Content Analysis of Student Assessment Exams

Priscila da Silva Neves Lima
Instituto de Informática
Universidade Federal de Goiás
Goiânia - Goiás - Brasil
Email: priscilasilva@inf.ufg.br

Ana Paula Laboissière Ambrósio
Instituto de Informática
Universidade Federal de Goiás
Goiânia - Goiás - Brasil
Email: apaula@inf.ufg.br

Igor Moreira Félix
Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo - São Paulo - Brasil
Email: igormf@ime.usp.br

Jacques Duílo Brancher
Departamento de Computação
Universidade Estadual de Londrina
Londrina - Paraná - Brasil
Email: jacques@uel.br

Deller James Ferreira
Instituto de Informática
Universidade Federal de Goiás
Goiânia - Goiás - Brasil
Email: deller@inf.ufg.br

***Abstract*—This Full Paper presents a methodology for classification and analysis of test questions by domain content. Today, performance indicators are used to assess the quality of education. Often, these indicators are obtained through large scale tests applied by public assessment organizations, and their results contribute to the quality analysis of institutions, courses and students. To contribute to a more comprehensive educational assessment analysis, which considers the program content of the test items, we propose an automatic domain classification of exam questions and procedures to report results that allow a better comprehension of student performance based on the domain knowledge represented in the exam. The process starts by automatically classifying the test questions into knowledge domains using dictionaries. Using this classification, the mechanism separates the student's results by domain and generates reports that give a better understanding of the test structure and student outcome by knowledge content. This data can be used by data mining algorithms to predict student outcome in the exam based on their results during the course, allowing preventive actions to be taken to address the problems. A prototype has been implemented and is being used to classify and analyze data from the Computer Science exams of the National Examination of Student Performance (ENADE) – an exam applied in Brazil to assess the quality of all higher education institutions. Results have shown that relevant information can be extracted using this methodology that allows to identify test's domain emphasis, areas where students do better or worst, and how the exam reflects course structure.**

Keywords: ***education, student performance, educational analysis, enade, exam questions*.**

I. Introduction

One of the main objectives of educational evaluation is to ensure the quality of education. In this context, it is essential to obtain indicators for quality control of educational institutions. The measure of the quality of the institutions can be defined by indicators of how much each contributes to the development of academic skills, professional competencies and increase in knowledge of their students. In democratic societies, such evaluations serve as instruments of accountability, used to examine whether organizations, to which certain roles are determined, have fulfilled their obligations [1].

These indicators can be obtained through large-scale tests applied by public evaluation organizations, and their results contribute to the analysis of institutions, courses and students. Educational assessment systems collect and store detailed information, concerning the tests, the students and their results, and generate reports that inform and sustain decisions at several levels [2]. These reports are usually the result of descriptive statistics analysis that describe and summarize the information collected, reporting the s obtained by students, often classifying them by socioeconomic parameters such as gender, age, etc.

Content analysis of exams are not always explored in official evaluations, often due to the difficulty of generalizing analysis procedures and standardizing a methodology applicable to all domains. Eventually, punctual content analyses are made by researchers interested in perfecting specific exams [3] [4], [5], in improving pedagogical projects [6], in training teachers [7] [8] [9], but do not present a generalized methodology and are not part of a systematized analysis like the one proposed by this research.

To contribute to a more comprehensive educational assessment analysis, which considers the program content of the test items, this research aims to define a methodology that allows classification of exam questions into knowledge domain areas represented in the tests and to identify analysis that can be performed from this classification. Using this classification, the mechanism separates student's results by domain and generates reports that provide a better understanding of the test structure and student outcome by knowledge domain.

A prototype, using this methodology, has been implemented and is being used to classify and analyze data from the Computer Science exams of the National Examination of Student Performance (ENADE) – an exam applied in Brazil to assess the quality of all higher education institutions.

Analyzing the content of the exam questions allows a better understanding of the exam structure and to know the content considered most important by the evaluation organizations. It also provides information that allows to improve the test

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

and gives the market a clearer view of the professional they receive. From the point of view of educational institutions, they can analyze their curricula and identify areas in which their students stand out or are deficient.

The information provides decision makers with a new perspective that can have a wide impact. As procedures can be defined for reporting results that allow you to understand the student's performance based on the domain knowledge represented in the exam. Furthermore, the use of data mining techniques may help predict the student's result in the exam based on their results during the course, allowing preventive actions to be taken to solve the problems.

The article is structured in 6 sections. Initially, the methodology for content analysis is discussed. Next, the National Assessment of Student Achievement (ENADE) Exam is presented. Section IV shows the applicability of the methodology using ENADE data for a computer science course, giving rise to the implemented SysEnade prototype . Section V presents the results obtained applying this methodology to the Computer Science course, providing an example of the information that can be extracted. Finally, conclusions on the methodology are presented.

II. Methodology

To perform exam content analysis one must understand the exam structure and the domain knowledge being verified by the exam. Often, widely applied tests release documents stating the content that will be evaluated, as well as its structure and type of questions. Information is also given as to how scores are calculated, including maximum score for each question and weight in the final grade. The proposed methodology takes into account all these factors and can be applied to any large-scale assessment. It is structured in three phases according to Figure 1. Each phase is detailed below.

