2014 International Conference of Educational Innovation through Technology

Using Learning Analytics Technologies to Find Learning Structures from Online Examination System

Qingtang Liu
School of Educational Information Technology
Central China Normal University
Wuhan, China
liuqtang@mail.ccnu.edu.cn

Guilin Fan
School of Educational Information Technology
Central China Normal University
Wuhan, China
fanguilin@163.com

***Abstract*—The field of learning analytics has the potential to enable education institutions to increase their understanding of their students' learning needs and to use that understanding to positively influence student learning and progression. Analysis of data relating to students and their engagement with their learning is the foundation of this process. In this work we investigate the database of learners' answers to the asked questions by applying the Markov chains. We want to understand whether the learners' answers to the already asked questions can affect the way they will answer the subsequent asked questions and if so, to what extent. Results showed that influential structures were identified in the history of learners' answers considering the Markov chain of different orders. The results could be used to identify undergraduates who have difficulties after couple of steps and to optimize the way questions are asked for each undergraduate individually.**

***Keywords-** learning analytics; Markov chain; learning structure*

I. INTRODUCTION

With the rapid development of information technology, the amount of data on the web has exploded. Several PB data in internet can be produced every day, including log, micro-blog, photos, video, etc., which update in real-time and constantly.

In some sectors, the relatively recent emergence of analytics is now viewed as having the potential to transform economies and increase organizational productivity [1] and increase competitiveness [2]. In the field of education, more and more software systems have been deployed and there stores massive educational data. Unfortunately, education systems—primary, secondary, and postsecondary— have made limited use of the available data to improve teaching, learning, and learner success. Hey, Tansley, and Tolle (2009) arguing that data analytics represent the emergence of a new approach to science [3]. How could we fully use the massive educational data and transfer the data into useful information and knowledge has become the content concerned by educators and learners. Learning analytics could be used for realizing the value of learning process data.

Analytics in education can also be viewed as existing in various levels, ranging from individual classroom, department, university, region, state/province, and international. Buckingham Shum (2012) groups these organizational levels as micro, meso, and macro analytics layers [4]. Learning analytics has become a necessity in education in the big data era with a very broad application prospect. Therefore, teachers in the era of big data should have strong data analysis ability. Learning analytics can promote online professional development of teachers such as promoting learning efficiency and stimulate their sense of autonomous professional development; improving the teaching efficiency and develop online teaching practice wisdom; enhancing the performance of research and improve the ability of the e-learning services.

Research questions

This study aimed to answer the following research questions:

1. Did the students form some learning structures when they did exercises through online system?

2. How the structures can be applied to connect systems to practices if it exists?

We start by reviewing the learning analytics problem and the approach that is used to solve it in Section 2. We then describe our methodology in Section 3. In Section 4, we get the results by experimental analysis. We then discuss the work in Section 5.

II. LITERATURE REVIEW

A. Learning analytics (LA)

LA [5] is a type of data mining that has gained traction in recent years. As the field of LA is further refined and established, an authoritative definition will emerge. At present, the vast majority of LA literature has begun to adopt the following definition offered in the first international Conference on Learning Analytics and Knowledge (LAK

978-1-4799-4230-5/14 $31.00 © 2014 IEEE
DOI 10.1109/EITT.2014.38

2011) and adopted by the Society for Learning Analytics Research:

Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs.

As this is a newly defined field, papers relating to learning analytics draw on a diverse range of literature from fields including education, technology and the social sciences [6]. Hendler and Berners Lee (2010) state that "the problems that our society faces today are such that only the concerted effort of groups of people, operating with a joint power much greater than that of the individual can hope to provide solutions"[7].

Baker and Yacef (2009) address the technique dimension of LA in listing five primary areas of analysis [8]:

- Prediction
- Clustering
- Relationship mining
- Distillation of data for human judgment
- Discovery with models

Through statistical analysis, neural networks, and so on, new data-based discoveries are made and insight is gained into learner behavior. This can be viewed as basic research where discovery occurs through models and algorithms [9].

Table 1. Three different fields and their researches

fields	focusing on	research
Educational data mining	technical challenge	How can we extract value from these big sets of learning-related data?
Learning analytics	educational challenge	How can we optimize opportunities for online learning?
Academic analytics	political/economic challenge	How can we substantially improve learning opportunities and educational results at national or international levels?

