S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education

A Framework of User-Driven Data Analytics in the Cloud for Course Management

Jie ZHANG[1], William Chandra TJHI[2], Bu Sung LEE[1], Kee Khoon LEE[2], Julita VASSILEVA[3] & Chee Kit LOOI[4]
[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]A*STAR, Institute of High Performance Computing, Singapore
[3]Department of Computer Science, University of Saskatchewan, Canada
[4]Centre of Excellence for Learning Innovation, National Institute of Education, Singapore
{zhangj@ntu.edu.sg}

Abstract: In this paper, we describe our goal of an effective course management system for assisting course managers to make informed decisions about what materials should be most appropriate to be presented to students (learners) and what learning strategies or methods should be used for the students. The system is supported by our design of a novel framework for user-driven data analytics in the cloud. Different modules of the framework will be illustrated in detail in the context of course management.

Keywords: Course Management, E-Learning, Data Analytics, Cloud Computing

Introduction

Modern technologies have the potential of transforming traditional ways of learning. Instead of gaining most of knowledge from classes, they can now learn knowledge from many resources scattered on Internet. In Singapore, in 2008 the government has invested in a 4-year Future School project to create a futuristic engaging learning environment with the help of Infocomm Technology. Many e-learning systems have also been implemented for online education purposes [10], e.g. Blackborad deployed by Nanyang Technological University. A crucial element for these systems is effective course management. Course managers (i.e. lecturers, course coordinators, and designers of e-learning systems) need to make informed decisions about what materials should be the most appropriate to be presented to students (learners) and what learning strategies or methods should be used for the students [2]. Most existing commercial course management systems, e.g. Blackboard, provide only basic information about student access statistics to the e-learning system. To achieve this, students' learning behaviors need to be modeled by taking into account different factors from distributed information sources and multiple domains: students' performance records from university databases, relevant policies from education authorities, student models from another country or university, general educational trends in a country, and rich secondary data (such as students' social networking patterns). The heterogeneity of these secondary data demands both robust data fusion and intuitive summarization or visualization to make possible interpretation and interaction by course managers. A robust user-driven framework is thus needed to integrate these different aspects involved.

In this paper, we describe our goal of an effective course management system that is supported by our proposal of a novel framework for user-driven data analytics in the cloud. This framework provides course managers with various data analytics services, including an intelligent crawler to find relevant data, a meta miner to recommend the best workflow together with the transfer learning for producing different models (such as student models and course material/learning object models) possibly across different domains, the cloud

compute service to support computation and storage in heterogeneous and distributed environment, and the visual analytics service to allow course managers to interact with our system to find extra relevant data and refine produced models in order to perform more effective course management. We also describe in detail the different modules and components of our proposed framework in the targeted context of course management to clearly illustrate its supported procedures/functionalities towards an effective course management system.

1. Course Management

Our course management system will help course managers to determine which learning materials should be selected for students. Students' usage of the e-learning system kept in the log data will be used for usage mining to predict the students' performance. One important factor for prediction is the seasonality (i.e. access period) to represent student learning behavior in time series. Based on the prediction results, course managers need to personalize the materials that will be presented to the individual students who will likely have different grade levels. Relevant materials will be crawled from the cloud, such as Wikipedia, online discussion forums, various e-learning systems, digital libraries and so on. Data from other domains (i.e. IT service domain) will also provide additional indication for students who are studying in IT area about the demanding IT knowledge and skills they should have in order to be well-prepared for future career. These data will then be used to produce data models for managers to make decisions on which materials will be the most appropriate and relevant for which students. Along this process, the course managers may possibly refine the data analytics workflow, for example, to use different machine learning algorithms and to use a different data mining process through interactive visual analytics.
Our course management system will also help course managers to determine appropriate and effective methods to be used for students to learn. The social relationships among students can be analyzed for peer learning or group learning to enable students to get help from each other and promote interactive and collaborative learning, based on the relevant data from social networking sites in the cloud, such as Facebook and Twitter. Information about different mobile devices preferred by students will also be elicited through student survey. This information can be used to enable proactive pushing learning materials to students so that they can learn from anywhere and at anytime. Our vision of the effective course management system is supported by the design of a user-driven data analytics framework in the cloud, which will be illustrated in the next section.