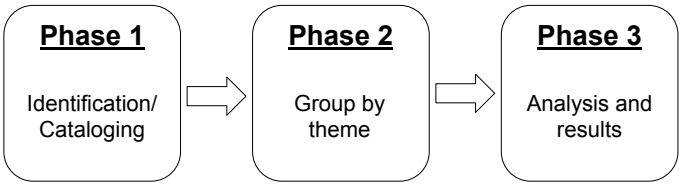

Figure. 1: Phases of the methodology used

Phase 1 - Identification/Cataloging

This phase includes the identification of the exam questions to be analyzed and their classification in previously defined themes. Identifying exam questions means extracting the wording of each exam question (including choices in multiple-choice questions) and identifying how it is graded and its weight in the final exam score. Any other information provided by the examiners about individual questions may also be taken into account, for example percentage of hits and misses. Each question must then be classified in one of the predefined Themes.

These themes are non-overlapping and complementary subsets of the programmatic content of the exams and may be created by domain knowledge experts. These themes correspond to the sub-domains in which analysis will be undertaken. Therefore, they must be specific enough to be representative of the important content domains, but generic enough to allow several questions do fall into that theme.

Classification of the exam questions into themes may be done manually by an expert, or automatically using text mining techniques. For an automatic association through algorithms, without the use of artificial intelligence, each theme is identified by a dictionary of keywords. Association is performed by checking the number of occurrences of dictionary words for each theme in the questions. The question is associated to the theme with the greater number of occurences. Other more sophisticated approaches can also be used.

Phase 2 - Group by theme

In this phase, we must calculate, for each student, a grade for each theme. This is done by identifying the student's result for each exam question, grouping the questions by theme and calculating a score obtained by that student in that theme. If questions have different weights in the test's final grade, it may be advisable to maintain these weights when calculating the scores by theme.

How this grade is calculated can be defined according to the situation, but they must be normalized as to allow comparison between themes. An option is to use a weighted arithmetic mean, that is, for each theme, multiply the student's grade by the weight of the question, add the result for all the questions in the theme and divide by the sum of weights of the questions related to the theme.

Other aspects of data transformation must also be taken into account, such as missing values and outliers. If the dataset has not been pre-processed to treat these aspects, this must be done before proceeding to analysis. Implementation of this procedure will depend on the adopted approach and how the data containing the grade for each question is stored.

Phase 3 - Analysis and results

In this last phase of the methodology, the data is already transformed and grouped. Data analysis techniques such as descriptive and inferential statistics can be applied. Descriptive statistics includes a set of measurements and graphical representations, which summarize and describe a data set. Inferential statistics encompasses statistical tests, which make possible conclusions about the target population, based on results obtained in the selected sample [10].

Another powerful analysis technique, data mining proposes to abstract relevant information from large volumes of data, allowing to establish relationships and interpretations that serve as the basis for the construction of new knowledge [11]. Data mining techniques also provide predictability of the outcome of a future observation.

What type of analysis is best will depend on users' expectations and the available data. As most often happens in reports,

analysis starts with descriptive statistics that allow users to have an overview, or picture, of the current state. The use of graphs and tables help to rapidly assess how groups are faring. The same applies to the analysis of student outcome by theme. With a better understanding of your dataset, several inferential statistics tests can also be used to determine correlations between themes and groups. If the students that took the test are a representative random sample of the total population, thru inferential analysis conclusions can be made about the overall population based on the sample results. Information of how students fare in each theme and how themes relate to each other can be analysed.

Taking it a step further, data ming techniques can be used to predict student outcome. By incorporating data about how students scored by theme, predictions can be made using this information. If results in the test reflect academic results, this type of analysis will allow to predict exam outcome based on student performance for example.

To validate this methodology and address specific implementation aspects, a prototype has been developed that applies the methodology to the computer science ENADE exam. This prototype is called SysEnade and considers all the exam characteristics.

III. National Assessment of Student Achievement - ENADE

National Assessment of Student Achievement (ENADE) – is an exam used in the assessment of undergraduate programs in higher education institutions throughout Brazil. It consists of an exam administered to students who are entering or finishing their undergraduate courses. The programs are grouped in three representative areas (Health and Agriculture, STEM, Social Sciences) and each year one group is assessed, meaning that programs in these areas are assessed every three years. Each program has its own test, with domain specific questions to evaluate how students perform nation wide, which allows comparisons between equivalent courses offered in different institutions.

ENADE is comprised of two instruments: a socioeconomic questionnaire, used to define student profile, and a test, composed of objective (multiple-choice) and essay type questions, divided into two parts. The first part, called "general formation (GF)", presents itself as a "common component" to the tests of the different domains being evaluated in that year, investigating general skills and knowledge expected from undergraduate students independent of their course. The second part, called "specific component (SC)", contemplates the specificities of each domain, both in the field of knowledge and in the skills expected for the professional profile. At the end of the test book, there is a questionnaire on test perception that investigates the students' perception of their trajectory in the course and in their educational institution, through objective questions that explore the social function of the profession and the fundamental aspects of professional formation. The specific component of the tests is defined by the guidelines and matrices prepared by the domain's assessment advisory committee and the ENADE general assessment advisory committee. These matrices comprise content in accordant to the national curricular guidelines for undergraduate courses.

