Expert Systems with Applications 41 (2014) 594–606

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An empirical study on the quantitative notion of task difficulty

Ricardo Conejo, Eduardo Guzmán, Jose-Luis Perez-de-la-Cruz, Beatriz Barros *

E.T.S. Ingeniería Informática, Universidad de Malaga, 29071 Malaga, Spain

ARTICLE INFO

Keywords:
Web-based educational system
Intelligent tutoring systems
Knowledge assessment
IRT
CAT
Item difficulty
Item calibration
Difficulty estimation

ABSTRACT

Most Adaptive and Intelligent Web-based Educational Systems (AIWBES) use tasks in order to collect evidence for inferring knowledge states and adapt the learning process appropriately. To this end, it is important to determine the difficulty of tasks posed to the student. In most situations, difficulty values are directly provided by one or more persons. In this paper we explore the relationship between task difficulty estimations made by two different types of individuals, teachers and students, and compare these values with those estimated from experimental data. We have performed three different experiments with three different real student samples. All these experiments have been done using the SIETTE web-based assessment system. We conclude that heuristic estimation is not always the best solution and claim that automatic estimation should improve the performance of AIWBES.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

The advent of the Internet has entailed the apparition of several kinds of tools. From an educational perspective, the Internet is a repository of information that both teachers and students can use to their own benefit. However, the evolution of technologies used to develop web-based tools has led to the use of new and more sophisticated systems. These new tools offer the student a tutored learning process, emulating the behavior of a teacher in a classroom. Such systems are called *Adaptive and Intelligent Web-based Educational Systems* (AIWBES) (Brusilovsky & Peylo, 2003) and they are the evolution of two families of systems: Intelligent Tutoring Systems and *Adaptive Hypermedia Systems* (Brusilovsky, 2001). The first have emerged as a result of applying Artificial Intelligence techniques to *Computer-Assisted Learning* (CAL) Systems. *Intelligent Tutoring Systems* (ITS) are also influenced by two other knowledge areas, like Cognitive Psychology and Educational Research. Initially, they were intended to partially automate the task of providing the student with individualized and self-paced learning instruction.

In AIWBES, the learning process is adapted to student needs. This adaptation requires the elicitation and updating of student models. A *student model*, also called *leaner model*, (LM) represents the perception of the system about the learner (VanLehn, 1988). Selecting the right task or question to pose according to the student model is a central topic in intelligent learning systems. (Barla et al., 2010). The quality of an AIWBES will be determined by the scope and quality of the data stored in the learner model, and by the ability of the system to update this model appropriately. This update is usually carried out on the basis of evidence generated from student examinations. Student responses to tasks are raw data which should be converted into information and used to update learner models. The selection of the most appropriate task and the process of updating learner models depend on the properties of the task. Perhaps one of the most relevant properties is task difficulty. Everybody has a subjective notion of what difficulty means and, if we asked a set of persons to give a precise definition of it, they would surely supply related but different statements.

One of the most used techniques for student knowledge diagnosis is testing. There are well-known psychometric theories that relate observed student responses to his/her knowledge state. Most of the tests we find in AIWBES are based on the Classical Test Theory (CTT). This theory, although easy to apply, does not guarantee reliable and invariant diagnosis. Item Response Theory, (IRT) appeared later to solve some of those problems.

Both theories, i.e. CTT and IRT, provide statistical definitions on the concept of difficulty and use data-driven mechanisms to compute the *difficulty* value. However, we can find that in practice most systems use *estimations provided by human "experts"*. There are also some "mathematical" proposals to estimate the *difficulty* from a set of features of the task, such as its complexity or the number of concepts involved, by means of a formula that predicts the *difficulty* or the student performance.

To sum up, there are three different approaches for estimating the "*difficulty*" of a task:

- *Statistical*, that is, estimating the *difficulty* from a previous sample of students.
- *Heuristic*, that is, by human "experts" direct estimation.

* Corresponding author. Tel.: +34 952 13356.
E-mail address: bbarros@lcc.uma.es (B. Barros).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.084

- *Mathematical*, given a formula that predicts the *difficulty* in terms of the number and type of concepts involved in the task.

Statistical approaches require a previous definition of the concept of *difficulty*. So it is commonly associated with CTT or IRT assessment (see Section 2.1), but there is an increasing interest in the ITS and AIWBES community for data mining methods to adjust and fine tune system performance (Romero & Ventura, 2010).

On the other hand, *heuristic* approaches are common in ITS and AIWBES, (see Section 2.2), but IRT assessment sometime use *heuristic* estimation of the item parameters. Teachers, or course creators, are commonly the "experts" that estimate the *difficulty* but there are some experience of using the students as "experts" (see Section 2.3).

What we have called *mathematical* approach can also be viewed as a complex form of heuristic, because the formula itself and the parameters involved are also given by human experts. This approach is mainly used in ITS and AIWBES (see Section 2.2), but also in IRT assessment, for instance to predict the parameters of an item generated from a template (Geerlings, van der Linden, & Glas, 2013). However, *mathematical* approaches are commonly related to complex tasks or problems. In this paper we will focus on simple tasks, like test questions and compare the statistical and heuristic approaches to the *difficulty* parameter estimation.

