Computers in Human Behavior 65 (2016) 9–13

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

A Tutor Assisting Novel Electronic Framework for Qualitative Analysis of a Question Bank

Nagashree N.*, Nitin V. Pujari

PES University, Department of Computer Science and Engineering, 100 Feet Ring Road, Banashankari Stage III, Bengaluru, Karnataka 560085, India

ARTICLE INFO

Article history:
Received 10 June 2016
Received in revised form
28 July 2016
Accepted 2 August 2016

Keywords:
Learning analytics
Intelligent tutoring systems
Question bank
Clustering
Classification

ABSTRACT

The traditional approach followed by tutors to assess the students is through a set of questions. The quality of a question bank has an impact on the effectiveness of evaluation in educational institutions. Determining the coverage of these questions with respect to a set of prescribed text/reference books helps in evaluating students efficiently. In this paper, we describe a Tutor Assisting e-Framework (TAeF) that enables the tutors to analyze the quality of a question bank. Initially, it clusters all individual topics of each of the input text/reference books according to their dependencies. Later, the questions are classified into these topics. The result is a set of topics, each containing the topic title and the probability by which the question is related to it. Lower the accuracy of the predicted topics, higher is the quality of the question. In other words, if question contains the topic title unaltered, it has a higher probability of being related to the topic; this degrades the quality of question. Furthermore, the congruence relation between the questions and the set of topics is found. This gives the question coverage of each topic. Finally, with this relation, the percentage of understanding the students have developed in each of these topics is computed. The Tutor Assisting e-Framework (TAeF) helps to improve the quality of a question bank, to check the topics covered by each question and the knowledge gained.

1. Introduction

Analytics is a term used for fact-finding or exploratory and predictive models and statistical analysis to gain perception and inference of data generated by the ecosystem of the domain under consideration. In simple terms, it means to uncover meaningful patterns in dataset. The value of analytics is especially significant in domains that are information intensive.

Learning analytics (LA) is a genre of analytics that aims to accomplish much greater success in the learning process (Brown, 2012). Learning analytics is defined as the act, analysis or inspection, collection or representation, measurement or quantification and reporting of large data about learners and their circumstances, with the intent to interpret and optimize the activity of learning and the domains in which it prevails. Learning Analytics points at:

- optimizing the overall performance of tutor and learner,
- fine-tuning various pedagogical policies,
- smoothing the incurred institutional costs (educational expenses assessed on students),
- finding the engagement of students with the course,
- projecting the possibly struggling students and adapting pedagogy accordingly,
- improving the grading systems with the help of real-time analysis and
- allowing the tutors, instructors to evaluate their own efficacy in education whether it is by using interactive visualizations, statistical procedures, predictive modeling, taxonomies or framework.

In short, it makes an effort to improve pedagogy and learning by analyzing performance data of students (Larusson & White, 2014, pp. 1–38).

Learning Analytics consists of five different steps. Campbell and Oblinger have framed the definition of academic analytics as an engine that aids in making decisions or to guide some of the actions consisting of five steps (a) act, (b) capture or represent data, (c) report or record, (d) predict and (e) refine. The LA stage begins with the capture of raw data which is later processed and translated into information in the report phase. In the next phase, based on the

* Corresponding author.
E-mail addresses: nagashree.n.iyer@gmail.com (N. Nagashree), nitin.pujari@pes.edu (N.V. Pujari).

http://dx.doi.org/10.1016/j.chb.2016.08.008
0747-5632/

knowledge captured and sagacious action, prediction is made. Finally, the refine phase views learning analytics as a self-enhancement process or project in which its impact is monitored continuously i.e., on a regular basis.

Learning Analytics technologies enable tutors to capture the novel approach to any of the educational issues. The advancement in technology has an influence in the gradual transition of education contexts from content-centric to student-centric (Larusson & White, 2014, pp. 1–38). The use of educational data mining (EDM) and learning analytics to research and develop models in several areas has influenced the online learning systems. In the field of LA, in education and learning, research has developed rapidly over the past few years. The rapidly inflating availability of large datasets from various institutional learning systems provides great potential for data interrogation with the goal of enhancing teaching and learning environment. Nevertheless, the learning analytics tools fall short of their potential if there is a lack of clear understanding of academic needs, thereby missing the mark. Learning analytics allows improvement of educational outcomes by analyzing the data about learners and their activities. LA focuses on the process of learning at the personal, course, degree, departmental or institutional level. Student success, usually measured as completion of degree, is given primary significance. With its rise, there is an increase in the demand for institutional accountability. LA enables the institutions to address this while fulfilling the academic needs. There are numerous academic systems available to generate a wide range of data to predict retention and graduation of the students (Bollenback, 2014).

