Automatic Evaluation of Quiz Test

Dan Laurentiu Lacrama

Elisabeta Bacala

Ovidiu Crista

Emil Sebastian Ciorba

"Tibiscus" University of Timisoara, Romania

Summary: This paper is focused on the automatic evaluation of quiz test using Pattern Recognition methods. The subjects' answers are numerically encoded in a descriptor vector. This vector is then compared with some pre-defined models and the differences are computed using Manhattan distance formula. Evaluation is finally made tacking account of the shortest distances found.

Key words: quiz test, pattern recognition

1. Introduction

The quality assessment quiz test given to each of an university's student during every year's teaching quality evaluation produce a large quantity of individual responses that have to be checked and interpret during a short time in order to elaborate the annual report over that institution's teaching quality level. Hence a lot of work is done to fulfill this task and a method to automatically evaluate the students' level of satisfaction over the education process is highly welcomed.

Any implementation of such a computerized technique of evaluation has to take account of some important characteristics of these quiz tests [1]:

a. Test is complex because it has to evaluate all the characteristics of a modern education process: curricula level of usefulness, teaching materials value and effectiveness, university's infrastructure quality, teachers' skill, administration's competence etc. Therefore tests contain many questions, some of them being relevant for more than one of the above mentioned problems. It is advisable that distinct test sections to correspond to each of the major topics included in the research;

b. Tests must be opportune for studying the satisfaction level with all categories of students (bachelors, master and even PhD.) educated in quite different area of expertise (i.e. computers, low, economics, modern languages etc.). Thus the questions and their set of alternative answers have to be well fitted in order to cover all the diversity currently existing inside a modern university;

c. Some students do not respond entirely earnest to all the questions (even if they are anonymous) or pay not enough attention to the importance of quality assessment and give speedy responses without carefully reading the questions. Consequently, cross-correlations between questions or check points have to be added in order to detect non-valid (i.e. dishonest or superficial) answers.

d. The "don't know / don't answer" response is an alternative existing for each question and must be considered even if it does usually correspond to situations described above;

e. The individual results of this kind of test can be put in a set of standard format sentences like: "the student is very content / somehow content / somehow discontent / very discontent of the educational infrastructure / teaching materials / professors' skill / …";

f. Each of these partial conclusions is important for the corresponding chapter of the quality evaluation, but it is also considered in the global decision over the student's general level of satisfaction. Therefore, we have to keep in the database both partial and global verdicts for each respondent.

g. The number of students in each faculty is various and the department's infrastructure is different according to their level of finance, management quality, staff research performances etc. There are some teaching facilities and activities shared by more than one field of education (i.e. sports infrastructure, multimedia lab etc.). Thus, pure quantitative statistics will sure lead to wrong conclusions. Past experience shows that the final report authors need both the detailed information and the global statistics from the database for detecting the good and the bad parts of the teaching process in the university.

Considering the above features the authors thought that it is workable to use the classification methods employed in the Pattern recognition for conceiving a technique and a computer program able to evaluate the individual test answers and give decisions on the students' opinions over both the particular aspects and the global stage of the education process in the institution.

First step in conceiving such a methodology is to establish a procedure to encode into digital descriptors the following items:

- the value of the answers to the individual questions in the quiz test
- the weight of the questions
- the cross-correlations between different questions
- the check points
- the classes of subjects and their descriptors' vector
- the supplemental parameters (e.g. speed of the response)

For example, a question as "Are you satisfied with the multimedia lab infrastructure?" usually have five alternative answers: very content / somehow content / somehow discontent / very discontent / don't know or don't answer. They should be encoded as shown in the table 1.1.

Table 1.1.

	Answer	**Digital descriptor**
a.	Very content	+1
b.	Somehow content	+0.5
c.	Somehow discontent	-0.5
d.	Very discontent	-1
e.	Don't know / don't answer	0

When the test has questions with different weight this is materialized as a multiplicative factor applied to the above result. Therefore a "somehow discontent" answer (-0.5) to a question weighted with a double value will be equivalent to a "very discontent" (-1) answer at an usual question. This kind of weighted questions will be particularly useful if the analysis and evaluation is made not on the answer list but also over the relations between different responses.

Similarly with the above, the cross-correlations and the check points are implemented as logic operations between answers. If, for example, questions Q3 and Q7 are searching the same aspect but the text logic is reversed in order to discover non-sincere or superficial answers, the computer will execute the operation described in eq. 1.1

if (R(Q3) = = (- R(Q7))) validate(R(Q3)); (eq. 1.1)

where R(Q3) and R(Q7) are the response to the responses to Q3 and Q7 respectively and (- R(Q7)) performs the compensation of the reversed logic of the second question.

Such verifications can imply more than two questions and can implement quite clever traps able to detect non-valid answers inside tests. The best strategy to interpret such "artificial" responses to the questions is to abort it from the automatic evaluation and left the human expert to decide if it is still worth to be tacking into consideration.

