Chapter 13

A Rule-Based System for Test Quality Improvement

Gennaro Costagliola
University of Salerno, Italy

Vittorio Fuccella
University of Salerno, Italy

ABSTRACT

To correctly evaluate learners’ knowledge, it is important to administer tests composed of good quality question items. By the term “quality” we intend the potential of an item in effectively discriminating between skilled and untrained students and in obtaining tutor’s desired difficulty level. This article presents a rule-based e-testing system which assists tutors in obtaining better question items through subsequent test sessions. After each test session, the system automatically detects items’ quality and provides the tutors with advice about what to do with each of them: good items can be re-used for future tests; among items with lower performances, instead, some should be discarded, while some can be modified and then re-used. The proposed system has been experimented in a course at the University of Salerno.

INTRODUCTION

E-testing, also known as Computer Assisted Assessment (CAA), is a sector of e-learning aimed at assessing learner’s knowledge through computers. Through e-testing, tests composed of several question types can be presented to the students in order to assess their knowledge. Multiple choice question type is frequently employed, since, among other advantages, a large number of tests based on it can be easily corrected automatically.

The experience gained by educators and the results obtained through several experiments (Woodford & Bancroft, 2005) provide some guidelines for writing good multiple choice questions (items, in the sequel), such as: “use the right language”, “avoid a big number of unlikely distractors for an item”, etc.

It is also possible to evaluate the effectiveness of the items, through the use of several statistical
models, such as Item Analysis (IA, 2008) and Item Response theory (IRT). Both of them are based on the interpretation of statistical indicators calculated on test outcomes. The most important indicators are the difficulty indicator, which measures the difficulty of an item, and the discrimination indicator, which represents the information of how effectively an item discriminates between skilled and untrained students. More statistical indicators are related to the distractors (wrong options) of an item. A good quality item has a high discrimination potential and a difficulty level close to tutor’s desired one.

Despite the availability of guidelines for writing good items and statistical models to analyze their quality, only a few tutors are aware of the guidelines and even fewer are used with statistics. The result is that the quality of the tests used for exams or admissions is sometimes poor and in some cases could be improved.

The most common Web-based e-learning platforms, such as Moodle (Moodle, 2008), Blackboard (Blackboard, 2008), and Questionmark (Questionmark, 2008) evaluate item quality by generating and showing item statistics. Nevertheless, their interpretation is left to the tutors: these systems do not advise or help the tutor in improving items.

In this article we propose an approach and a system for improving items: we provide tutors with feedback on their quality and suggest them the opportune action to undertake for improving it. To elaborate, the approach consists of administering tests to learners through a suitable rule-based system. The system obtains item quality improvement by analyzing the test outcomes. After the analysis, the system provides the tutor with one of the following suggestions:

- “Keep on using the item” in future test sessions, for good items;
- “Discard the item”, for poor items;
- “Modify the item”, for poor items whose defect is originated by a well-known cause.

In this case, the system also provides the tutor with suggestions on how to modify the item.

Though item quality can be improved after the first test session in which it is used, the system can be used for subsequent test sessions, obtaining further improvements.

Rule-based systems are generally composed of an inferential engine, a knowledge-base and a user interface. Our system follows this model. The inferential engine works by exploiting fuzzy classification: the items are classified on the basis of the values of some parameters calculated on test outcomes. Fuzzy classification has been successfully employed in technological applications in several sectors, from weather forecast (Bradley et al.; 1982) to medical diagnosis (Exarchos et al.; 2007). In our system, it has been preferred over other frequently used classification methods based on machine learning due to the following reasons:

- Knowledge availability. Most of the knowledge is already available, as witnessed by the presence of numerous theories and manuals on psychometrics.
- Lack of data. Other types of classification based on data would require the availability of large data sets. Once they have been gathered, in such a way to have statistically significant classes to perform data analysis, such methods might be exploited.

The knowledge-base of the system has been inferred from IA and other statistical models for the evaluation of the items.

The system has been given a Web-based interface. Rather than developing it from scratch, we have preferred to integrate the system in an existing Web-based e-testing platform: eWorkbook (Costagliola et al.; 2007), developed at the University of Salerno.

An experiment on system’s performances has been carried out in a course at the University of