The test structure is the same for all courses and contains 40 questions: 10 general formation questions (2 being essay type questions) and 30 specific component questions (3 being essay type questions). Student's final score is the weighted average obtained in each part, with 25% for the general formation, and 75% for the specific component [12]. However, the calculation of the grades also considers the nature of the question, assigning different weights to objective and essay type questions. Table I details the weights of questions in each component.

Table I: Distribution of questions and weights in ENADE

Components	Number of questions	Weight of questions	Weight of components
GF: Essay	2	40%	25%
GF: Objective	8	60%	
SC: Essay	3	15%	75%
SC: Objective	27	85%	

The first ENADE was held nationwide on November 6, 2005, with the evaluation of 20 knowledge domains [13]. Since then it has evolved in how its structured and the knowledge domains it tests. In this study we have chosen to analyse results for the Computer Science course students. In Brazil, this course has a strong mathematical and theoretical basis, and is aimed at students that want to develop and research computer science technology. The computer science domain was evaluated in the years 2005, 2008, 2011, 2014 and 2017, but data for 2017 was not available at time of writing of this paper.

As expected, several adjustments were made during these years, and had to be taken into account when selecting the data to be used. In the 2005 and 2008 versions, there was one test for Computer Science that included groups of students selected from the Information System, Computer Science and Computer Engineering courses, selected by sampling, who were at different stages of their undergraduate course. A group, considered "entering" the course, were at the end of their first year; and another group, considered "finishing", were at the end of the last year of the course. The two groups of students were submitted to the same test.

In the 2011 version, students in the computing domain who took the test, belonged to undergraduate degrees in Computing, Computer Science, Computer Engineering and Information Systems. In 2014, there was the dismemberment of the computing domain and each course had a distinct specific test. Therefore, for the 2014 version, the tests used in this study were exclusively applied to bachelor degree students in computer science. In 2011 and 2014, only concluding students took the tests.

IV. SysEnade

SysEnade is a prototype developed for analyzing ENADE data. The prototype is being developed in JAVA [1]. MySql [2] is used to store the questions, themes and dictionaries.

As a case study, information on Computer Science students was used. However, as all the exams in other domains follow the same structure, this system can be easily adapted to analyze ENADE data from all higher education courses offered in Brazil. It mainly involves defining the themes and the associated dictionaries for that course.

Micro-data on students' results is provided by the National Institute of Educational Studies and Research Anísio Teixeira (INEP) [3]. In addition, copies of the exams and oficial reports emitted by INEP were obtained from the INEP website on Higher Education/ENADE [4].

SysEnade was developed using the methodology proposed in section II. How the methodology was adapted to the ENADE exam is specified in subsection IV-A. Subsection IV-B introduces SysEnade's architecture, comprising the domain model and the component diagrams. It is important to emphasize that SysEnade was created specifically for ENADE, however, its architecture can be adapted according to the structure of other tests, extending its applicability.

A. Methodology Applied to ENADE

This methodology was applied according to the general structure of section II comprising the three phases described. For the ENADE context, each phase was divided in three steps as presented in Figure 2.

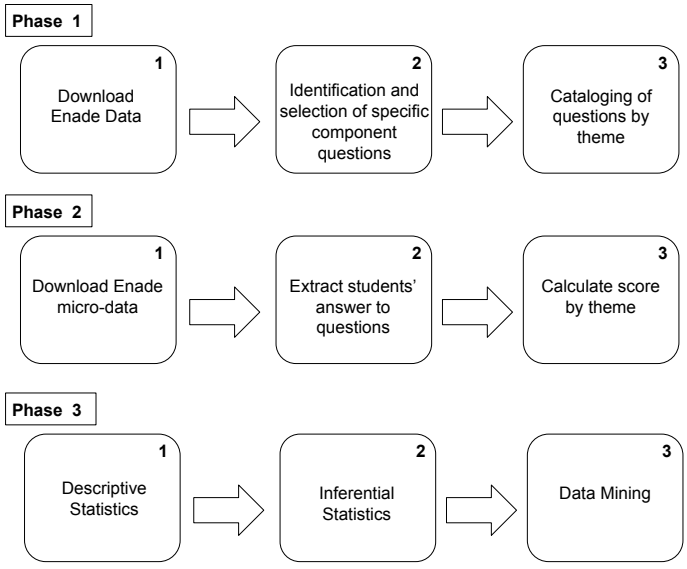

Figure. 2: Methodology phases applied to ENADE

[1]Available in: https://www.java.com/en/. Accessed: 05/01/2018.
[2]Available in: https://www.mysql.com/. Accessed: 05/01/2018.
[3]Available in: http://portal.inep.gov.br/microdados. Accessed: 04/18/2018.
[4]Available in: http://portal.inep.gov.br/web/guest/enade. Accessed: 04/18/2018.

All data used in this research are open and made available by INEP, which is responsible for promoting studies, researches and evaluations about the Brazilian educational system [14]. This research focused students from bachelor degrees in computer science, who were entering or finishing their courses. In the following, each phase will be described, detailing exam characteristics and how they were incorporated in the system.

