USING HIGH PERFORMANCE COMPUTING AND OPEN SOURCE TECHNOLOGIES FOR SOLVING BEHAVIOUR ANALYTICS PROBLEMS IN E-LEARNING

Laimonis Zacs

Anita Jansone
Liepaja University, Latvia

Abstract. *In this paper the authors describe solution for solving various analytical problems in E-learning, Course Management Systems like Moodle by using HPC (High Performance Computing) and Apache Hadoop open source technologies in Liepaja University. The problem is that nowadays there are collecting huge amounts of analytics data from several gigabytes to petabytes, which is hard to store, process, analyse and visualize. This article reflects one of the solutions concerning distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that can store and process the data, can scale without limits and provides technological opportunities of reliable, scalable and distributed computing.*

Keywords: *Apache Hadoop, Big Data, E-Learning technologies, High Performance Computing, online learning platform, open-source software.*

Introduction

Nowadays the count of internet users and the role of internet appliance in the studies rapidly increase. At the moment the Internet is the first source, where students search materials for their researches and theses etc. The large part of students use the Internet as compulsory tool for studies at school, as well as for attending online libraries and conferences. Therefore, the number of those students increases, who use the Internet as the primary technical auxiliary tool in order to search and use the necessary information. Therefore, it is possible to say that it is very important to use technological solutions and software tools, in order to summarize and collect information, which is necessary for knowledge acquisition and to provide effective knowledge transfer between teachers and students at any university. As one of the solutions it is necessary to mention application of E-Learning education methods and means in the education process of higher schools.

The term “e-learning” was defined in alignment with a definition by Rosenberg (Rosenberg, 2001). According to Rosenberg, the first and the most important feature of e-learning is that it takes place in the network environment. This means that computer of the learner is in constant communication with the central server. Also e-learning materials are accessible via an Internet browser on a personal computer (Ninoriya et al., 2011). So in our opinion, E-Learning is online process in a network where students use computer equipment with

http://dx.doi.org/10.17770/sie2015vol4.405

installed Internet browser to connect to the central education server to get online E-Learning materials and collaborate with educator.

E-learning is a method of education which utilizes a wide spectrum of technologies in the learning process, mainly the Internet or is computer-based. It is naturally related to distance learning, but nowadays it is commonly used to support face-to-face learning as well. *Learning Management Systems* (LMS) provides effective maintenance of particular courses and facilitate communication within the student community and between educators and students (Drazdilova et al., 2010).

Nowadays, the major part of the educational centres (universities, institutes, colleges and schools) are use some e-Learning tools as an integral part of their learning systems; to enhance their traditional learning systems or to use an alternative approach for virtual learning environment. These tools may be based on content management or learning content management (Rosenberg, 2011).

Distance learning education process in Liepaja University have used certain technological solutions, such as Blackboard Learning System since 2005.

BlackBoard Inc. *Virtual Learning Environment* (VLE) is an online web server software, which provides course management, customizable open architecture, and scalable design that allows educators to communicate with students through information systems with authentication.

The alternative to Blackboard Learning System solution is Moodle online learning platform, which has been used at Liepaja University from 2010. Moodle has proven to be an immensely popular and important tool in distance education. One feature provided by Moodle is a rich source of information about student access to online materials. The open-source software learning platform is focused on managing students, tracking progress and delivering online courses, not authoring the contents. It's expected by the year 2020 that 50% of all university courses will be provided and delivered online.

Moodle, the popular *Virtual Learning Environment* (VLE) or, using the term *Course Management System* (CMS), is a flexible open-source software and online learning platform. Moodle is currently one of the most popular open-source course management systems in online education. Some evaluations have also indicated that Moodle is one of the top-rated programs when compared to other open-source course management systems (Graf&List, 2005). Moodle's unique focus on pedagogy allows online learning to cross over from the traditional educational area of factual recall and memorization into the area of social networking (Rogers et al., 2009).

By 5 March 2015:

- Moodle had a user-base of 54,049 registered sites with 71,509,781 users in 7,699,073 courses in 226 countries (Moodle.net, 2015).
- In Latvia there are 75 registered Moodle websites totally and Liepaja University Moodle site is one of them with 2500 users in total, 500

new users are added every year, approximately 300 users are active online every month.