2. Literature review

The novel but rapidly evolving field of learning analytics (LA) has grabbed the attention of many researchers. From academic analytics (AA) to business analysis to social network analysis (SNA), learning analytics has remained prominent.

EDUCAUSE and SOLAR have been active on the learning analytics research from the past few years. Learning analytics, however, is not one of the less active research areas. It takes from different fields that are related and synthesizes several existing techniques. Considering the six critical dimensions – stakeholders, internal limitations, external constraints, instruments, objectives and data, Greller and Drachsler (2012) have designed a generic framework for LA.

LA focuses specifically on designing applications for student assessment. In "A Learning Analytics Approach to Academic Assessment", student performance reports are based on all outcomes, which are aligned to assignments, tests, questions, forums, or group assignment (Bollenback, 2014). This describes the way LA can be utilized to view a snapshot of performance at the level of course, program, or institution based on student learning outcomes. This system is designed to capture data such as college, department, course number etc. from the students. Based on these, a set of learning outcomes are generated. The students choose the learning outcomes they have achieved, one at a time, and justify them.

3. Methodology

The Tutor Assisting e-Framework (TAeF) enables the tutors to enhance the quality of a question bank. It clearly distinguishes good questions from the poor ones based on the key words present in the questions. If there is a 100% match between the topic title and the question, the system classifies it as poorly framed. Apart from this, the system also gives the question coverage of every individual topic in the reference book(s).

The flowchart of TAeF is depicted in Fig. 1. TAeF has three modules: (i) Parser, (ii) Clustering System and (iii) Classification System. The parser processes the reference book input as a single, editable PDF file and generates an XML file for the table of contents. For each topic in the book, it generates a text file that contains its title and the following text. Given an arbitrary number of clusters, the clustering system groups the individual topics in the book according to their dependencies. The classification system classifies the input questions into various topics. It generates a probability by which each question is related to each class in the set. It calculates the congruence relation and the percentage of understanding gathered by the corresponding students in these topics.

3.1. Parser

The e-copy of the reference book – a single, editable PDF file, is input to the parser. Also, the first and last chapters' titles are fed to the parser. The parser generates an XML file for the table of contents in the eBook. The knowledge is usually organized in levels between 1 and 3; if not, it is shrunk to level 3. Visualizing the entire eBook as a tree of chapters, topics and sub-topics, the XML file is processed to generate a temporary file containing the level of every individual topic, the title of topic and the page number of the topic. Further, this helps to process and store the content of the topics into text files. For every topic, pointers are assigned to the titles of current topic and following topic; also to their corresponding page numbers. These pages are converted to a temporary text file. The text file is read and the content between current topic title and following topic title along with the title of current topic are written into a text file. This is executed for every individual topic in the eBook. Clearly, the number of text files generated is equal to the number of topics between the specified first and last chapters. The irrelevant topics such as references, summary, to do etc., are ignored. This contributes to the training dataset.

3.2. Clustering system

The clustering algorithm used is K-means. It segregates n data points into k clusters such that each data point is a part of the cluster with nearest mean as shown in equation (1).

$$\arg\min_s \sum_{i=1}^{k} \sum_{x \epsilon S_i} \|x - \mu_i\|^2 \quad (1)$$

where, $S = \{S_1, S_2, .., S_k\}$ is the set of k clusters, $x_1, x_2, .., xn$ are n observations and μ is the mean.

This system groups the topics into the given number of clusters. The training dataset is loaded and processed. Hyphenated words are concatenated, stop words are removed and stemming is done. A term frequency-inverse document frequency (tf-idf) matrix is generated for it. The tf-idf representation of a document d is shown in equation (2).

$$d_{tfidf} = (tf_1 \cdot \log(N/df_1), \quad tf_2 \cdot \log(N/df_2), \ \ldots, \ tf_n \cdot \log(N/df_n)) \quad (2)$$

where, N is the total number of documents in the dataset, tf_i is the term frequency, df_i is the document frequency and $\log(N/df_i)$ is the inverse document frequency of i^{th} document. Later, this matrix along with the number of clusters is input to K-means to generate clusters. A detailed statistical report is produced. A scatter plot is generated for better visualization.