The quality quiz tests have as explained before a limited number of probable conclusions. For example, regarding each specific topic (i.e. curricula level of usefulness, teaching materials value and effectiveness, university's infrastructure quality, teachers' skill, administration's competence etc), the respondent can be:

a. Very content
b. Somehow content
c. Somehow discontent
d. Very discontent

There are also possible cases when the subject select not to have an opinion (don't know / don't answer), but those are usually expelled from the interpretation as long as they are just rare exceptions.

The C_i classes of responses for each topic are the same with the above enumeration (i.e. Very content / Somehow content / Somehow discontent / Very discontent), hence the author of the quiz test must only define the models of the descriptor vector V_{Ci} that better describes the typology of those categories of opinions. In the specific case of quality assessment tests, the authors' experience shows that is better to perform a two stage evaluation:

- **Step I:** separate evaluation of each test section researching an individual topic because habitually one subject has different opinions on different aspects (e.g. ***very content*** on the teachers skill and ***somehow discontent*** on the university's educational infrastructure),
- **Step II:** detection of the subject's opinion over the whole educational process in the institution.

The best way to do this is to use a set of already evaluated answers and to extract the state of the features from them. The procedure is called iterative learning and its convergence to the solution is mathematically proved as achievable in the Pattern recognition and Neural networks literature [2][3].

The transposed descriptor vector V_{Xj}^{T} encoding the subject's X_j test responses has the form given in eq. 1.2.

$$V_{Xj}^{T} = \left[w_1 R(Q_1) \quad w_2 R(Q_2) \quad w_3 R(Q_3)... \begin{matrix} 0 & if & unvalidated \\ w_4 R(Q_k) & if & validated \end{matrix} \quad ...w_i R(Q_p)... \quad w_n R(Q_N) \right]$$

(eq. 1.2)

where 1 ≤ k,p ≤ N, $R(Q_p)$ are the answers to ordinary questions, $R(Q_k)$ are the answers to questions which have cross-correlations or check points for validation and w_p are the weights of the questions as established by the sociologist or the psychologist when the test was created.

Additional parameters a_p (e.q. speed of response, subject's temperature etc.) can be added to some or all the questions as multiplicative factors if the test creator do need them and if the computer where the subject type his responses has the devices able to measure them.

Each class of subjects C_i has a predefined descriptor vector V_{Ci}^{T} which has the same dimension as vector V_{Xj}^{T}:

$$V_{Xj}^{T} = \left[R_1^{Ci} \quad R_2^{Ci} \quad ...R_p^{Ci}... \quad R_N^{Ci} \right] \qquad \text{(eq. 1.3)}$$

where 0 < p < N R_p^{Ci}

Therefore the Manhattan distance between the two vectors is easily computed with the formula:

$$D_{XjCi} = \left\| V_{Xj}^{T} \quad - \quad V_{Ci}^{T} \right\| = \sum_{p=1}^{N} \left| R_p^{Ci} - R_p \right| \qquad \text{(eq. 1.4)}$$

and the decision is made detecting the minimum distance D_{XjCi}:

$$X_j \in C_q \; where \quad D_{XiCq} = \min(D_{XjCi}).$$

This means that the set of answers of the X_j subject are closer to the C_q class predefined model than of any other class.

In cases where there is no clear differentiated minimum D_{XjCi} and there are two or three close values, the corresponding classes are given as alternate solutions for a subsequent human evaluation and decision.

After categorizing the subject's opinions on each of the partial topics, the application passes to the second step and estimate his global attitude over the education process quality in the university. This is done using in essence the same procedure described above, but instead of the $R(Q_p)$ in eq. 1.2 the vectors to be compared contain the partial results from Step I.

2. Program implementation

The application AQI was written in Visual C++ paying special attention to make simple and easy to understand Graphic User Interfaces. This was considered very important by the authors because both quiz tests implementation and its application is done by operators not specialized in Computers Science.

Basically the program has three main parts:

a. The application core,

b. The link with the database;

c. The interfaces manager.

In the pre-operational calibration phase the application core computes the model vectors V_{Ci} for every C_i class of answers in the case of all test's sections. The same is done for the global evaluation classes vectors. All these predefined vectors are constructed using a set of learning data which where previously evaluated and classify by a human specialist in sociology or psychology (usually the quiz test's author). Those model vectors are put in the database in order to use them for the later evaluation of the subjects' responses.

During the operational stage the program core is responsible for the actions regarding the subject's opinions evaluation both for each test section and for the whole quiz. Therefore this module performs the operations needed for:

- encoding each individual answer according to the rules given in section I;
- performing validations using the cross correlations and the check points;
- including weights w_p and supplemental parameters a_p;
- constructing the V_{xj} vector for each test section;
- computing the distances D_{XjCi} and detection of their minimum – meaning the detection of the sections' conclusions;
- determining the global opinion of the analyzed subject using the set of partial conclusions from above.