Phase 1 - Identification/Cataloging

Step 1) Download Enade Data: ENADE Computer Science exams for 2005, 2008, 2011 and 2014, were downloaded from the INEP/ENADE website [5]. Reports provided by INEP for each exam were also downloaded. These reports contain statistical analysis of students' profile and outcomes. Based on these results, facility and discrimination indexes are calculated for each question and detailed in the reports.

The facility index is calculated taking into account the correctness index of each question, ie., the percentage of students that got the question right. The correlation between correctness index and facility index is described in Table II.

Table II: Facility Index

Correctness Index	Facility Index
>= 0,86	Very easy
between 0,61 e 0,85	Easy
between 0,41 e 0,60	Medium
between 0,16 e 0,40	Difficult
<= 0,15	Very Difficult

To be able to evaluate the students of a course, a question must have more hits by students who had good performance than those who had bad performance. To measure this power of discrimination of the question, ENADE uses the biserial point correlation. Questions with a weak discrimination index, with values less than or equal to 0.19, are annulled in the final score [15]. Table III presents the classification of questions according to their power of discrimination.

Table III: Discrimination Index

Discrimination Index	Classification
>= 0,40	Very good
between 0,30 e 0,39	Good
between 0,20 e 0,29	Medium
<= 0,19	Weak

Step 2) Identification and selection of specific component questions: As previously discussed, only in 2014, Computer Science courses had a separate exam. In previous years, different areas of computing shared the same exam. However, the set of questions students from each course had to answer varied. While some questions were shared by more than one course, others were specific. Therefore, for each exam, those questions answered by the Computer Science students were identified and selected for further treatment.

[5]Available in: http://portal.inep.gov.br/web/guest/provas-e-gabaritos3. Accessed in: 03/25/2018.

Step 3) Cataloging of questions by theme: Each question was associated to a theme of computer science. Themes were defined based on the reference domain contents published by the INEP domain advisory commission, using the national curricular course guidelines, approved and instituted by the National Education Council (CNE) of the Ministry of Education (MEC). 11 themes were identified and used in the cataloging process: (1) Algorithms, Data Structure and Programming; (2) Computer Architecture and Digital Circuits; (3) Database; (4) Computer Graphics and Image Processing; (5) Software Engineering; (6) Ethics, Computer and Society; (7) Artificial Intelligence; (8) Formal Languages and Automate, Compilers and Computability; (9) Mathematical Logic, Discrete Mathematics, Statistics and Graphs; (10) Computer Networks, Distributed Systems and Telecommunications; and (11) Operating Systems.

For an automatic association of questions to themes, each theme is identified by a dictionary of keywords. Association is performed by checking the number of occurrences of dictionary words for each theme in the questions. The question is associated to the theme with the greater number of occurences.

Output for phase 1: This phase generates a file containing for each selected question, the year of the exam, the question, its number in the exam, its facility index, its discrimination index, type of question (objective or essay), the theme to which it was associated, and if it was used in the final score or if it was annulled.

Phase 2 - Group by theme

Step 1) Download ENADE micro-data: ENADE micro-data, for the years 2005, 2008, 2011 and 2014, was downloaded through the INEP website. These files contain data for all students who completed ENADE in that year. Thus, it was necessary to extract from these worksheets the data referring to the computer science students. The number of participants varied from 8000-14500 per year, yielding a total of 47000 records in the computer science dataset.

This was done using information contained in the worksheets themselves, identifying the domain, and sub domains where appropriate, of the student's course. Students that did not complete the test or did not answer the social-economic questionnaire were also excluded. This way, entries with missing values were eliminated.

Step 2) Extract students' answer to questions: The ENADE micro-data worksheets contain a column with vectors that indicate the correctness of the objective part of the specific component questions in the exam. Each position in the vector identifies a specific component question in the exam. The number of the question in exam identifies its position in the vector. The possible values for each position in the vector are: 0 = Wrong, 1 = Right, 8 = anulled by the commission, 9 = anulled by the discrimination index. When a student does not answer a certain question, it is considered a wrong answer, that is, 0 is assigned to the question. Essay questions have separate columns containing the score for that question. Scores vary from 0 to 100.

Step 3) Calculate score by theme: Having identified the score the student obtained for each objective question in the exam taken, it is possible to calculate a raw score for each theme in that exam by adding the number of questions in that theme they answered correctly and dividing by the number of questions in that theme. This information is stored in a new column added to the spreadsheet. Anulled questions are disconsidered in this calculation. Another column was used to store the student's average in the essay questions. If no essay question was associated to a theme, the value 0 was atributed to all students. Finally a column was also added containing the final score for that theme calculated using the weighted average considering 85% for the objective questions and 15% for the essay questions, adopting the same proportion used to calculate the final score in the exam. Each new column is a new attribute that can be used in the analysis phase.

Output for phase 2: This phase generates a general spreadsheet where each line contains the year of the exam, selected information about the student from the original micro-data file, and information about theme scores. This file compiles the information obtained from all the downloaded exams, including socio-economic data, institution where the student did the course, etc. As the micro-data exam files had different structures, and in some cases different information, specially in the questionnaires answered by the students, an integration process was undertaken as to unify the information and create a file that could be used in the analysis phase.