With log data growing so rapidly and rise of structured and unstructured data today, it is necessary to find effective technological solutions of data storage, management and analysis Legacy systems will remain necessary for specific high-value, low-volume workloads, and contribute to the use of Hadoop ecosystem, optimizing the data management structure in Liepaja University by putting the right *Big Data* workloads in the right systems. Hadoop can handle all types of data from disparate systems: structured and unstructured data, log files, pictures, audio files, communications records, emails – just about anything you can think of, regardless of its native format. Even when different types of data are stored in unrelated systems, you can dump it all into your Hadoop cluster with no prior need for a schema. In other words, you do not need to know how you intend to query your data before you store it; Hadoop lets you decide later and over time it can reveal questions you never even thought to ask. Apache Hadoop is an open-source software, which pioneered a fundamentally new way of storing and processing data. With Hadoop, no data is too big. And in modern hyper-connected world where more and more data are created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless (Cloudera, 2015).

The aim of this study is to find potential solutions of high performance computing technologies together with data warehousing and large storage database system tools, applications of cluster networks and the modern information technologies in order to promote massive behavioural statistics data storing, processing, analysing and visualizing.

The research object of this research is to explore various data mining and analytics tools for analysing students' behaviour in the distance learning process. *Course Management Systems* like Moodle platform are successful e-learning tools to provide students access to courses in distance learning. E-learning systems store large amount of data based on the history of user interactions with the system, so we explore advanced *HPC* (High Performance Computing) and Apache Hadoop open source technologies for collecting huge amounts of analytics data from several gigabytes to petabytes in future.

Objectives of the research are: to analyse pedagogical and technical literature as well as articles on computer science that evaluate e-learning process; to study the existing behavioural user activity log data mining and analysing tools and implementation thereof in Liepaja University; to define effective data storage and also reliable, scalable and distributed computing in distance learning; to identify the performance tools of behavioural user activity log data mining and analysing tools.

Expected results is anticipated that implementation of *HPC* and Apache Hadoop open source technologies in E-Learning Moodle platform will significantly contribute to improvement of the efficiency and also will provide

unlimited storage about registered users activities and behaviours data as well as will provide high performance statistics analysis and visualisation.

According of this research results the next step will be necessary to develop a solution based on industry-standard servers that can store and process data, scale data without limits and provide technological opportunities of reliable, scalable and distributed computing with *HPC* and Apache Hadoop open source technologies and implement them in the Learning Moodle platform.

General position

The tables are filled with data, using various statistics tools, which can be obtained about Moodle user activities: Moodle statistics, AWstats, MooDog, Gismo and technical implementation thereof.

In this paper, we examine behaviours of Moodle data statistics from Liepaja University, using various technological tools. When students interact so intensively with Moodle platform, the question arises: *How huge amounts of analytics data is it possible to transfer/deliver to provide fast data storing, processing, analysing and effective management?*Moodle system record users behaviours in the log files. There is a great amount of data stored in the Moodle system about the activities of users, both educators and students. The stored activity information typically is „who, what, where, when was doing". Moodle records each action within the *VLE*, it registers each user who initiated that action, the time of initiation and location. The recorded data column refers to all the data that have been recorded in the log file. A typical example of user behaviour information from a Moodle log can be extracted with such options:

1) Total Page views - records the number of times any page within the course space was accessed.
2) Total unique users - number of unique users that have accessed the course space at least once.
3) Total unique actions - number of unique actions that have been carried out in the course space at least once and the nature of each action depends on the context of the activity or resource used. Typical Moodle actions are View, Add, Delete etc.
4) Total unique pages - number of pages that make up the structure of a course space.
5) Total IP addresses - unique IP addresses which have been recorded to have accessed the course space at least once.
6) Mean session length - refers to the average amount of time that users spend inside the course space. It should be noted that Moodle log files do not record the login and log off data of users. Therefore, this metric is an approximation and is based on the automatic log off time, as set by the administrators. This metric is an approximation, it is perhaps

more meaningful to focus on its fluctuations over time, rather than on the value of the metric itself.