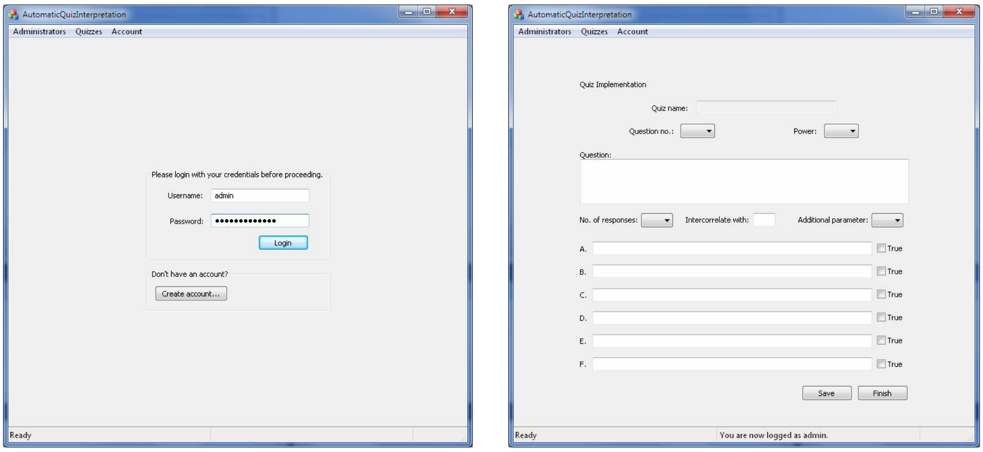
Both the partial conclusions for test sections and the final one concerning the whole quiz are memorized in the MySQL database using the link module. This application part is also responsible for the final statistic computations performed when all subjects' tests are evaluated and categorized.

Inside the Link module the authors did implement basic security mechanisms using the standard facilities provided by SQL. If needed this security policy can be upgraded in order to better guard the tests, the predefined data and the results of the evaluations.

The interfaces manager is the one that govern the operational link with the users, both the quiz authors and tests' subjects. The most important interfaces implemented in AQI are:

- the start window – where the users identify themselves as test's authors or subjects;
- the test implementation window – containing the tools kit needed by the authors in order to create the tests;
- the answers window – where subjects answer to the quiz;
- the evaluation window – shows the evaluation partial and global results;
- the statistics window – contains the statistic tools to process data and the place to display the results.

Some of these Graphical User Interfaces are depicted in Figure 2.1.



Please login with your credentials before proceeding.

Username: admin

Password: ●●●●●●●●●●●●●●

Don't have an account?

Ready

Quiz Implementation

Quiz name:

Question no.:

Power:

Question:

No. of responses:

Intercorrelate with:

Additional parameter:

A. □ True

B. □ True

C. □ True

D. □ True

E. □ True

F. □ True

Ready

You are now logged as admin.

a. b.

Figure 2.1. Graphical User Interfaces of the AQI

a. Start window

b. Quiz implementation window

3. Conclusions

The application AQI is still in the implementation phase, but some experimental results show that it has the potential to became a useful tool in dealing with the difficult task of processing the quiz tests results. Anyway the use of Pattern recognition methods in order to evaluate the opinions expressed by the quiz tests subjects proved to be a good idea.

The preliminary experimental results over a set of quiz test's answers show that the automatic classifications are accurate enough to be used as a faster and suitable instrument to help substantially reduce the human effort consumed to evaluate, organize and statistically process the answers.

The errors are highly dependent on the selection of the threshold which separate the minimum distance from the other close value distances. If this threshold is small, a lot of single value decisions are taken (more than 94% of the studied cases), but the amount of errors is more than 5%. If the threshold is wider the single decision percentage fall to about 82%, but the errors decrease to less than 1% and their effect on the global statistics became insignificant. This of course increases the number of cases needing a human final decision, thus the speed is not as big as before. Even so, the authors believe that the second situation is preferable and selected this approach.

Another possibility that has to be further refined is the learning procedure, where a better strategy could provide useful in the later operational stage of the application. For example, the refresh of the classes' models after each classification will add more flexibility to the method and will probably provide more accuracy to the final results [4].

Bibliography

[1] Karnyanszky T.M., Lacrama L. D., Luca L, Iacob I., “Teacher's Evaluation – a Component of Quality Assessment System”, Anale. Seria Informatica Vol. 6, Fasc. 1, May 2008

[2] Baird H.S. "Recognition Technology Frontiers", Pattern Recognition Letters, North Holland, vol.14, No.4, April 1993, pp.327 334

[3] Schalkoff R.J., "Pattern Recognition: Statistical, Structural and Neural Approaches", Whiley, 1992

[4] Hastie T., R. Tibshirani R., “Discriminant Adaptive Nearest Neighbour Classification”, IEEE Trans. on PAMI, vol.18, No.6, June 1996