Phase 3 - Analysis and results

Step 1) Descriptive Statistics: A series of descriptive statistic procedures were defined. They include mode, median, mean, frequency, interquartile range, and standard deviation. These will be used to analyse the exam structure and student outcome.

Step 2) Inferential Statistics: Inferential analysis allows us to verify the correlation between variables, differences between groups or differences in different moments in time [16]. Tests that will be used include Pearson Correlation Coefficient and Spearman Correlation Coefficient for correlation analysis, and T-test and ANOVA for differences between groups and moments in time.

Step 3) Data Mining: Data mining techniques will be used mainly to predict outcomes. Several classification algorithms have been developed, and the choice of the best classifier for a problem is defined not only by the accuracy obtained, but also by how the classifier returns the final result to the users. Some, like Decision Trees, allow users to understand the impact of variables in the outcome. Other, like MultiLayer Perceptron, are black-boxes. Frequently used algorithms include *C45*, *J48*, *Naïve Bayes*, *Linear Regression*, and *KNN* [17] [18] [19] [20].

Output for phase 3: This phase generates a report containing the results obtained from the different analysis procedures applied to the transformed data.

B. SysEnade development

The domain model is a visual representation of real-world conceptual classes in a domain of interest. Using UML nota-

tion, the domain model for SysEnade is illustrated with a set of class diagrams in which no operations are defined. Figure 3 shows that a student takes an ENADE exam, that must be processed in at least one way. The exam contains one or more questions, and each question is associated to one theme. Questions can be a multiple choice or an essay type question. The domain model highlights the structure of the exam and guided the definition of the SysEnade architecture.

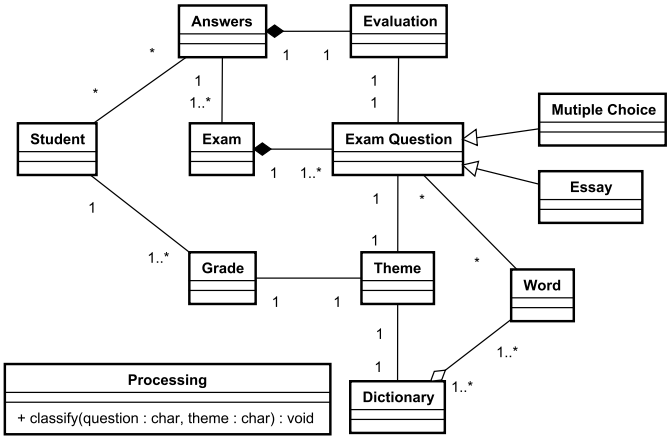

Figure. 3: Domain Model - Class Diagram

SysEnade development is presented as a activity diagram. It has four main components that fit the three phases defined in the general methodology defined. Figure 4 describes the development in a comprehensive way, that can be applied in other similar purpose tools.

The four main components are: Acquisition, Preparation, Analysis and Presentation. Each component has specific elements and functions within the architecture, and these will be detailed below.

Acquisition

Consists of introducing into the system the outside data it needs. This includes the list of themes with their dictionary, the exam questions and exam data.

Theme: For the ENADE exam, the theme list was based on the reference contents established by the INEP domain advisory commission, defined based on the national curricular guidelines of the courses, approved and instituted by the National Education Council (CNE) of the Ministry of Education (MEC).

Dictionary: Each theme has a dictionary that will be used in the process of cataloging the exam questions. For ENADE, the dictionary was created from keywords related to the knowledge domains. However, the dictionary can be built in other ways, for example, through questionnaires to specialists.

Micro-data: Demographic, socioeconomic, and academic data refer to the students who took the exam. In ENADE, these data are made available by the INEP website in ASCII format in the .csv extension (Spreadsheet).

Exam Questions: extract from the exam the questions that will be analyzed. ENADE is made up of general and domain specific questions. In this case, it was necessary to separate the specific component questions and insert them into the system manually.

Preparation

This component is responsible for manipulating the data entered in the acquisition component.

Cataloging: Associates the exam questions with the themes provided. This is being done automatically using dictionaries. The algorithm reads the question and compares it with the theme dictionary. Association is established with the theme containing the greatest amount of words appearing in the question.

Transformation: This is a selection of data for analysis. Only data that comprises the scope of the analysis is filtered and processed. In this case, the focus was on the computer science course. The worksheet provided by INEP provides data for all courses that take the exam that year. Therefore, it was necessary to create filters to extract only information from the students of the computer science courses. Other filters were also used to select and integrate information contained in the data from the different years as they are not structured and named exactly the same.

Score by Theme: This is the calculation of each student's score according to the theme. For ENADE, the grade was calculated individually for the 11 themes, applying the weights used in the exam (85% for the objective part and 15% for the essay part). Only questions from the specific component of the exam were used. It is important to observe that this is specific to ENADE and may be different for other tests. Calculations must be done obeying the rules of the test being analyzed.

Analysis

This component is responsible for analyzing the selected and transformed data. Analysis include statistics and data mining procedures.