Another dimension of the problem is time. Parameters need to be configured in some way before the system can be used. If difficulty parameters are estimated *heuristically* they mostly remain unchanged forever because the estimation requires a high costly human effort. On the other hand, there is a cold start problem for the *statistical* approach. This is the case of some IRT models, that require hundred of data to calibrate. Mixed approaches have been used in practice, like a heuristic initial estimation followed by a statistical updating (see Section 2.2). Other authors propose heuristic formulas to continuously update *difficulty* values, based on methods like the Elo rating (Klinkenberg, Straatemeier, & van der Maas, 2011).

This paper tries to contribute to some open research questions: Do statistical and heuristic estimations of difficulty correlate? Are heuristic estimations consistent? Do teachers' estimations and students' estimations correlate? Are heuristic estimations always reliable?

In this work we have carried out several experiments in order to study whether human expert (teacher/student) estimations are similar to *difficulty* values inferred by applying data-driven techniques. We have also explored the alignment of teacher and student viewpoints regarding the quantitative notion of task *difficulty*. Our aim is to focus on the relevance of having a clear understanding of what task *difficulty* represents, especially in AIWBES where educational instruction is adapted to the student needs.

In the next section, primary devoted to the background of this research, we introduce some notions about student modeling and knowledge diagnosis. Test theories and how they define the *difficulty* are considered. We also review some intelligent educational systems, focusing especially on how they manage the task *difficulty*. Section 3 introduces the SIETTE system, which has been used as a workbench to support these experiments. Section 4 describes three different experiments performed with real students and shows and discuss the obtained results. Finally, in Section 5 our results are summarized and some conclusions are drawn.

2. Theoretical background and related work

In this section we present some theoretical background related to the work presented in this paper and analyze different formal and informal definitions of the concept of *difficulty*. As we will see, it is closely related to the problem of knowledge diagnosis. Elements used for knowledge diagnostic purposes are generically called tasks. Tasks are the most interactive part of an assessment, and their main purpose is to elicit evidences (observables) about proficiencies (unobservables) (Shute, Graf, & Hansen, 2005).

Two main framework will be presented: formal test theories CTT and IRT where tasks are usually simple questions, and where *difficulty* has a clear meaning; and the ITS and AIWBES where the *difficulty* of tasks is defined and used in different ways.

The section continues with a summary of previous work about the estimation of the *difficulty* of assessment tasks either by teachers and/or students, analyzing the alignment between teachers' and students' point of view regarding problem solving complexity and strategies to estimate it. Although this is a very interesting question, we have not found many studies about task *difficulty* estimation. To find relevant studies a wide variety of computerized databases were used including Educational Resources Information Center (ERIC), The ISI Web of Knowledge, ScienceDirect, and Google Scholar. The following keywords were combined: *difficulty level, assessment difficulty, item difficulty, task difficulty, calibration and estimation*. Next, the 'snowball method' was employed and the references in the selected articles for additional works were reviewed, and also those articles that cite the previously found papers.

2.1. Task difficulty and knowledge diagnosis in CTT and IRT

2.1.1. Classical Test Theory (CTT)

CTT was first used at the beginning of the 20th century and has been used ever since. According to this theory, the knowledge (ability or true score) of a student is defined as the expected value obtained by a student in a certain test. Given a student s, who takes a test t, his/her knowledge can be expressed as follows:

$$Y_{st} = \tau_{st} + \varepsilon_{st}$$

where Y_{st} is a random variable representing the observed score of subject s when answering test t. This is also called the test score. It is composed of two parts: the true score (τ_{st}) and the measurement error (ε_{st}). Neither is observable. Y_{st} can be computed from the number of questions answered correctly or any other heuristic. In turn, the true score is a random variable with normal distribution with mean equal to zero and unknown variance.

CTT assumes that true score and error are not correlated. Therefore, if we take two different measurements, the errors we obtain are independent of each other. The error measurement is independent of the true score. In this theory items are characterized by two parameters: the *difficulty*, that is, the portion of students who answered the item successfully, and the *discrimination* factor, whose value is a correlation between the item and the test score.

CTT has several limitations, e.g. the knowledge measurement is strongly linked to test features. This means that when we measure student knowledge, we do not obtain an absolute quantitative measurement of his/her knowledge, but rather a value that depends on the test taken. This makes it very difficult to compare students who have taken different tests. Likewise, item parameters represent features of a certain population, therefore are not generic. As a result, the *difficulty* of an item will strongly depend on the knowledge levels of those individuals whose performance is used to infer the *difficulty* and vice versa.

On the other hand, CTT is easy to apply in several situations (Hambleton & Jones 1993). In addition, unlike other theories such as IRT, this theory has fewer requirements, e.g. it requires fewer examinees. Traditional test-based assessment criteria (percentage of success, score obtained, etc.) are in keeping with this theory.

Download English Version:

https://daneshyari.com/en/article/382569

Download Persian Version:

https://daneshyari.com/article/382569

Daneshyari.com