The idea of analysing student behaviour in Moodle is implemented with other middle-ware tools as well:

1) GISMO is a graphical interactive monitoring tool that provides useful visualization of students' activities in online courses to instructors. With GISMO instructors can examine various aspects of distance students, such as the attendance to courses, reading of materials, submission of assignments. Users of the popular learning management system Moodle may benefit from GISMO for their teaching activities. With respect to the standard reports provided by Moodle (which basically allow teachers to see if an individual student has viewed a specific resource or participated on a specific activity on a specific day), GISMO provides comprehensive visualizations that give an overview of the whole class, not only a specific student or a particular resource. With GISMO, instructors can perform analysis of the whole class, and may have a „clear picture" of what the class is doing, or has done within a period in the past (GISMO, 2015).

The GISMO project (Mazza and Milani) offers visualisations of various statistics in the Moodle log files, and performs little automatic analysis (Mazza & Milani, 2005).

2) Moodog (Zhang), another Moodle log file analysis tool, performs a similar role with an emphasis on visualisation of data rather than analysis. Moodog is superior to the original Moodle log file facility in several aspects: (1) it provides aggregated and meaningful statistical reports; (2) it visualizes the results, which make comparisons between multiple students much easier; (3) not only does Moodog display the activities that a student has performed, but also identifies the materials a student has not yet viewed; and (4) it has the capability to remind students to view those materials that they have not yet downloaded (Zhang et al., 2007).
3) AWStats log analyser for showing all possible information Moodle's log contains. AWStats is a free powerful tool with multiple features and it is able to generate advanced web, streaming, ftp or mail server statistics, also graphically. It can analyse log files from all major server tools like Apache log files (NCSA combined/XLF/ELF log format or common/CLF log format), WebStar, IIS (W3C log format) and a lot of other web, proxy, wap, streaming servers, mail servers and some ftp servers (AWStats, 2015).

Results

Certain descriptions with figures about registered and non-registered users activities for data measures and other parameters. The data for the analysis below were collected within the period between September 2011 and August 2014 for a number of modules in different full-time courses from Moodle log file data.

Moodle provides a rich set of information about how often and how intensely students interacted with Moodle. Fig. 1 shows all activity statistics for students and educators. The statistics shows six fields of activity data: guest, students, non-editing teacher, educators, course authors, manager and total summary. Fig. 1 shows that the most intensive activities take place within the academic year from September until August in Liepaja University. It can be concluded that the most users activity is beginning of study semesters that is from September to January and from February to August.

Fig. 2 shows the expanded information about general number of users, who are connected to the Moodle system, and expanded statistics of unique users connections is also shown. The number of unique users within these years was changing from 250 to 400 depending on the academic year in Liepaja University.