Statistics: Statistical analysis was performed using several libraries. JFreeChart [6], an open source framework for creating a wide variety of both interactive and non-interactive graphs. JFreeChart supports a variety of graphics, including mixed graphics. Commons Math [7], a library of lightweight, self-contained mathematics and statistics components addressing the most common problems not available in the Java programming language or Commons Lang. JRI [8] a Java/R Interface, which allows to run R inside Java applications as a single thread. Basically it loads R dynamic library into Java and provides a Java API to R functionality.

[6]Available: https://www.gnu.org/software/pspp/http://www.jfree.org/jfreechart/. Accessed: 04/03/2018.

[7]Available: http://commons.apache.org/proper/commons-math/. Accessed: 05/01/2018.

[8]Available: http://www.rforge.net/JRI/. Accessed: 05/01/2018.

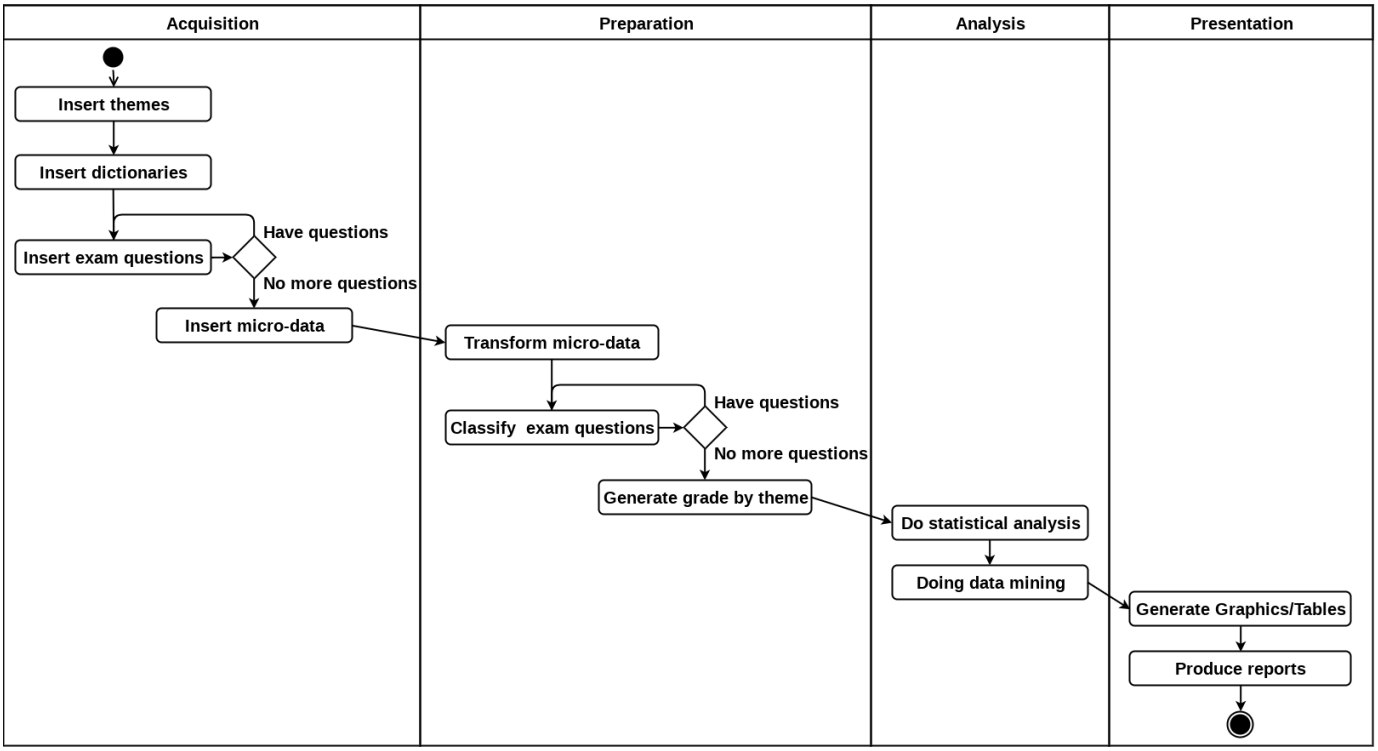

Figure. 4: Activity Diagram - UML

Data Mining: Data mining in SysEnade is performed using WEKA [9]. It is an open source software, developed in 1993, at the University of Waikato, New Zealand [21]. WEKA has an API that enables it to be incorporated into any application allowing data mining tasks to be carried out in an automated way from the software itself.

WEKA uses ARFF (Attribute-Relation File Format) files as input. Furthermore, depending on the algorithm that will be used, other information, such as class attributes must be defined. Thus, the selected data must be transformed and adapted to be used.

Presentation

Final component in development, responsible for viewing the tool's results information.

Reports: Reports relating to the knowledge domain (user defined) are presented. Reports are composed of a brief description of the analysis being done and how results may be interpreted, followed by the analysis output obtained from running the appropriate algorithm.

Graphics/Tables: Tables and graphs are elements that make up the reports. These are inserted as to facilitate the interpretation of the reports, making access faster and improving the visualization of the information.

V. ENADE RESULTS

This section describes the results obtained from SysEnade for the Computer Science course. It serves as an example of the type of information that can be obtained from a theme based analysis.

