

Extraction of Relevant Terms and Learning Outcomes from Online Courses

<https://doi.org/10.3991/ijet.v11i10.5928>

I. Guitart, J. Conesa, D. Baneres, J. Moré, J. Duran, D. Gañan
Open University of Catalonia, Barcelona, Spain

Abstract—Nowadays, universities (on-site and online) have a large competition in order to attract more students. In this panorama, learning analytics can be a very useful tool since it allows instructors (and university managers) to get a more thorough view of their context, to better understand the environment, and to identify potential improvements. In order to perform analytics efficiently, it is necessary to have as much information as possible about the instructional context. The paper proposes a novel approach to gather information from different aspects within courses. In particular, the approach applies natural language processing (NLP) techniques to analyze the course's materials and discover what concepts are taught, their relevancy in the course and their alignment with the learning outcomes of the course. The contribution of the paper is a semi-automatic system that allows obtaining a better understanding of courses. A validation experiment on a master of the Open University of Catalonia is presented in order to show the quality of the results. The system can be used to analyze the suitability of course's materials and to enrich and contextualize other analytical processes.

Index Terms—Information retrieval, analytics, learning analytics, natural language processing, eLearning, learning outcomes discovery.

I. INTRODUCTION

Higher education sector had begun a changing process during the last decades in order to improve their efficiency and productivity. In 90s decade, for example, some countries reformed the university sector in order to improve their management and finance [1]. In the first decade of twenty-first century, universities felt a higher competition and the pressure of governments. Such pressures also motivated universities to improve their management [2]. Today, the crisis has even worsened the panorama for universities, placing new constraints: less public funds, more competition and exigent students with more expectations and higher economical restrictions. In this new panorama universities have to be managed efficiently, but they also have to excel in teaching in order to survive. To do so, several approaches can be followed. One possible way is to improve university processes by using analytical systems that provide contextual information of the instructional process and detect potential improvements in teaching activities.

Universities have access to a vast amount of educational data and even more when they use virtual learning environments (VLE). In such cases VLE can gather navigational data [3][4][5] and the way students learn can be analyzed [6][7]. However, universities do neither have the culture, nor the capacity to analyze and take advantage of

the available data [8]. In spite of their low analytical culture, several approaches that perform analytics over educational data have appeared in the last years [9][10][11]. Most of them are focused to analyze the data related with students; the background of students [12], their ratings or navigation to propose personalized learning paths [13], personalized activities or automating the assessment feedback [14]. However, as far as we know, there is no system that analyzes the learning resources within courses.

Getting information from students is relevant, but without information of the context where the students work (the course), the knowledge about students cannot be efficiently exploited. Having information about the course (the contents taught, its organization, their activities and the learning outcomes students are supposed to obtain) is essential in order to 1) perform rich analytics related to students; and 2) provide analytics focused to help teachers and programme coordinators to improve their curricula and teaching. This kind of analysis requires to have an extensive knowledge about several facets of the course, such as what concepts and competences are taught, how the course is taught, what the more difficult topics or exercises are to students, whether the course lacks to address efficiently expected competences or which activities are done.

Obtaining the required information to improve the planning and teaching of a course is hard, but maintaining such information updated is even a more demanding task. For that reason, an automatic system is necessary 1) to gather the main information related to courses, and 2) to keep such information updated.

In that way, the information can be extracted, automatically and keep updated easily. However, a manual intervention would be needed in terms of a refining process in order to improve the quality of the output information. The research described in this paper works in that direction: the semi-automatic inference of relevant concepts and learning outcomes related to a course. This output should allow to analyze the information under the point of view of a course, but also in higher abstraction levels, such as the degree where the course belongs.

The presented work proposes a semi-automatic system that infers the pieces of knowledge and competences that are taught in a course and how this information is aligned. Such information is gathered by using natural language processing (NLP) techniques over the learning resources of the course.

In Section 2, the context of the proposed system is presented, that is the relationship between courses, their learning resources and academic programs, such as minors, degrees and masters. Later, Section 3 presents the

proposed system to extract relevant content and Section 4 describes the extension to extract learning outcomes and to analyze the alignment with the extracted terms of the learning resources. Section 5 summarizes the experiment performed to validate the proposal in the master program of Business Intelligence from the Open University of Catalonia. Section 6 discusses how to evolve the system in order to improve its performance. Finally, Section 7 presents the main conclusions of the work and the planned further research.