2. Design of the User-Driven Data Analytics Framework

Figure 1 shows the high-level view of the proposed user-driven data analytics framework. The framework comprises two main modules: Data Analytics, which is responsible for the data analytics operations; and Data Infrastructure, which is responsible for the management of distributed compute and storage resources. The following briefly describes the Data Analytics module's components and their purposes:

- Intelligent crawler: gathers relevant information and services available in the cloud;
- Transfer learner: adapts analytics models from one relevant domain to another;
- Meta miner: recommends to users optimum data analytics workflows;
- Visual analytics: provides visualization and interaction features for users to refine data analytics workflows and output;
- Usage miner: mines collective patterns of usage by users for reusability and collaborative analysis.

The Data Infrastructure module includes the Hadoop framework for distributed computation, a distributed data storage management system, and the Data Broker Service. Hadoop framework ensures the efficiency and resiliency of computation in the cloud. It is also responsible for optimum Hadoop execution scheduling. Data storage management system provides cloud-based storage solutions by for example leveraging the existing cloud-based technologies such as Google File System, which can provide performance, scalability, reliability, and availability to our service. The Data Broker Service component utilizes the usage patterns to optimize the compute and storage needs. We now illustrate the use of the proposed framework in the context of course management.

Course management analytics seeding question is formalized: The course management analytics is seeded by a question concerning the effectiveness of a course design. Without the loss of generality, an example question that is used as our illustration throughout this section is "does providing customized course materials to students in different profile-groups help to improve the overall class grade?" To formalize this question, the proposed system requires three key elements: initial dataset, initial analytics workflow, and Predictive Modeling Markup Language (PMML) conversion of these two elements [5]. Good initial datasets in this context would be abstracted course characteristics (i.e. with abstract features like: learning curve and mathematical skill required), student profiles, and historical records of students' performances. The corresponding analytics workflow can be clustering of students based on their profiles and association rule mining to find strong associations between good student-group performances and course characteristics. These datasets and the workflows are converted into standardized PMML descriptions as queries to the proposed system.

Course management-related resources are gathered and compiled: Using the seeding question provided by course managers (formatted in PMML), the primary goal of this stage is to expand the question by finding relevant datasets, analytics software, knowledge base, and services (e.g. storage and compute resources) in the cloud. Referring to the illustrative question described above the student profiles can be enriched by mining patterns of their social networking, more descriptive clustering algorithm (with PMML descriptions) can be downloaded, known associations between courses and student performances reported in education journal abstracts can be extracted. The intelligent crawler of the proposed system, a component responsible for this expansion, crawls the cloud databases and the Web to gather the resources. The intelligent crawler compiles these resources by converting them into PMML formats. For unstructured data (e.g. journal abstracts), information extraction based on the fields used in the PMML descriptions of the seeding question is performed to extract relevant entities from the data [3]. For semi-structured and structured data (e.g. formatted algorithm descriptions), synchronization with the seeding question is performed by data fusion techniques [1]. To achieve integration with available cloud resources, the compilation of resources conforms to SOA protocols. The output of this stage is the synchronized form of information and services relevant to the seeding questions from the cloud.

Analytics workflow is optimized through human-system collaboration to predict effectiveness of course materials and delivery methods: Taking into account factors enriching resources from the cloud, the initial analytics workflow defined by course managers need to be adjusted for more optimal prediction of course effectiveness. The proposed system is equipped with a meta-mining component, which based on the PMML descriptions of the seeding questions and the relevant resources, proposes to course managers more optimized analytics workflow [9]. Some examples of meta-mining proposals in our context are: 1) with the student profiles enriched by social networking information, a clustering algorithm is more optimized for networked data such as MCL [4] is suggested; 2) making use of Hadoop, a faster parallel version of MCL is deployed; 3)

based on the existing publications on some known course-student performances associations, the support and confidence levels of the association rule miner are fine-tuned. Course managers monitor and refine proposals from the meta-miner through the visual analytics interface of the proposed system. The visual analytics screen shows in real time the graph representation of the student clusters [8], through which course managers can adjust for example the granularity levels of student clusters (i.e. smaller or bigger groups of students are more practical). The mining of associations between courses and student performances is also visualized using tools such as FpVAT [7], which enables course managers to handpick unexpected association rules missed out by the system.