In total, the four ENADE exams have 120 computer science specific domain questions, being 109 objective and 11 essay-type questions. Each exam is composed of 27 objective and 3 essay questions, except for 2005, with 28 objective and 2 essay questions.

There is a predominance of the theme Algorithms, Data Structure and Programming, with 22 % of the total questions, followed by Formal Languages and Automate, Compilers and Computability, with 14%. On the other hand, the themes that were least present among the questions analyzed are Ethics, Computer and Society; and Artificial Intelligence, both with 3%, followed by Computer Graphics and Image Processing, with 4%. Comparison of this distribution with the number of hours allocated to each theme in the Computer Science course in our university showed that they are closely related.

The research also showed that the theme of Ethics, Computer and Society, presents 100% of the questions classified with average facility index, that is, more than 40% of students have got the questions of Ethics, Computer and Society right. The themes Computer Graphics and Image Processing, and Artificial Intelligence presented questions with index difficult and very difficult, that is, a maximum of 40% of the students were able to answer these questions, and the very difficult questions were correctly answered by a maximum of 15% of the students. Other themes such as Database and Software Engineering had a more homogeneous distribution of facility indexes.

[9]Available: http://www.cs.waikato.ac.nz/ml/weka/. Accessed: 04/06/2017.

As for the discrimination index, questions for the theme Ethics, Computer and Society presented the best power of discrimination, with 100% of the questions classified with the very good discrimination index. Artificial Intelligence had 50% of the questions classified as weak discrimination. 17 objective questions were annulled by the discrimination index criterion, 1 objective question (2011) was disregarded by the advisory committee of the computer domain due to problems in the question, and 1 essay question (2005) was annulled because the question wording contained inaccuracies that hindered the solution. The largest number of questions canceled (4 questions) is related to the Formal Languages and Automate, Compilers and Computability theme. Percentage wise, themes with the greatest percentage of annuled questions were Operating Systems (33%), followed by Computer Architecture and Digital Circuits (27%).

An interesting observation when analysing the correlation between the Facility and Discrimination indexes, is that questions with facility indexes "easy" and "medium" have "good" or "very good" discrimination indexes. This means that questions that were correctly answered by more than 40% of the students were better in evaluating than those few students answered.

Analysis of the essay questions showed that the entering students had difficulty in answering the proposed questions, with a high index of grades zero. The exception appears in computer architecture and digital circuits where the percentage of grades (0) was only 71%. This is because these students are starting the course and do not have previous knowledge about the subjects. The essay questions that obtained 71% of grades (0) were about contents related to computer organization and memory hierarchy, usually seen in the first year of the course.

The finishing students, however, obtained better results, as expected, with improvements around 10 %. However, the percentage of zero (0) grades remains high. Operating Systems, a theme that was not addressed in an essay question in the exams taken by entering students, had 98 % of grades 0 among the graduating students.

VI. Conclusion

Although large-scale evaluations are already part of the education scenario [22] [23], it is not possible to create highly accurate large-scale evaluation processes. However, it is the role of all society, and not just specialists, to contribute to improvements in the quality of education. This implies thinking of covering all the dimensions that make up the concept "Quality of Education". One way of doing this is to carry out research that will contribute to increase the efficiency and effectiveness of these evaluations.

This paper presents a methodology for the analysis of exams by content knowledge, offering decision makers another perspective. The methodology allowed the development of an architecture that can be adapted to any evaluation exam. This enables an automatic analysis of tests and educational data.

A bibliographical survey did not uncover other studies on the automatization of content analysis of exam questions. Therefore, this study is innovative in that sense. By exploring the content of exams, it is possible to identify other types of analysis focused on the knowledge domain, including which domains are predictive of success in the final exam result.

The results presented by the research in the case study demonstrated that the methodology allows the analysis of the content of a test. The analysis was based on 120 questions from four exams, classified in eleven themes. This classification highlights the exam structure, identifying themes considered more important and how this structure relates to the curricular matrices of undergraduate courses. That is, it is possible to verify that what is being covered, in terms of quantity, in the test, is in accordance with the hour load of what is being offered in the course.

The analysis also shows how students perform in different themes. Themes or questions with low discriminating indexes or very difficult facility indexes should be given greater attention. They may indicate that students are not learning what they should about that particular theme, or it may indicate that the questions being asked are not adequate for the knowledge level of the students or in accordance with the content being evaluated.

The analyzed data point to the most recurrent themes, Algorithms, Data Structure and Programming; Formal Languages and Automate, Compilers and Computability; and Mathematical Logic, Discrete Mathematics, Statistics and Graphs. Although no comprehensive study has been done on the subject, there are indications that the ENADE exams in computing reflect the importance of each theme, or the emphasis given by the curriculum in the distribution of the disciplines. This may be further investigated.

The implementation of the methodology for the ENADE exams showed it is viable and adequate for the proposed objectives, specially for the analysis of exams that span many years and are taken by a significant number of students as is the case of ENADE. Although the prototype was developed for the computer science course, it can easily be adapted to be used by other courses evaluated by ENADE, making it extremely relevant in the Brazilian educational context. However, for the analysis of other tests, that may have completely different structures, the methodology may be reused, but a new, adapted implementation would be necessary.