II. THE CONTEXT: COURSES AND THEIR RELATED LEARNING RESOURCES

Relevant content about a given course can be obtained from different sources: the description of the course, the activities of the course, their communication forums, their lectures, and etcetera. In our context, relevant content is the information that can be used to provide context about the course or allow detecting flaws that denote potential improvements.

In this section we will describe the kind of resources that can be taken into account to infer relevant content in virtual learning environments. We will focus in the more relevant ones, according to our experience, and describe, for each of them, what kind of knowledge they can provide. Figure 1.a) shows the selected resources. Since we plan to use NLP techniques to infer knowledge from resources, non-textual resources, such as video tutorials – unless they provide subtitles–, will be discarded.

The different kinds of data sources we propose to analyze are detailed next:

Academic programs/syllabus: they provide general information about the course, its motivation, its goals, its contents, its calendar and methodology. Their short length makes difficult to extract trustable information from them. However, a lot of information can be extracted when their structure is predefined. Relevant information that can be extracted are the skills, the learning outcomes or traversal competences.

Materials: this kind of resources includes all the textual artifacts that contain the contents students have to learn during the course. They usually include textbooks, lectures, case studies and tutorials. Note that, it is important to have the resources in digital formats (i.e. eBooks instead of printed books) to be able to analyze the required text and avoid OCR (optical character recognition) processes that will add more complexity and noise to the analysis. Learning resources can provide information related to the concepts students must know at the end of the course.

Communication forums: they encompass communication mechanism used by students and teachers to interact. They can be analyzed to gather information about the students' evolution during the course. They can be audio or video forums, but for analytical reasons it is better when they follow textual formats. Similar to printed textbooks; textual formats will avoid speech-to-text processes that will add complexity and noise to the analysis. The potential information that can be extracted are emotional states of students (positive or negative), main questions, doubts, the more problematic concepts (in terms of understanding), complaints and points of improvement.

Learning activities: they include all the activities provided to students, such as assessment activities and exam-

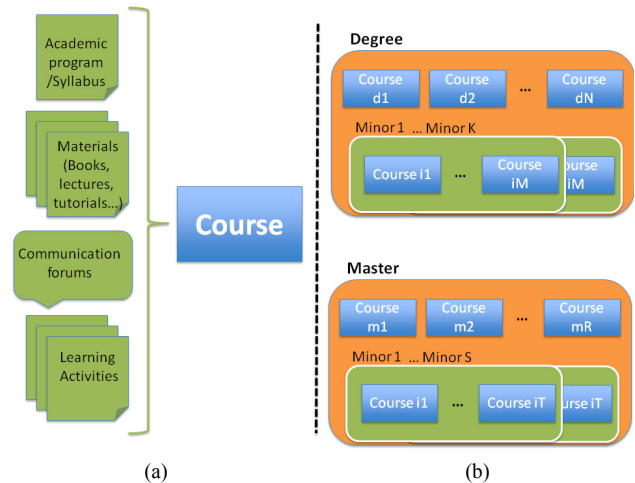


Figure 1. Different kind of resources related to courses and the relation between courses and academic programs.

ples. Learning activities can be used to find out what elements are taken into account in the evaluation of the course. This information can be very useful to detect improvements in the assessment activities by detecting mismatches between the more relevant content (extracted from the material) and the topics used to evaluate the student.

These resources can be used to gather information and to improve the quality of a course from different points of view (planning, content, assessment activities, communication and etcetera). Such information can be used also to provide similar benefits at higher abstraction levels. An academic program, which is the term we use to generalize both masters, degrees and minors, can be seen as an aggregation of courses. Aggregating information may provide information about the main topics dealt in the academic program, check whether the contents are coherent with the program design or detect the courses where students do not fulfill expectations (usually related to some negative emotions). In order to aggregate the information at the different levels of abstraction, it is necessary to know the configuration of the courses on each academic program. Figure 1.b) shows one of the possible configurations for academic programs.

In this paper we plan to deal only with the content of the materials. This is the first approach since concepts can be more objectively grasped and evaluated as a resource of course quality. In terms of this context, the materials may include examples comprised mainly for implemented programs and case studies. With this analysis, we plan to find out information about the main topics for each course (and each academic program). There are multiple approaches related to text analysis using machine-learning techniques [15][16][17]. However, the applied NLP techniques are related to textual analytics that filter types of semantically relevant data from unstructured texts [18][19][20] such as text categorization and concept extraction.