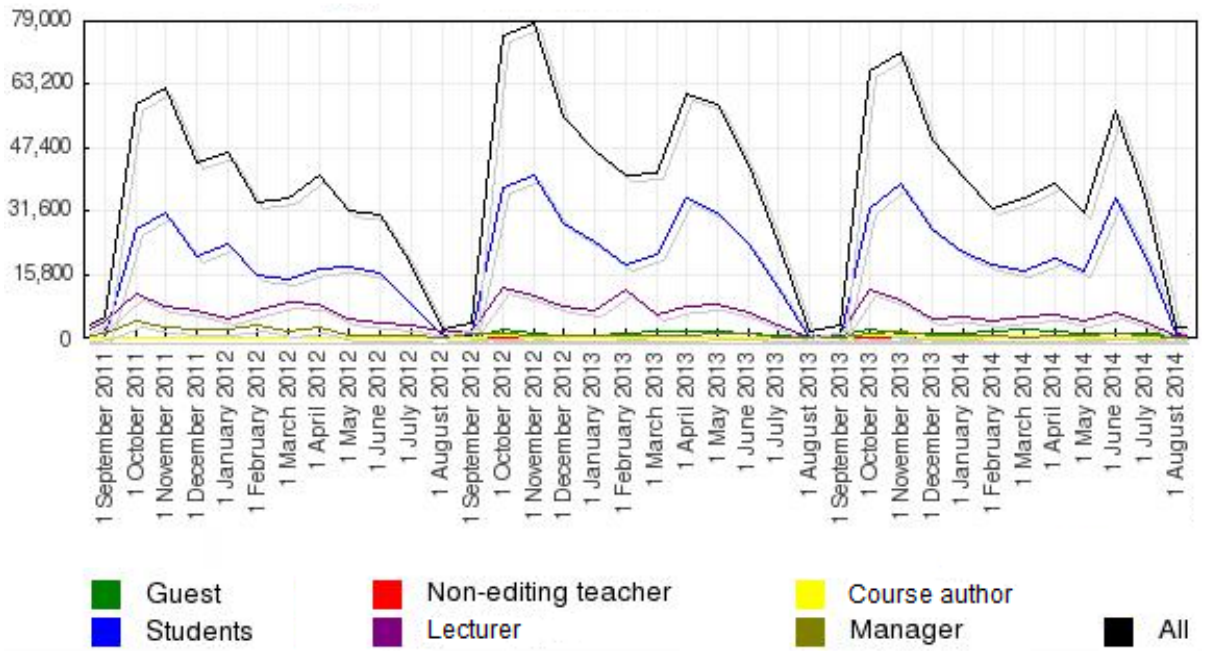


Figure 1. All activity (all roles) as measured by Moodle

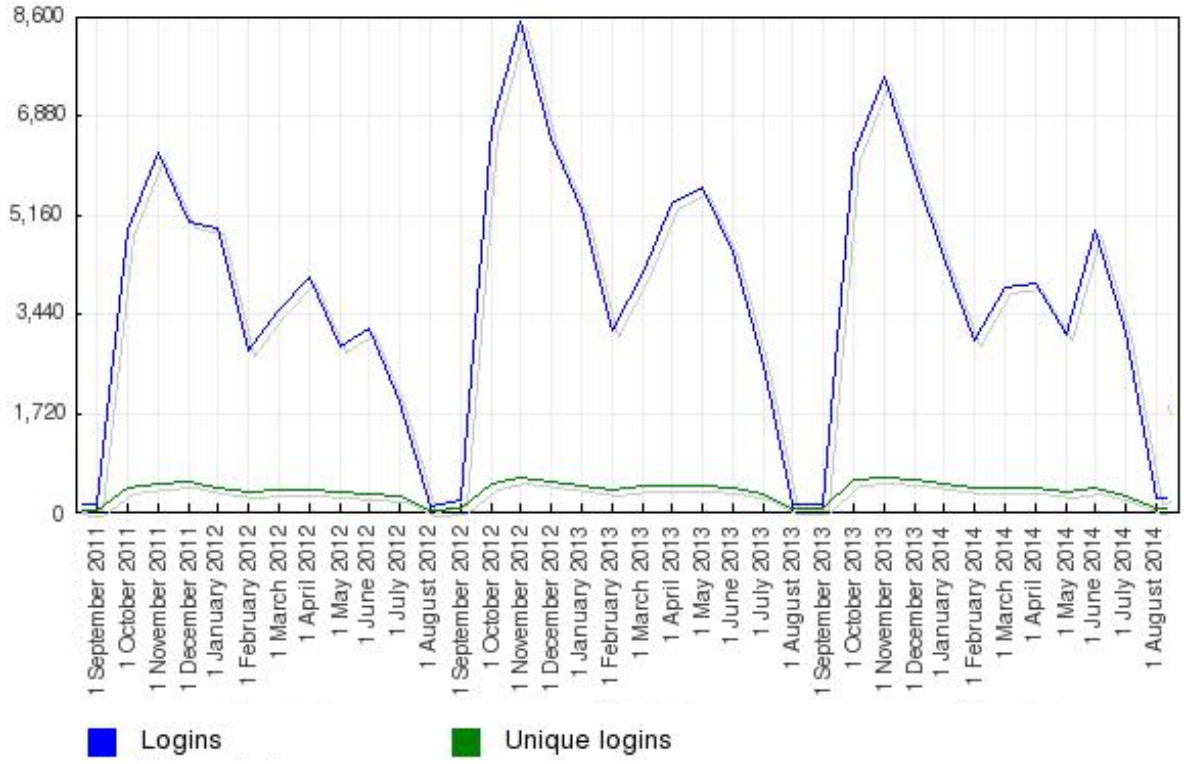


Figure 2. All logins and unique logins activity as measured by Moodle

Additional expanded statistical data can be obtained by using AWStats log file records. Fig. 3 shows the monthly report („Mēnešu atskaite") for the period from January 2014 until December 2014. This report shows unique users per month („Unikālie apmeklētāji"), the number of visits („Vizīšu skaits"), the number of used pages („Lapu skaits"), the hits („Trāpījumi") and volume in bytes („Baiti").