References

[1] R. Primi, C. S. Hutz, and M. C. R. Da Silva, "A Prova do ENADE de Psicologia 2006: Concepção, Construção e Análise Psicométrica da Prova," *Revista Avaliação Psicológica*, vol. 10, no. 3, pp. 271–294, 2011.

[2] C. L. Dias, M. De Lourdes, M. Horiguela, P. S. Marchelli, U. São, and M. Resumo, "Políticas para avaliação da qualidade do Ensino Superior no Brasil: um balanço crítico," vol. 32, no. 3, pp. 435–464, 2006. [Online]. Available: http://www.scielo.br/pdf/ep/v32n3/a02v32n3.pdf

[3] S. A. Razmjoo and H. H. Tabrizi, "A Content Analysis of the TEFL M . A . Entrance Examinations (Case Study : Majors Courses) Seyyed Ayatollah Razmjoo and Hossein Heydari Tabrizi," *Applied Linguistics*, vol. 14, no. 1, pp. 159–170, 2010. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ920511.pdf

[4] "XAT 2018 Analysis: In-depth XAT Exam Analysis by MBA expert | MBAUniverse.com," 2018. [Online]. Available: http://www.mbauniverse.com/xat/exam-analysis.php

[5] V. A. H. A. Hypolito, “Uma análise do conteúdo das provaas da OBMEP Nível 3,” *Technical report*, 2016.

[6] J. P. D. C. Costa and M. I. Martins, “O ENADE para a licenciatura em física: Uma proposta de Matriz de Referência,” *Revista Brasileira de Ensino de Física*, vol. 36, no. 3, pp. 3401 – 3401/9, 2014. [Online]. Available: www.sbfisica.org.br

[7] I. C. M. De Lara, “Exames Nacionais e as "verdades" sobre a Produção do Professor de Matemática,” *Tese (Doutorado) - Universidade Federal do Rio Grande do Sul*, 2007.

[8] S. Novossate, “O ENADE e os Documentos Curriculares: Um Estudo sobre a Formação de Professores de Biologia,” *Dissertação (Mestrado) - Universidade Federal do Paraná*, 2010.

[9] L. Schwengber, “Exame Nacional de Desempenho de Estudantes: Problematizando Verdades sobre a Formação do Professor de Matemática,” *Dissertação (Mestrado) - Universidade de Santa Cruz do Sul*, 2013.

[10] D. C. Howell, *Statistical Methods for Psychology*, 7th ed., Belmonte: Wadsworth, 2010.

[11] M. G. Kulkarni, M. C. Rampure, and M. Yadav, “Understanding Educational Data Mining,” *International Journal of Electronics and Computer Science Engineering*, pp. 773–77, 2013.

[12] “Nota Técnica Nº 2/2017/CGCQES/DAES,” *INEP*, 2017. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362006000300002&lng=pt&tlng=pt

[13] INEP, “Relatório de Síntese - Computação,” *Disponível em: <http://portal.inep.gov.br/relatorios>. Acesso em: 17 de abr. 2017.*, 2008.

[14] S. O. d. Fonseca and A. A. Namen, “Mineraçao em Bases de Dados do INEP: Uma Análise Exploratória para Nortear Melhorias no Sistema Educacional Brasileiro,” *Educação em Revista*, vol. 32, no. 1, pp. 133–157, 3 2016.

[15] INEP, “Relatório de Síntese 2005,” *Disponível em: <http://portal.inep.gov.br/relatorios>. Acesso em: 01 de mai. 2018.*, 2005.

[16] C. Martins, *Manual de análise de dados quantitativos com recurso ao IBM SPSS*, 2011.

[17] S. Anupama Kumar and M. N. Vijayalakshmi, “Efficiency of Decision Trees In Predictiing Student’s,” pp. 335–343, 2011.

[18] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining*, 2008, vol. 14, no. 1.

[19] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” *International Journal Of Computer Science And Applications, ISSN: 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.

[20] M. Pereira, “Mineração de Dados - Conceitos , Aplicações e Experimentos com Weka,” *Dados*, vol. 11, pp. 10–18, 2009. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Minera??o+de+Dados+-+Conceitos,+Aplica??es+e+Experimentos+com+Weka#0

[21] M. Hall Eibe Frank, G. Holmes, B. Pfahringer Peter Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009. [Online]. Available: http://www.cms.waikato.ac.nz/~ml/publications/2009/weka_update.pdf

[22] S. Penin and W. Messias, “Debate: O Enade é eficiente como forma de avaliação da USP? – Jornal do Campus,” 2009. [Online]. Available: http://www.jornaldocampus.usp.br/index.php/2009/09/debate-o-enade-e-eficiente-como-forma-de-avaliacao-da-usp/

[23] K. Uchôa, H. Sobral De Matos, S. Oliveira Chagas, and C. R. Conceição De Menezes, “ENADE: O Desafio de uma avaliação do ensino superior eficaz para as instituições de ensino,” *Encontro Internacional de Formação de Professores e Fórum Permanente de Inovação Educacional*, vol. v.9 n.1, no. 2179-0663, 2016. [Online]. Available: https://eventos.set.edu.br/index.php/enfope/article/viewFile/2349/566