In the literature, there are multiple works related to the analysis of forums based on the analysis of discourse analytics [21]. Authors in [22] explored the dialogue produced on the forum. Unnecessary dialogue for learning such as discussion is discarded and the technique is able to analyze the exploratory dialogue commonly associated

with the construction of knowledge. Sentiment analysis is analyzed in [23] to classify learners based on the emotional state. This information may serve as a satisfaction indicator. A different analysis is performed in [24] to predict student retention. This analysis using other models also based in NLP techniques may predict whether the learner will finish the course based on the comments posted in the forum. Another analysis proposed in [25] focused on classifying the posts depending on objective (question, reflection, reference, statement, feedback, etc.).

In the future, we will deal with more subjective information using previous described techniques. Emotional states of students, flaws in the course or mismatch between the expected concepts to learn and the concepts dealt in the practical activities could be analyzed by correlating concepts in the learning resources with concepts described in the communication channels.

III. A SEMI-AUTOMATIC SYSTEM FOR INFERRING RELEVANT TOPICS

The purpose of the proposed system is to analyze the learning resources within a course in order to extract a set of terms (of different length) that describe the relevant topics. Here, we define a relevant topic as a term that describes a learning content of the course or allows the detection of flaws that denote potential improvements. As aforesaid, in this proposal only textual materials, which include theoretical concepts and examples, will be taken into account.

The inputs of the system (see figure 2) are mainly two: 1) a file that contains the metadata about courses and 2) the learning resources that describe the theory related to the course (materials) and examples about the content (learning activities). The output will be analytical data with information about the relevant terms aggregated by course and academic program. Additionally, some metrics will be added as metadata for each term in order to support the evaluation of its relevancy.

The proposed system is composed of three different processes detailed next:

- **Extractor:** it is the responsible to process the courses and to get a first list of the keywords they contain.
- **Analyzer:** the keywords extracted in the previous step will be filtered taking into account three different criteria: linguistic relevance, course relevancy and subsumption of terms. At the end, a list of relevant keywords will be sent to next process.
- **Aggregator:** the relevant keywords will be aggregated by course and academic program. Finally, keywords will be stored in a database that has been specially designed to facilitate the analysis over the extracted information.

Next subsections describe in more detail each of these processes.

A. Extraction of relevant terms from resources

Before the extraction, the system needs to determine the courses that should be analyzed. The analysis does not have to be meaningful for all courses, due to several reasons, such as the lack of textual information or the obsolescence of materials. The list of courses to be analyzed is determined by reading the input metadata file. This file should contain the list of courses, the academic programs

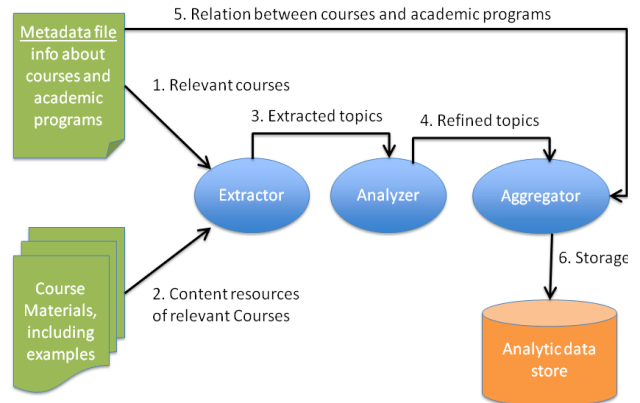


Figure 2. Sketch of the semi-automatic system to infer relevant knowledge about courses from their learning resources.

where they belong and how they are aggregated within a program (by minors, by specialties, etc.).

The metadata file is represented in CSV format and includes the following information (fields):

- The name of the course,
- the academic programs where it belongs; note that a course may belong to more than one academic program. For example “Data Warehousing” may be in the programs “Degree of Computer Science” and “Master of Business Intelligence”,
- the different aggregations where it belongs, understanding an aggregation as a meaningful group of courses related with a purpose, such as a minor or a specialization, and
- the materials related to the course. For each material it is important to get the general information (title, language of the resource, etc.). Here, we assume that it is a material in digital textual format and it is accessible via an institutional repository or an URL.

In this stage of the research, the metadata file has been created manually since the institution does not have any administrative system that produces this information and, moreover, there is no central repository where all resources are stored. However, it can be created semi-automatically when the institution provides information about their academic programs available in some data source.

