Knowledge-Based Systems 000 (2016) 1–9

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Content-based methods in peer assessment of open-response questions to grade students as authors and as graders

Oscar Luaces[a,*], Jorge Díez[a], Amparo Alonso-Betanzos[b], Alicia Troncoso[c], Antonio Bahamonde[a]

[a] Artificial Intelligence Center, University of Oviedo, 33204 Gijón, Spain
[b] Dept. of Computer Science, Faculty of Informatics, University of A Coruña, 15071 A Coruña, Spain
[c] Dept. of Computer Science, Pablo de Olavide University, 41013 Sevilla, Spain

ARTICLE INFO

Article history:
Received 15 February 2016
Revised 9 May 2016
Accepted 19 June 2016
Available online xxx

Keywords:
Peer assessment
Factorization
Preference learning
Grading graders
MOOCs

ABSTRACT

Massive Open Online Courses (MOOCs) use different types of assignments in order to evaluate student knowledge. Multiple-choice tests are particularly apt given the possibility for automatic assessment of large numbers of assignments. However, certain skills require open responses that cannot be assessed automatically yet their evaluation by instructors or teaching assistants is unfeasible given the large number of students. A potentially effective solution is peer assessment whereby students grade the answers of other students. However, to avoid bias due to inexperience, such grades must be filtered. We describe a factorization approach to grading, as a scalable method capable of dealing with very high volumes of data. Our method is also capable of representing open-response content using a vector space model of the answers. Since reliable peer assessment requires students to make coherent assessments, students can be motivated by their assessments reflecting not only their own answers but also their efforts as graders. The method described is able to tackle both these aspects simultaneously. Finally, for a real-world university setting in Spain, we compared grades obtained by our method and grades awarded by university instructors, with results indicating a notable improvement from using a content-based approach. There was no evidence that instructor grading would have led to more accurate grading outcomes than the assessment produced by our models.

1. Introduction

Massive open online courses (MOOCs) offer promising new educational opportunities and have focused the attention of many researches in terms of improving the educational experience of students. Since MOOCs attract thousands of students, assessment in particular – in order to provide feedback to students and to guarantee the quality of qualifications – is a problematic issue, since the vast numbers of students enrolled implies a huge or even impossible burden for instructors and teaching assistants. Assessment is therefore one of the most complex and challenging applications of big data in education.

We tackled the challenge of evaluating open-response questions, adopting, as our basic strategy, peer assessment [1–8], whereby students evaluate the anonymized answers of other students participating in the same course. Students, in their role as graders, receive a set of detailed rules (called a rubric) designed to ensure uniform assessment. However, since students typically have no experience of assessing peers, grading must deal with the effects of inconsistent and subjective evaluation. Yet peer assessment also has an important pedagogical function in that a deeper understanding of course content is obtained when students read and are required to assess other students' answers. The two main peer assessment streams are cardinal and ordinal.

In cardinal peer assessment, grades are numbers or categorical labels with straightforward numerical semantics. If we have a sufficiently large number of grades for each assignment, then the correct grade could be approximated by computing the mean or the median [2]. Means have been reported to be more consistently accurate with respect to the rubric than staff grades [3]. However, one problem with the cardinal approach is that students cannot be charged with the job of grading large numbers of answers and another issue is that it is affected by the lack of student experience in assessment.

* Corresponding author.
E-mail addresses: oluaces@uniovi.es (O. Luaces), jdiez@uniovi.es (J. Díez), ciamparo@udc.es (A. Alonso-Betanzos), atrolor@upo.es (A. Troncoso), abahamonde@uniovi.es (A. Bahamonde).

http://dx.doi.org/10.1016/j.knosys.2016.06.024
0950-7051/

In the ordinal approach to peer assessment, graders rank answers in terms of their quality [5,7,8] – clearly an easier task for inexpert graders than cardinal grading, as evidenced by the considerably higher reliability reported for ordinal compared to cardinal assessment [1,9–12]. (see [13] for an interesting discussion of cardinal and ordinal peer grading from a psychological point of view).

Another approach to assessment is content-based methods, which use information retrieval techniques, for instance, a preference approach to learning the relevance of documents [14]. These methods require some shallow linguistic processing and also frequently require assistance from the instructor. Methods include comparing several ideal answers (references) with student answers or labelling a subset of answers with correct grades that are then extended to the whole set of answers using a machine learning algorithm.

As far as we are aware, no existing peer-assessment method takes into account the content of student responses to open questions. Since peer-assessment methods function like collaborative filters that recommend a grade for each answer, their predictive power could be enhanced using available information about answers.