B. Answer type

The answer type is picked from a hand-built taxonomy having dozens to hundreds of answer types [10][11].With the increasing popularity of statistical NLP, [12], [13] and [14] used supervised learning for question classification

Li and Roth (2002) used a Sparse Network of Winnows (SNoW) with features included tokens, parts of speech (POS), chunks (non-overlapping phrases) and named entity (NE) tags. They achieved 78.8% accuracy for 50 classes [12].

Hacioglu and Ward (2003) used linear support vector machines (SVMs) with question word 2-grams and error-correcting output codes (ECOC)—but no NE tagger or related word dictionary—to get 80.2–82% accuracy [13].

Zhang and Lee (2003) used linear SVMs with all possible question word q-grams, and obtained 79.2% accuracy [14].

All of them have got a good accuracy in answer type classification. However, they do not give a prediction when undergraduates answer a question. In the following Section, we use the Markov chain to give the probabilities of answer types.

III. METHODOLOGY

The use of data for improving learning is common in universities. Much of this activity currently happens at a small scale in individual classrooms, where educators use data collected manually or through analysis of server logs to provide individual educators with feedback on which exam questions cause learner confusion or which learning activities or lectures need greater clarity as measured by learner performance on exams or tests.

A. Markov chain

Markov chain is a sequence of random variables X0, X1, X2,…, Xn for which the following Markov property holds:

P(Xn+1=x|X0,X1,X2,…,Xn)=P(Xn+1=x|Xn).

The Markov chain of first order is called memoryless, meaning that the next state depends only on the current state. Considering the Markov chain, the probability of the next state depends on the previous states. A transition matrix P of all stochastic transition probabilities between the states represents the Markov model. In this paper we defined 6 answer types to build the states of Markov chain model.

B. Dataset

In this paper we implement our learning analytics on the dataset form the database of online examination system for C programming language. Our data set contained 178620 answered questions by 362 sophomores. Their ages ranged from 18 to 22. A first manual analysis of data made clear that there was some noise in the dataset. We select 10 knowledge points about C pointer for the reason that they form a big knowledge unit as a whole. The dataset reduced to 17920 answered questions by 356 sophomores. 97 of them were females and 259 were males. The online examination system puts each knowledge point to the undergraduates over 3 times with different type. The answers for each knowledge point are classified based on the correctness / incorrectness of the submitted answers. If a undergraduate answers a question correctly for the first time, we mark the answer type T (T stands for true), another type question of the same knowledge point is asked again later to ensure that the knowledge point is truly known by undergraduate. If the other question of the same knowledge point is answered correctly for the second time we mark the answer type TT. On this kind of situation the online examination system assumes that the undergraduate had already know well of the knowledge point. In addition, if the undergraduate answers incorrectly in the second time we mark the answer type TF (F stands for false), the system keeps on asking the same knowledge point question with different type later until the undergraduate answers it correctly and we mark the answer type FT. The following example shows how the answer types are marked. Assuming a undergraduate answers the same knowledge point question for 4 times and the results as follows: T, F, F, and T. The answer types for this knowledge point would be: T, TF, FF, and FT. These answer types build the states of Markov chain model in our analysis. Table 2 shows the 6 answer types we marked.

Table 2. Six different answer types and their definition
“T” stands for “True” and “F” stands for “False”

Answer type	Meaning	Preceding answer	Current answer
T	First true answer	-	T
F	First false answer	-	F
TT	First true answer and second true answer	T	T
TF	First true answer and second false answer	T	F
FT	First false answer and second true answer	F	T
FF	First false answer and second false answer	F	F

IV. Results

The answer types to each knowledge point represent the states and probabilities in the sequences as transition links between the states. There are 6 answer types we have defined. So we built a Markov matrix of the size 6*6 representing the transition probabilities.

A. Markov model of entirely all data

We analyze the Markov model entirely over all dataset. Fig. 1 shows the Markov chain probabilities of transitions by growing order n. For the sake of description, we only depicted the first 5 orders (n=1, 2, 3, 4, 5).