Course management analytics is made reusable for future and collaborative analyses: To support institution-wide use of the proposed system, current analytics activities and usage patterns (e.g. the workflow and the workflow design process) are stored as resources. Other course managers in the future can benefit from collective rules generated by the usage miner component, such as "80% of course managers managing a course attended by more than 300 students prefer student clusters with higher granularity". Aside from the usage miner component, the proposed system also deploys a transfer learning component for reusability [6]. Transfer learning performs adaptation of datasets and analytics workflows from one domain to another (e.g. from clustering of science students adapted to clustering of business students). Relevant analyses from past and cross-domain (e.g. from science to business) course management analytics can be taken into account to enhance the accuracy of current analysis.

Cloud compute system is activated to support course-management analytics: Behind the scenes, a cloud compute system, supported by the Hadoop framework and cloud-storage, serves as the engine of the large-scale and distributed analytics activities involved in the course management analytics. The computation also benefits from the mining of usage patterns that allows optimal allocation of resources, e.g. "Crawling of relevant information for business courses on average requires 10% more storage than engineering courses". With this cloud compute system, course managers are not limited to the processors and storage of their personal computers.

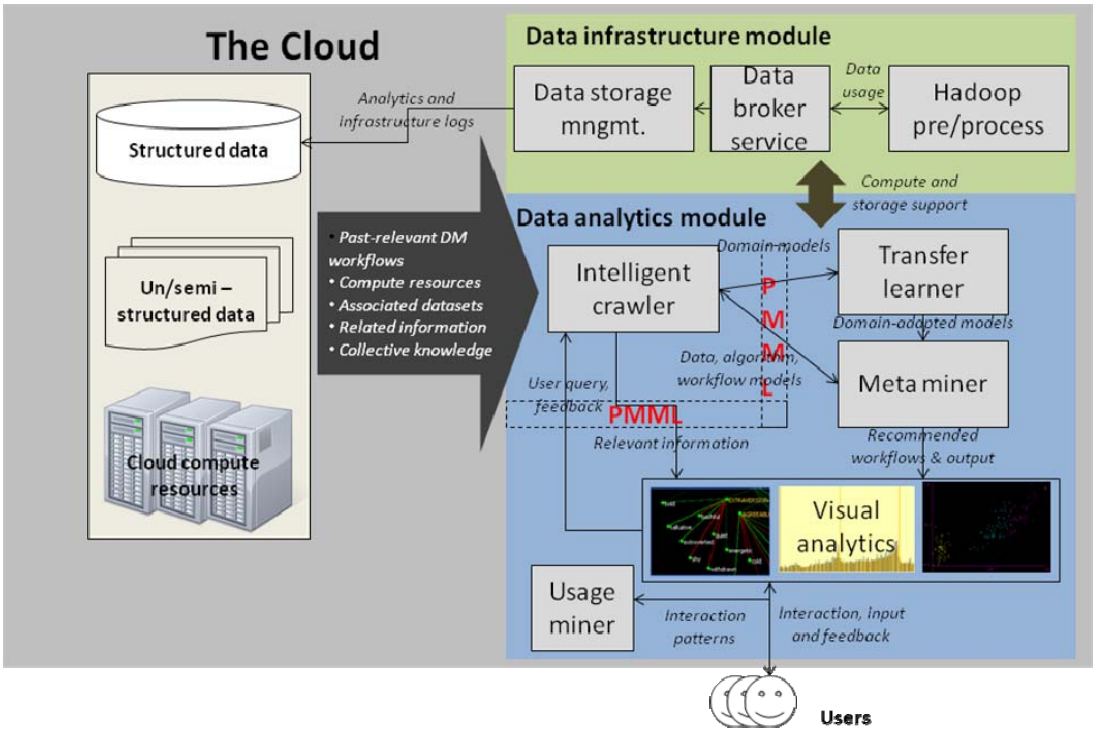

Figure 1. The Proposed Framework for User-Driven Data Analytics in the Cloud

In summary, a package of user-driven data analytics services is offered by our proposed framework to provide the following functionalities of the effective course management system: a) discovering new learning materials through crawling in the cloud (i.e. the web, learning object repositories, and e-learning systems); b) discovering global patterns in students records and profiles to improve existing course materials and learning objects, and optimize learning objects repositories and e-learning systems performance; c) discovering local patterns in the histories of students, by learning models of individual student's knowledge, learning style, motivation level, and social network; d) in combination with other domains, discovery of new important skills and knowledge that need to be learned, to constantly update the teaching goals and keep the skills of the students in sink with the market demands; e) discovering gradually patterns of successful personalization or learning sequences and learning methods based on successful learning sequences and methods for groups of students or classes of students (i.e. collaborative filtering); f) visualizing recommended analytics workflows and learned models or rules of students, learning materials and learning methods to allow course managers to interact with the data analytics results, in order to for example refine the course management analytics workflows, speed up the learning process, or adjust the learned results.