We describe an approach that combines the strengths of ordinal collaborative filters and content-based recommenders. We use a factorization method to train a utility function that estimates consensus in rankings of answers. This approach – inspired by a preference learning framework [9,15] – was used in previous research by us [6,7,16,17]. Answers can be represented by vectors of features, which have been acknowledged to be crucial for the success of peer assessment [18,19]. If no other information is available, features only capture a binary identification of answers and graders, reflecting a pure collaborative approach. However, our factorization method allows representations that include other information about the answers. Unlike other approaches, our proposed method does not need any self-grading of answers or any previous grading by instructors.

We also propose a method to grade students as graders, as student grading is a potentially powerful motivational aspect in learning. This would require announcing, before starting a course, that students' final grades would be calculated as a linear combination of answers both authored by and graded by students.

Below we formally describe our assessment method and results for a real-world data set based on a computer science assignment issued to students at three Spanish universities, reporting discrepancies for our methods with instructor grades that were similar or lower than discrepancies between instructors. We tested both collaborative filtering and content-based representations, finding that the latter achieved considerably better results. Our proposal for grading graders also obtained good results, which improved, furthermore, in line with the number of answers evaluated by graders. The use of content also improved scores in most cases.

The paper is organized as follow: in the next section we present some related state-of-the-art works. Then, in Section 3 we formally introduce our approach, including detailed equations and explanations. This section is followed by a detailed description of the experimental setting and the results obtained, together with an analysis of the performance of the content-based and collaborative filtering approaches, as well as performance of the grader assessment. The paper ends with a short summary and some conclusions derived from this research.

2. Related works

As with recommender systems, automatic assessment methods can be split into two groups: those that use answer content provided by students and those that function as collaborative filters.

Some interesting content-free assessments have been described [3,5] with authors emphasizing the importance of assessing grader accuracy. In fact, accurate evaluations are crucial to obtaining reliable data so one way to encourage good-quality grading is to include grading of peer-grading assessments as part of the student's final grade.

Shah et al. [19] propose using methods that include some kind of *dimensionality reduction*, e.g., clustering, and using features to represent the issues involved in assessment. Although their proposals are very abstract, the factorization method proposed here offers a suitable framework for implementing both approaches.

In the area of content-based systems, the most widely used option is to combine shallow Natural Language Processing with Machine Learning, that is, methods borrowed from the Information Retrieval field. Broadly speaking, we can distinguish between matching and categorization methods.

Matching methods compare students' answers against some reference (ideal answer) or template; Pérez-Marín et al. [20] made a detailed survey of published algorithms which used this paradigm. Rodrigues and Oliveira [21] matched students, answers with references by computed cosine similarity after preprocessing. Both references and students' answers were represented using the vector space model (VSM) [22], in which each word is the index of a vector whose values – which may be weighted using different strategies – record the presence or frequency of a word in a document.

To deal with answer content, some authors have used matching methods that exploit coincidences between groups of words, with the aim being to take into account the syntactic structure of documents without penalizing the process with a deep analysis. A key tool in this case is a metric of document similarity called BLEU [23], devised to assess the quality of machine translations. Given a set of reference translations, BLEU computes scores for candidate translations based on the co-occurrence of n-grams in the references and candidate translations. A modified version of BLEU was used by Noorbehbahani and Kardan [24] to build a system for automatic assessment of open-ended answers.

The main disadvantage of content-based methods is that they do not consider synonyms. Since we cannot reasonably expect students to use exactly the same words as used in reference answers, a certain degree of semantic analysis is necessary to fairly compare students' answers with references. One way to overcome this problem is to use *Latent Semantic Analysis* (LSA) [25], which projects the matrix of VSM representations of all answers (usually called the *term-document* matrix) into a smaller dimensional space using the singular value decomposition (SVD) of the matrix. This robust information retrieval method thus manages to capture the implicit semantics of a set of documents.

A pilot LSA study that evaluated six students' answers to three questions in the computer science domain reported high precision despite a small data set [26]. LSA was also used to assess participants in a professional development program according to five attitudinal categories of free-form text responses [27], with preprocessing – based on standardization, stop-word removal and Porter stemming – implemented in order to obtain the term-document matrix. Pérez et al. [28] proposed combining BLEU and LSA to assess open-ended answers. In our factorization approach – a generalization of the SVD matrix decomposition method – the decomposition aims to optimize a loss function and so improve predicted outcomes.

Rodrigues and Oliveira [21], mentioned above, included semantic analysis in their proposed cosine similarity method, whereby two words were considered to be similar if they were related in the WordNet semantic network.

Another content-based approach is an adaptation of text categorization, whereby a reduced set of answers is graded by the instructor and then processed by an ordinal classifier that learns

Please cite this article as: O. Luaces et al., Content-based methods in peer assessment of open-response questions to grade students as authors and as graders, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.06.024