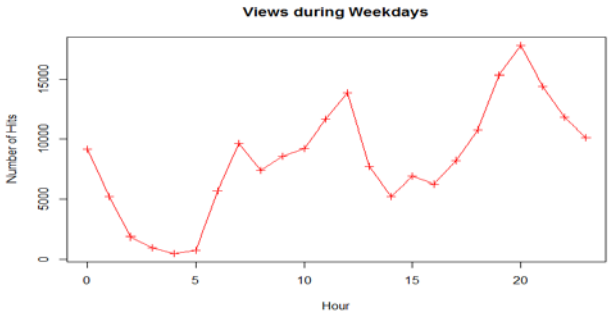

Fig.3 Views from Monday to Friday

It is shown in Figure 3 the views of every hour during each day from Monday to Friday, you can see there is a big difference in the views of different periods, it is because basically, in the early morning, the students are resting, views are very few; but from 8: 00 to 10: 00 and from 13: 00 to 17: 00, views are not that many, either, because some students are having classes, which affects the numbers of visitors; whereas from 19: 00 to 21:00, the numbers of visitors reach the top, with over 14,000 views.

However, the results are completely different during some periods on the weekends, the peak of views is from 15: 00 to 16: 00, which can reach about 4000, but compared with the one from Monday to Friday, there is still a difference, which, of course, is an inevitable result owing to a certain decrease of students' concern over the website on weekends. The fewest views of the day also come in the early morning, and there are no major fluctuations during other time points, when views basically remain at about 2000-3000, as is shown in Figure 4.

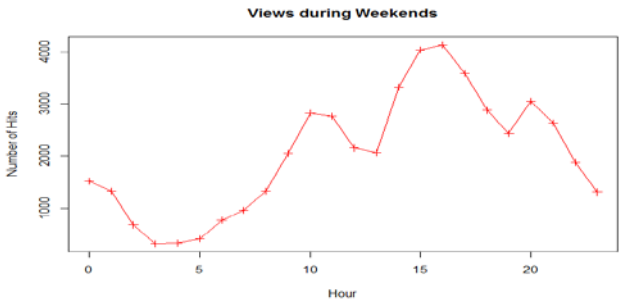

Fig.4 Views on Weekends

Discussion and Conclusion

Through the analysis above, we can better reflect the time of students' access to the website, like which day of a week and what time during a day. Through the analyzed results above[5], we can draw a time table of the highest and lowest views of every week, as is shown in Table 1.

Table 1 Comprehensive analysis of the results

Time & P/T	Weekdays	Weekends
Peak	19:00-21:00	15:00-16:00
Trough	03:00-05:00	03:00-06:00

According to the table above, teachers can know more clearly about when the most and least views from Monday to Friday and from Saturday to Sunday come, which helps teachers better adjust the periods of teacher-student communication. Accordingly, on the one hand, teachers can try to communicate with students between 19: 00 and 21: 00 from Monday to Friday, and between 15: 00 and 16: 00 on weekends. On the other hand, try to avoid periods during which views are relatively fewer. Thus, it not only saves time for both teachers and students, but also makes interactions between them more efficient.

Future Work

The analysis above is somewhat representative, it can not only help teachers better choose the period of time to communicate with students[6], but can also make good use of the data in these fields of education. Of course, other than analyze those conclusions; these data can also help you get more interesting things. I hope in our future work, we can make use of more technology or tools to dig out more valuable conclusions which can be better applied to the field of education to help schools and teachers make appropriate decisions, and furthermore, to help teachers and students to communicate more effectively.

Acknowledgements

This work is supported by the National Natural Science Fund Project of China (61300147, 61472159), Electronic Commerce Engineering Laboratory Project of Jilin Province (2014N143), the Science Technology Development Projects of Jilin Province (20121805, 20140101180JC), China Postdoctoral Science Foundation (2014M551185) and the Science Technology Project of Changchun (14GH014).

Reference

[1] Hao Xu, Chang-hai Zhang, Yu-an Tan, Jun Lu, An improved evolutionary approach to the Extended Capacitated Arc Routing Problem. Expert Systems with Applications, 38(4): 4637-4641,2011

[2] Jin Xiong,Yiming Hu,Guojie Li,Rongfeng Tang,Zhihua Fan, Metadata Distribution and Consistency Techniques for Large-Scale Cluster File Systems. IEEE Transactions on Parallel and Distributed Systems,2011

[3] Yang Lai,Shi ZhongZhi, An Efficient Data Ming Framework on Hadoop using Java Persistence API. 2010 10th IEEE International Conference on Computer and Information Technology (CIT .2010) .

[4] Hao Xu, Yue Zhao, Li-ning Xing, The Novel Heuristic for Data Transmission Dynamic Scheduling Problems. Journal of Applied Mathematics, Volume 2013.

[5] Das S,Sismanis Y,Beyer K S,Gemulla R,Haas P J,McPherson J.Ricardo, Integrating Rand Hadoop. Proceed-ings of the ACM SIGMOD International Conference onManagement of Data,2010.