Generalised information about log file data sizes:

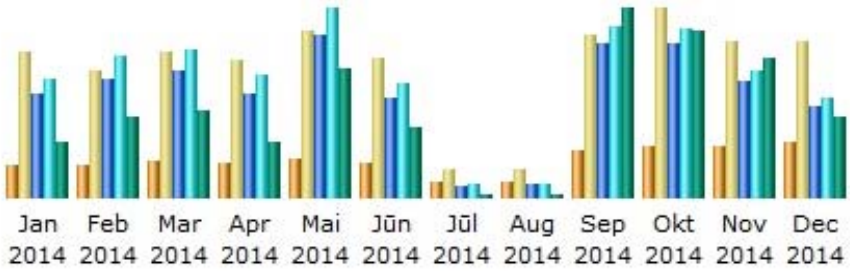
1. Moodle system stores approximately 150 MB log files in average per year with the existing load, see Fig. 1 and Fig. 2.
2. AWStats statistical system stores 50 MB log files in average per month with the existing load, see Fig. 3.

Summarizing the above-mentioned information, in order to compute the size of the all the log files for the period from 2010 until 2015 it is obtained:

1. Moodle statistics data reach 800 MB.
2. AWStats statistics data reach 2500 MB.

Assuming that the number of users can increase by ten times, then the log file size will reach:

1. Moodle yearly statistics data file will reach 1500 MB.
2. AWStats statistics monthly data file will reach 500 MB and 6000 MB in a year.



Mēnesis	Unikālie apmeklētāji	Vizīšu skaits	Lapas	Trāpījumi	Baiti
Jan 2014	949	4164	177916	203732	8.71 GB
Feb 2014	940	3582	201765	240780	12.55 GB
Mar 2014	1049	4161	218937	254817	13.46 GB
Apr 2014	972	3885	176701	209608	8.54 GB
Mai 2014	1125	4741	277374	323339	19.86 GB
Jūn 2014	970	3941	168942	194863	10.81 GB
Jūl 2014	445	809	19053	23037	641.90 MB
Aug 2014	453	785	21788	24464	648.83 MB
Sep 2014	1339	4615	263907	291791	29.26 GB
Okt 2014	1464	5367	263285	290795	25.85 GB
Nov 2014	1464	4457	199585	217662	21.48 GB
Dec 2014	1597	4417	157794	172223	12.44 GB
Kopā	12767	44924	2147047	2447111	164.22 GB

Figure 3. Monthly activity for period from January to December 2014 as measured by AWStats

If such intense activities are anticipated, it can be concluded that in order to collect data for the next ten years, the volume of the log files will reach approximately:

1. Moodle statistics data file for 10 years will reach 15000 MB.
2. AWStats statistics data reach 60000 MB.

A large part of the volume in the Moodle system database is taken by study materials and students works. With the increase of the interest of using the offered functions of Moodle in the study process management, the number of courses and learning materials also increases, therefore the volume of the data base also becomes larger. Large data volume is also created by student works, which are uploaded and collected within the study process. The data volume in Moodle database is increased by the growing number of study works each academic year, because the works from the former years are not usually deleted from databases, but are kept in the same database as archive. In future all of this information could be eligible as to *Big Data* and it could be also necessity providing high throughput access to application data.

Such large volume data are not only to be stored and managed, but also analytical actions should be done in order to acquire the full information about

the online education in Liepaja University. This means that it would be advisable to be prepared for it in advance and introduce hybrid architecture with *HPC* and Hadoop ecosystem technologies for storing the users activities and processing of *Big Data*. As opposed to a standard supported databases for Moodle system, HadoopDB is an open source parallel database capable of performing high speed analytics and *Big Data* management problems as well as combine the scalability of Hadoop with the high performance of relation databases on structured data.