3. Conclusion and Future Work

In this paper, to support our goal of an effective course management system that is crucial for better e-learning, we designed a framework for user-driven data analytics in the cloud. It is a new scalable distributed data analytics framework that complements users of analytics by retrieval, integration and summarization/visualization of relevant heterogeneous information from external sources and facilitates user interpretation, interaction and collaboration to achieve domain-specific solutions.

Our next step is to implement the course management system based on the framework design. Then, we will evaluate the performance of this system. Students and course managers of e-learning systems will be involved. By employing our course management system, it is expected that students' learning will be improved and better feedback will be provided to course managers.

References

[1] Bleiholder J. & Naumann F. (2008) Data fusion. *ACM Comput. Surv., 41*(1):1–41.

[2] Champaign, J. & Cohen, R. (2010). A Model for Content Sequencing in Intelligent Tutoring Systems Based on the Ecological Approach and Its Validation Through Simulated Students. *Proceedings of the Ninth Florida Artificial Intelligence Research Symposium (FLAIRS)*. Daytona Beach, Florida.

[3] Chang C.H., Kayed M., Girgis M.R., & Shaalan K. (2006) A Survey of Web Information Extraction Systems. *IEEE Trans. on Knowledge and Data Engineering*, 0475-1104.R3

[4] Enright A.J., Van Dongen S., & Ouzounis C.A. (2002) An Efficient Algorithm for Large-scale Detection of Protein Families. *Nucleic Acids Research, 30*(7):1575-1584

[5] Guazzelli, A., & Stathatos, K., & Zeller, M. (2009) Efficient Deployment of Predictive Analytics through Open Standards and Cloud Computing. *SIGKDD Explor. Newsl., 11*(1):32–38.

[6] Gupta, A., & Ratinov, L. (2008) Text Categorization with Knowledge Transfer from Heterogeneous Data Sources. *Proceedings of National Conference on Artificial Intelligence (AAAI)*.

[7] Leung C.K. & Carmichael C.L. (2009) FpVAT: A Visual Analytic Tool for Supporting Frequent Pattern Mining. *SIGKDD Explorations, 11*(2), 39-48.

[8] Shen, Z., & Ma, K., & Eliassi-Rad, T. (2006) Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. *IEEE Trans on Visualization & Computer Graphics, 12,* 1427-1439.

[9] Smith-Miles, K. A. (2008) Cross-disciplinary Perspectives on Meta-learning for Algorithm Selection. *ACM Comput. Surv., 41*(1):1–25.

[10] Vassileva, J. (2009). Towards Social Learning Environments, *IEEE Transactions on Learning Technologies, 1*(4), 199-214.