Then, the relevant terms of the courses are extracted. The extraction performed by means of NLP techniques that segment the text, lemmatize the words and produce candidate terms, which are word lemmas, bigrams and trigrams of lemmas. Not all the candidates are relevant because of their part of speech (POS), and their importance in the course. The importance depends on the appearance frequency of the candidates in a course in comparison to their frequency in other courses. So the extractor uses a POS tagger to analyze each document in order to identify the grammatical category of each term and lemmatize it. Then, the candidates are produced by calculating the single lemmas and lemmatized n-grams of the document. Finally, for each pair <candidate, document>, the frequency of appearance and a score are calculated to determine the importance of the candidate in the document. In our case, the score is computed using the *Lucene* score [26] and it is used in order to discard irrele-

vant elements by defining a threshold. Note that at the end of the process we will have a set of records of the form:

<documentID, candidate, POS, frequency, score>

B. Analyzer: Cleansing relevant terms

The previous process produces a large set of terms. Most of them are useless, not relevant in the domain of the course. For example, the bigram “such as”, which denotes the style of the author, is not relevant as a piece of knowledge in courses. The obtained terms are cleansed according to three different criteria:

- **Linguistic:** its purpose is to merge equivalent keywords and to discard stop words. The system uses Freeling [27] to determine when different keywords are different representations of the same lemma. If so, the different keywords are unified, the frequencies are aggregated and the scores are recomputed. For example, the keywords “analyzing”, “analyzed” and “analyze” are semantically equivalent. Therefore, they can be unified in one term “analyze” and their metrics can be integrated. In the deletion of stop expressions, language-specific rules should be used to identify what expressions can be deleted. For example, a rule will be “When a bigram or a trigram contains a pronoun, then it can be deleted”. Domain synonyms detection has not been dealt in this stage since they require a high manual intervention.
- **Relevancy:** the frequency of terms for each document can be used in order to detect (and discard) terms with few-relevance to the course and detect (and keep) terms that are domain-specific to a given document. According to the number of selected elements, a threshold has been set in order to discard terms whose frequency is under certain number. TFIDF (term frequency – inverse document frequency) metric [28] has been used to find out the domain-specific terms for each document. If a term is considered very specific to a given document, we can say that it is potentially specific to the domain of the document and, in extension, to the course. The TFIDF has been calculated from a custom-built corpus that contains thousands of news from several newspapers during the period 2003-2007 classified in several categories such as politics, sports, or culture. The aim of this corpus is to be able to identify domain-specific terms. If a term appears very frequently within several categories then it will be classified as a general term element and it can be discarded, otherwise it will be classified as a domain specific and should be kept. Results have been satisfactory using the selected corpus. However, in the future it is advisable to use a corpus thematically closer to the academic fields to deal with.
- **Inclusion:** dealing with keywords of different length provides more terms and potentially more quality terms, but it can also produce some noise. This process reduces partially such noise by detecting whether one term is a subset of other term. When we have two terms and one is a subset of the other, we can discard the subsumed one when the frequencies of both terms are very close. For example, suppose the terms “big” with a frequency of 19 and “big data” with a frequency of 18. In such case the system can discard the term “big” without losing precision (only one apparition of “big” will be lost).

```
"subjects": [
  {
    'title' => 'Supply Chain Management',
    'languages' => [
      {
        'lemmas' => [
          {
            'score' => '0.0492235',
            'frequency' => 56,
            'lemma' => "actividades log\x{ed}sticas"
          },
          {
            'score' => '0.037181',
            'frequency' => 29,
            'lemma' => "sistema log\x{ed}stico"
          },
          {
            'score' => '0.0297685',
            'lemma' => 'filiales de multinacionales',
            'frequency' => 6
          }
        ]
      }
    ]
  }
]
```

Figure 3. JSON fragment of the output of the analyzer process.

At the end of the process the output will have the same structure than before, but the size of the set will be drastically reduced. Figure 3 shows an excerpt of the JSON file resulting of this step for a course titled “Supply Chain Management”.

C. Aggregator: Storing relevant terms to support analytics

This process aggregates the relevant information collected from previous processes at different level of granularity.

First, the process will consult, from the metadata file, the set of courses that belong to each academic program. Later, the different thematic aggregations (minors, specialization...) of courses are detected. Then, the terms of each course can be grouped to the different levels of thematic aggregations and the academic programs. Note that the *Lucene score* will be lost in this process since the semantic of this kind of scores does not allow its addition. A possibility would have been to recalculate them by doing a new parsing of all the documents grouped by academic program. Such option has been discarded since we believe that its benefits will be limited but the required computational resources very demanding.