Big Data is the equivalent of HPC, which could also be called high-performance commercial computing or scientific supercomputing. *Big Data* can also solve large computing problems, but it is less about equations and more about discovering patterns. Hadoop enables distributed data processing for *Big Data* applications across a large number of servers. Hadoop cluster runs Hadoop's open source distributed processing software on low-cost commodity computers. Typically one machine in the cluster is designated as the NameNode and another machine as the JobTracker; these are the masters. The rest of the machines in the cluster act as both DataNode and TaskTracker; these are the slaves (Hadoop, 2015). The idea is that distributed, parallel processing will result in redundancy and stronger application performance across clouds to prevent failure. A Hadoop cluster is a special type of computational cluster designed specifically for storing and analysing huge amounts of structured and unstructured data in a distributed computing environment. A Hadoop cluster can be build with various tools and techniques such as StackIQ Warehouse-grade Automation Platform and Rolls like tools. StackIQ Cluster Manager integrated with Hortonworks Data Platform provides a software solution that comes with everything needed to install, configure, deploy and manage Hadoop cluster.

Conclusion and future work

It is necessary work on other researches about moving users activities data to the *HDFS* (Hadoop Distributed File System) and processing with Hadoop ecosystem tools, in order to improve online education processes and management.

The most problematic factor shaping the future of online education is something we cannot actually touch or see that is *Big Data* and analytics.

The main aim of the *Big Data* data analytics is to make decisions in Liepaja University to able to give more effective and thoughtful solutions in respect to the online learning strategy, allowing users behavioural analytics to data scientists and predictive modellers to analyse large volumes of transaction data, as well as other forms and types of data, therefore for managing large volumes of both structured and unstructured data. That could include Moodle system platform and Web server log files, Web site click-stream data, social media

content and social network activity reports, text from educators and students emails and survey responses and so on.

Big Data can be analysed with using *HPC* and various open-source software tools commonly used as part of advanced analytics disciplines such as data mining, predictive analytics, text analytics and statistical analysis.

Future work of our project will focus on three areas: a) Moodle system integration with advanced *HPC* technologies, b) Hadoop ecosystem using in E-Learning c) intelligent agents development to provide effective behavioural analytics.

Acknowledgements

This text has been elaborated within the framework of the ESF funded project „Development of the Doctoral studies at Liepaja University" (Agreement No. 2009/0127/1DP/1.1.2.1.2./09/IPIA/SEDA/018).

References

AWStats (2015). *What is AWStats.* Retrieved from http://www.awstats.org/

Cloudera (2015). *Hadoop and Big Data.* The Platform for Big Data and the Leading Solution for Apahe Hadoop in the Enterprise – Cloud Retrieved from http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html

Drazdilova P., Obadi G., Slaninova K., Al-Dubaee S., Martinovic J., and Snasel V. (2010). Computational intelligence methods for data analysis and mining of elearning activities. In F. Xhafa, S. Caballe, A. Abraham, T. Daradoumis, and J. Perez, editors, *Studies in ComputationalIntelligence For Technology Enhanced Learning*, *volume 273*, pages 195–224. Heidelberg, Germany: Springer-Verlag

GISMO (2015). *Graphical Interactive Student Monitoring Tool for Moodle.* Retrieved from http://gismo.sourceforge.net/

Graf, S., List, B. (2005). An evaluation of open source e-learning platforms stressing adaption issues. *In Proceedings of 5th IEEE International Conference on Advanced Learning Technologies* (pp. 163-165).

Hadoop (2015). *Hadoop 1.1.2 Documentation.Cluster Setup.* Retrieved from http://hadoop.apache.org/docs/r1.1.2/cluster_setup.html

Mazza R., Milani C. (2005). Exploring Usage Analysis in Learning Systems: Gaining Insights from Visualisations. *AIED Workshops (AIED'05)*

Moodle.net (2015). *Current Moodle Statistics.* Retrieved from https://moodle.net/stats/

Ninoriya Suman, Chawan P.M., Meshram B.B. (2011).CMS, LMS and LCMS For eLearning. *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2*, ISSN (Online): 1694-0814, www.IJCSI.org

Rogers P., Berg G., Boettcher J., Howard C., Justice L., Schenk.Hershey (editors). (2009). Course Management Meets Social Networking in Moodle. M. Crosslin. *The Encyclopedia of Distance Learning, Second Edition*. Information Science Reference New York, Idea Group Inc (IGI).

Rosenberg M. J. (2001).*E-learning: Strategies for delivering knowledge in the digital age*. New York: McGraw-Hill.

Zhang H., Almeroth K., Knight A., Bulger M., Mayer R. (2007). C. Montgomerie & J. Seale (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007* (pp. 4415-4422). Chesapeake, VA: AACE.