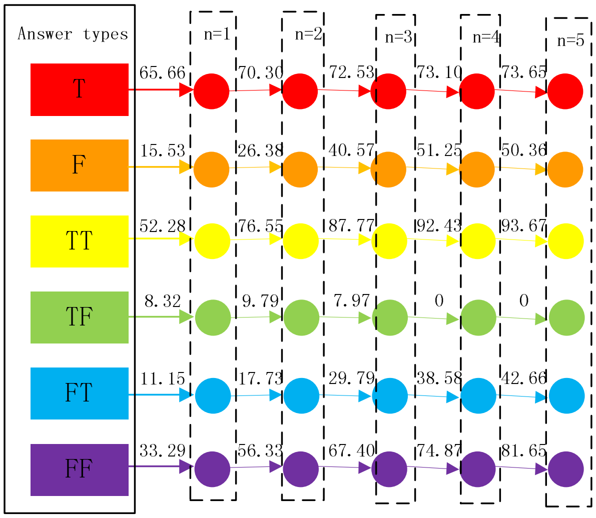

Figure 1.Markov chain probabilities of entirely over all dataset self-transitions (in %) for each answer type.

We can find that the probabilities for all answer types increase by ascending n besides TF. For example by n=1 the probability of transition from FF to FF state is about 33.29%. But by n=2 the probability increases to 56.33%, by n=3 it increases to 67.4%. It means that assuming a undergraduate’s last answer is of type FF. He/she will answer the next asked question of the same knowledge point with the probability of 56.33% incorrectly for the second time. However, the probabilities for answer type T show that undergraduates have already mastered the knowledge point. He/she will answer the next asked question of the same knowledge point with the probability about 70% correctly for the second time. The same thing happens in TT answer type. The probabilities of answer type TF decreases after the second step and reaches 0 in the fourth step. This implies that if undergraduates answer the questions correctly, they rarely give incorrect answers in the second time. The probability decreases in the next step and then to 0.

B. Markov model of every undergraduate

In 4.1, we analyze the Markov model entirely over all dataset. However, are there any differences between the top and bottom ranking by undergraduate’s score? Firstly, we sorted the undergraduates according to their score. Secondly, we selected the top 30% as high group and the bottom 30% as low group. Thirdly, imitate the method in 4.1, we got Fig. 2 Markov chain probabilities of high group self-transitions (in %) for each answer type and Fig. 3 Markov chain probabilities of low group self-transitions (in %) for each answer type.

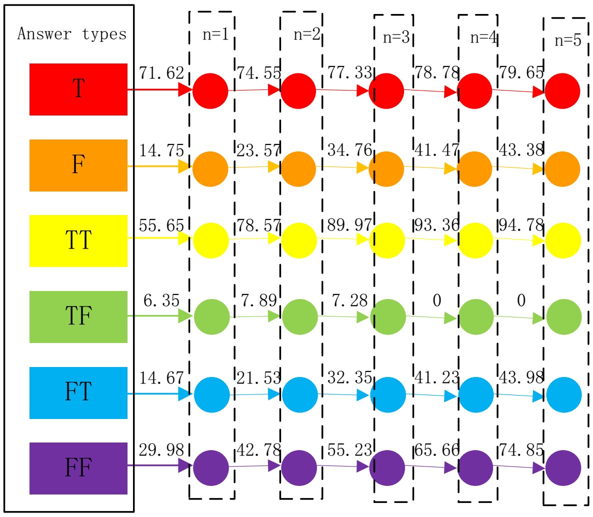

Figure 2.Markov chain probabilities of high group self-transitions (in %) for each answer type.