The aggregated information is stored in a database following the conceptual schema summarized in Figure 4. As we can see in the schema, the system will store information about all the keywords detected for each resource, both lemmas (the selected keywords) and keywords that have been discarded due to linguistic reasons. The point is to provide information of equivalences between the different ways relevant elements appear in the documents (synonymy, dialectal variations, etc.). Such information resulted to be useful in some way to detect the discourse speech used on each domain. Note that the system is also prepared to take into account equivalences among lemmas. Equivalences are not yet dealt with in this version of the system. However, it is an evolution proposed in Section VI.

The output of the system is a database that contains relevant information for analytics. The database should be another source that provides data for the analytic systems of the university. In our particular case, such information has been integrated (with other data sources) in a NoSQL graph database, due to the high number of relations between terms and the relevance of these relationships. In particular, Titan [29] has been chosen.

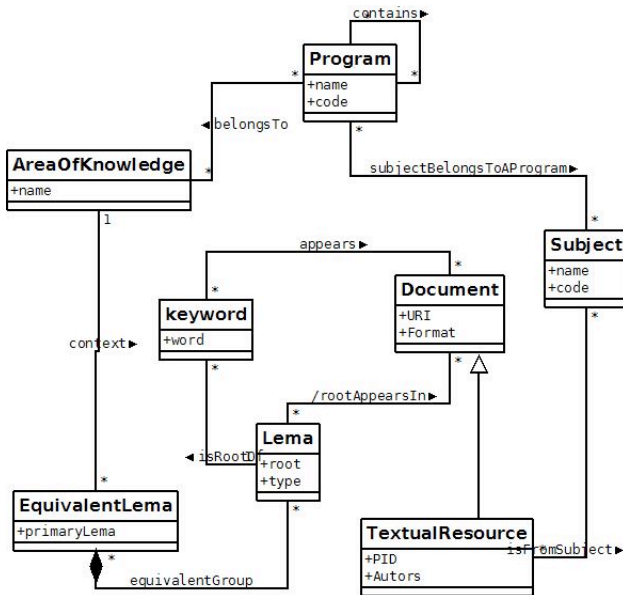


Figure 4. Summary of the conceptual schema of the generated information

Note that, the system will require processing a large amount of data coming from learning resources. Currently, the system has been developed as the pipeline described in Figure 2. From the list of analyzed programs, courses and the respective learning resources are processed sequentially. This could require a large processing time but this is not a critical issue at this stage. Learning resources tend to be a static resource that only may change at the beginning of the instructional process and, therefore, this analysis process is scheduled at this point of the course, once per semester in our case. Note that, only new or updated resources should be processed to optimize time and resources. Data privacy should be another critical issue. However, no sensitive information from student is extracted at this point.

On future evolutions of the system when information of the learner will be analyzed (forum, activities, etc), the schedule of the process and data privacy issues will be revisited.

IV. INFERRING RELEVANT TERMS WITH LEARNING OUTCOMES

In this section, we describe an evolution of the previous system by extracting learning outcomes of the course and inferring their alignment with the course contents. Note that, this type of information could be useful on revising the learning content of a course since a resource could be automatically analyzed to check whether the outcomes are in some extend addressed in the resource.

Our proposal assumes some restrictions: 1) the learning outcomes of a course are clearly defined and stated in the course's syllabus, 2) there are learning outcomes related to each program (or thematic aggregations of courses) and 3) syllabus (or academic program) is a semi-structured document that follows a textual format that can be parsed.

The system provides a semi-automatic process that filters the topics that are potentially related to learning outcomes. This process relates relevant concepts of a course (such as *OLAP* and *data mart* for the course "Data warehouse") with the learning outcomes they contribute (*be*

able to answer analytical questions in the context of a functional unit).

The proposed extension requires adding functionalities to the previous processes:

- **Extractor:** the learning outcomes of each course and program are automatically obtained. The learning outcomes of each course can be obtained from its syllabus by using the technique described in [30]. Therefore, syllabus of courses should also be considered as an input. The learning outcomes from academic programs and other levels of aggregation are more difficult to obtain since they are not systematically defined in structured and accessible documents. We added these learning outcomes in the metadata file. Therefore, a new field has been added in order to define the learning outcomes of each program. The program description will have three fields: the code and name of the program (or level of aggregation) and the list of associated learning outcomes.
- **Analyzer:** It is hard to automatically find out the concepts related to a given learning outcome. However, sometimes the description of learning outcomes may share a common pattern that can be used to detect knowledge and abilities related to learning outcomes. For example, let us assume the learning outcomes (LO) in "Business Analytics" course LO1: "Be able to make queries using SQL" and LO2: "Be able to perform analysis using R-Commander". As we can observe both outcomes start with the common pattern "be able to...", but there are other patterns commonly used to describe learning outcome that also can be used. In such cases, the system could derive the list of relevant terms after the common patterns. Following the previous example, it is easy to automatically infer that *SQL* and *R-Commander* are languages or tools students should practice to achieve LO1 and LO2. If one of the concepts that appear within the materials is *SQL*, then we can estimate that it is related with LO1, and therefore quite important in the context of the course. As a result of the analyzer process, each term incorporates new information about the related learning outcome. Note that, some terms might not be related to any learning outcome:

<documentID, candidate, POS, frequency, score, [Related LO]>

- **Aggregator:** The data scheme described in Figure 4 is extended as it is shown in Figure 5. The class *Learning Outcome* will be created to represent learning outcomes. A *Learning Outcome* can be further specialized as *specific*, when it is only associated to one course, or *generic* (also known as *traversal*), when it is associated with more than one course. The relations between the classes *Learning Outcome*, *Program* and *Course* will be extracted from the metadata file and the syllabus of courses respectively. The relation between learning outcomes and the keywords (*Lemma*) will be automatically extracted from the patterns commented previously.

With this extension, the final result will contain not only information about the relevant contents within materials, but also information about the learning outcomes, the relationship of such outcomes with courses and academic programs and the concepts relevant for each outcome.

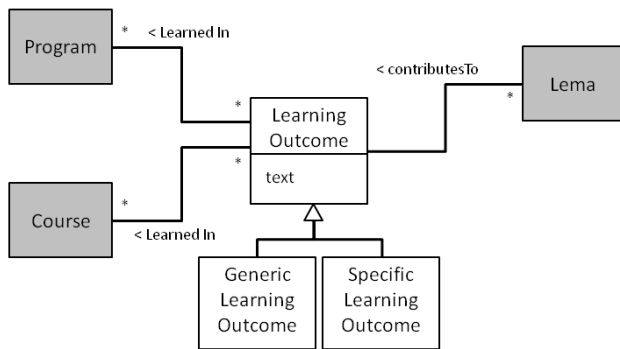


Figure 5. Extension to the original database to take into account learning outcomes.

V. SYSTEM VALIDATION

We have validated the system by conducting two experiments. The aim of the experiments is to analyze the quality of the extracted set of terms by reviewing the automatic system extraction using a manual validation phase. One experiment reviews the set of extracted keywords identified by the system. The other experiment validates the keywords related to the learning outcomes of the courses. The number of validated terms will denote the quality of the inferred keywords. Note that, a cross-validation experiment (i.e. manual extraction of terms) could not be done in all stages of the experiments due to the large set of potential relevant terms. In this case, a perception of relevant terms, which do not appear in the extraction, is performed.

The context where the experiments have been conducted is the Master of Business Intelligence from the Open University of Catalonia (UOC). UOC is the main virtual university in Spain, with over 45,000 students and the instructional process is performed exclusively using a proprietary virtual learning environment (VLE). The VLE has multiple functionalities but the most relevant for this experiment is that it includes virtual classrooms. Each classroom has a communication facility (enabling instructors to guide learning activities and students to ask questions), an area for storing learning resources, a system to submit assessment activities, among others. The learning resources usually contain a textbook (or a set of them) that has the learning contents of the course, case studies, complementary activities and exercises that provide real application of the contents of the course. These learning resources have been used as inputs for our system. For this experiment, learning materials in text document¹, and html format (wikis and blogs) have been taken into account.

The chosen master is one of the most popular from the UOC and has an average of 100 students per year. It takes two years (4 semesters) to finish and has sixteen courses. Students can enroll the full master or any of their semesters, called specialization. The experiment has been conducted in one specialization (composed by 4 courses) due to the large process of manual validation. The necessary time to perform the validation has been around 10 hours.

Table I reports the quantitative analysis of the validation phase of the first experiment. The table summarizes the number of extracted terms after the initial phase and

the cleansing phase. Additionally, the number of relevant terms validated by the manual review is also reported and the percentage of correct terms. Table shows significant results. There is a large detection on the initial phase. There is no validation process on this step due to the large number of terms. The cleansing process reduces dramatically the set of terms and the quality extraction ranges from 65% – 75% in three courses. The discarded terms included acronyms, words about exercises, singular/plural, author's name, website and other irrelevant terms. As we can observe, the course *Case Study in Data Warehouse* has a quality of 10,7%. The reason is that its material was related to a case study of hospitals. Hence, most of their terms were related to the health domain and therefore irrelevant to the data warehousing domain.

In terms of completeness of the selected terms the domain expert who performed the validation process concluded that the system selected most of the relevant terms for the tested courses and specializations. However, many irrelevant keywords were also selected, requiring an extensive refinement work to merge equivalent concepts and discard irrelevant ones. Domain experts pointed the necessity to improve such process by identifying automatically some of the more obvious erroneous terms or by providing a usable environment to validate terms.

Finally, the manual validation has identified few relevant terms that have not selected by the system. In most of the cases, the identified terms have a different semantic meaning in other domains.

The quantitative analysis of the second experiment is shown in Table II. Here, the number of learning outcomes extracted by the system and the number of aligned terms (terms related to a learning outcome) is shown. Similar to the previous experiment, we shown the number of correct terms related to learning outcomes in the validation phase. As we can observe, the learning outcomes are extracted correctly from the syllabus of the courses using the technique proposed in [30]. Only one outcome in the metadata of the specialization could not be extracted because the description of the outcome did not follow any of the predefined patterns to describe learning outcomes. If we refer to the lemma alignment, results show a higher quality in the validation phase because a smaller set of terms is extracted from the learning outcomes description. In the aggregation phase (terms aligned with the specialization), we can observe the quality is still high compared to the quality on individual courses.

Similar to the previous experiment, in terms of completeness, the expert concluded that few terms were missed. Basically, terms that in the previous experiment could not be retrieved. Moreover, all learning outcomes were described in terms of some relevant terms (there was no learning outcome without relevant terms).

These experiments should be analyzed with caution. We are aware that it is a small experiment within a specific domain and, therefore, the results can vary in other domains. However, this experiment gives initial insights about issues that have to be improved as we point out in the next section.

¹ The textbooks of the courses can be retrieved in html format from a private resource repository of the university.

PAPER
EXTRACTION OF RELEVANT TERMS AND LEARNING OUTCOMES FROM ONLINE COURSES

TABLE I.
VALIDATION RESULTS OF RELEVANT TERMS EXTRACTION. THE TABLE SUMMARIZES, FOR EACH ANALYZED SUBJECT, THE NUMBER OF EXTRACTED TERMS, THE NUMBER OF RESULTANT TERMS AFTER THE CLEANSING (SELECTED TERMS), THE NUMBER OF SURVIVING TERMS AFTER THE MANUAL VALIDATION (VALIDATED TERMS) AND THE PERCENTAGE OF CORRECT TERMS (%).

Course	Extracted terms	Lemmas after analyzer		
		Selected terms	Validated terms	%
Business Analytics	60.619	952	644	67,6
Data bases: preliminaries, physical design and benchmarking	83.280	778	561	72,1
Data warehouse	72.358	451	336	74,5
Case Study in Data Warehouse	84.734	1972	211	10,7

TABLE II.
VALIDATION RESULT ON ALIGNMENT BETWEEN RELEVANT TERMS AND LEARNING OUTCOMES IN THE MASTER OF BUSSINESS INTELLIGENCE. THE TABLE SUMMARIZES THE NUMBER OF LEARNING OUTCOMES AUTOMATICALLY DETECTED AFTER THE CLEANSING (SELECTED TERMS), THE NUMBER OF VALIDATED TERMS BY THE DOMAIN EXPERT (VALIDATED TERMS) AND THE PERCENTAGE OF CORRECT TERMS (%). SUCH INFORMATION IS ALSO PROVIDED FOR THE LEMMAS RELATED TO EACH LEARNING OUTCOME.

Course	Learning Outcomes [30]			Lemmas aligned after analyzer		
	Selected terms	Validated terms	%	Selected terms	Validated terms	%
Business Analytics	6	6	100	22	20	90,9
Data bases: preliminaries, physical design and benchmarking	8	8	100	25	21	84,0
Data warehouse	7	7	100	22	18	81,0
Case Study in Data Warehouse	5	5	100	18	17	94,4
Business Intelligence Information Systems	15	14	93,3	62	55	88,7

VI. PROPOSED EVOLUTION

The lessons learnt during the experimentation have lead to the proposal of the evolution of the system in two directions: the automatic detection of equivalent terms and the improvement of the refining process.