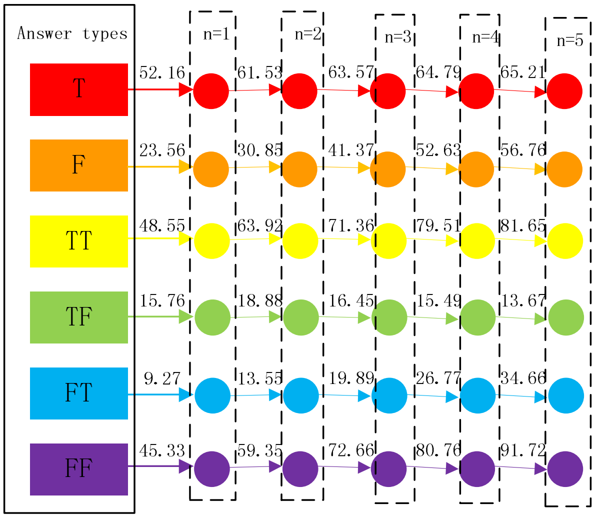

Figure 3.Markov chain probabilities of low group self-transitions (in %) for each answer type.

As can be seen, for the answer types of T, TT and FT, the probabilities in the low group are smaller than the probabilities in the high group. The other three answer types of F, TF and FF, the probabilities in the low group are bigger than the probabilities in the high group. There are some differences between the two groups particularly in answer types TF and FT. For example, in the high group, the probabilities of answer type TF decreases and reaches 0 by n=4. But in the low group, the probabilities of answer type TF fluctuate and not reach 0. They will answer incorrectly in the second round even they answer correctly in the first round. This implies that in the low group, undergraduates do not master the knowledge point strongly. It is similar in answer type FT. In high group, the probabilities of answer type FT increases but it is not obvious in low group. This implies that in the low group, if undergraduates once answer the question incorrectly, the probabilities they will give correctly answer for the next time is very small.

V. Discussion

In this work we analyzed the dataset from online examination system by applying the Markov chain and found some structures within undergraduates' answer types. However, a single data source or analytics method is insufficient when considering learning as a holistic and social process. Multiple analytic approaches provide more information to educators and students than single data sources.

The model we found by using Markov chain is of interest to teachers to know beforehand whether the undergraduates will have difficulties during answer questions online. The goal is to support teachers to discover undergraduates' weak points. However, different knowledge point has different difficulty coefficient. The difficulty should be taken into account when we calculate the transition probability.

The learning process is creative, requiring the generation of new ideas, approaches, and concepts [15]. In contrast, Analytics is about identifying and revealing what already exists. Learning analytics systems and software may in the future be capable of innovation in modeling learner creativity. The tension between innovation and analytics is one that will continue to exist in the foreseeable future.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NO. 61272205) and it was also supported by the excellent doctorial dissertation cultivation grant from Huazhong Normal University(NO. 2013YBYB51).

References

[1] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[2] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. 2011. Analytics: The widening gap. MIT Sloan Management Review.

[3] Hey, T., Tansley, S., & Tolle, K. (Eds.). 2009. The fourth paradigm: Data-intensive scientific discovery. Microsoft Research.

[4] Buckingham Shum, S., & Ferguson, R. 2012. Social learning analytics. Educational Technology and Science, 15(3), 3-26.

[5] Siemens, G. 2011. About: Learning analytics and knowledge. In 1st International Conference on Learning Analytics and Knowledge, 2011.

[6] Ferguson, Rebecca. 2012. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6) pages 304–317..

[7] Hendler, J., & Berners-Lee, T. 2010. From the semantic web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence, 174(2), 156-161..

[8] Baker, R. S. J. d., & Yacef, K. 2009. The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1).

[9] Herskovitz, Baker, R. S. J., Gobert, J., Wixon, M., & Pedro, M. S. 2013. Discovery with models: A case study on carelessness in computer-based science inquiry. American Behavioral Scientist.

[10] C Kwok, O Etzioni, and D. S Weld. 2001. Scaling question answering to the Web. In WWW Conference, volume 10, pages 150–161, Hong Kong..

[11] S Dumais, M Banko, E Brill, J Lin, and A Ng. 2002. Web question answering: Is more always better? In SIGIR, pages 291–298..

[12] X Li and D Roth. 2002. Learning question classifiers. In COLING, pages 556–562..

[13] K Hacioglu and W Ward. 2003. Question classification with support vector machines and error correcting codes. In HLT, pages 28–30..

[14] D Zhang and W Lee. 2003. Question classification using support vector machines. In SIGIR, pages 26–32..

[15] George Siemens. 2013. Learning Analytics: The Emergence of a Discipline. American Behavioral Scientist, 57(10), 1380–1400..