After analyzing results on Table I and Table II, we can observe that many non-relevant terms survived the cleansing process. This result requires a further analysis on equivalence detection. We detected the next issues related to the cleansing process.

1. *Some stop words still remain after the cleansing process* in the analyzer step (usually happened in terms with more than one word).
2. *Equivalent terms highly related with the domain of interest.* Additionally to the semantic meaning, we found that terms can be equivalent for different reasons. Table III shows some particular examples found in the validation phase and the reason for each equivalence.
3. *Terms unrelated with the domain of the course but related with the domain of their examples.* For example, in the learning resources of the course "Data warehouse" there was an exercise in which one data warehouse was created to analyze the results of Formula 1 championship. In consequence, keywords related with the domain Formula 1 (*championship, pilot, car, engine, etc.*) were selected as some of the most relevant concepts related to the course. The reason is that terms were very frequent in the materials and very domain-specific.

Related to the equivalence issue, it is important to note that equivalences can appear in combination, and may involve more than two keywords. Improving the analyzer process to deal with keywords with multiple words, basically in the cases of case sensitive and singular/plural can automatically detect some equivalences. Other particular equivalences related to a specific domain should be tackled by other methods. One option isto automatically de-

TABLE III.
EXAMPLE OF EQUIVALENCES FOUND IN THE MANUAL VALIDATION OF RELEVANT KEYWORDS.

Term	Equivalence	Reason
OLAP	Online Analytical Processing	Acronym
Data Warehouse	Data Warehouses	Singular/plural
Hadoop	Apache Hadoop	Generalization/specialization of terms
mapReduce	Mapreduce	Case sensitive
Time diagrams	Chronograms	Technical names
Business Intelligence	BI, Inteligência de Negócio, Inteligência Empresarial	Acronym, different languages
Almacen de Datos	Data Warehouse, Data Warehousing	Different languages, similar words

tect these terms from a domain specific corpus (similar to the TFIDF calculation). This improvement could help on equivalences related to *acronyms, technical names* and *different languages* reasons. Moreover, this domain specific corpus could also help to remove terms unrelated with the domain.

Note that, a manual refining process could finally improve the automatic cleansing process. This process could be performed in both systems (terms extraction and learning outcomes extraction). A manual intervention on initial stages of the deployment of the system could add rules to remove common stop words issues, to add equivalent terms on specific domains, and to specify new detection patterns in learning outcomes extraction.

VII. CONCLUSIONS AND FUTURE STEPS

In the current paper we have presented a solution towards the automatic analysis of relevant topics taught extracted from the learning resources used in the courses. Additionally, the system has been improved with a com-

plementary system to align learning outcomes with the relevant terms. A semi-automatic system implementing this solution has been created. We have also presented preliminary findings and conclusions from two experiments that we conducted in a domain specific master in Open University of Catalonia. The initial insights prove the potential of the system, in the sense that it can automatically select a set of keywords and the learning outcomes that define the main topics and skills taught in a course with a reasonable quality. After conducting the experiments, we have proposed a couple of improvements to increase the effectiveness of the proposed system and its generalization.

In order to improve generalization, authors are planning to work in three directions: 1) by implementing an editor that facilitates the manual refining process of the terms, 2) by improving the linguistic cleansing process in the analyzer step to deal more effectively with terms of multiple words (that would reduce the number of *singular/plural* and *case sensitive* equivalences), and 3) by apply new techniques that use crowd knowledge in order to find out possible equivalences in the cases of *acronyms*, *technical names* and *different languages*.

Obviously, another expected line of research is to use the extracted information for analytical purposes. We plan to use the selected information in an integrated system that helps teachers in the management and teaching process. Such system will give different information to teachers to provide information about courses from different points of view (content, mood of students in the communication forums [9], performance, and etcetera) and will permit teacher to take decisions with more trustable information.

REFERENCES

- [1] D.B. Johnstone, A. Arora, and W. Experton, "The financing and management of higher education: A status report on worldwide reforms", World Bank Washington, 1998.
- [2] C.L. Comm, and D.F. Mathaisel, "Less is more: a framework for a sustainable university", *International Journal of Sustainability in Higher Education*, vol. 4, no. 4, 2003, pp. 314-323. <http://dx.doi.org/10.1108/14676370310497543>
- [3] M. De Laat, V. Lally, L. Lipponen, and R.J Simons, "Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis." *International Journal of Computer-Supported Collaborative Learning*, vol. 2, no. 1, 2007, pp. 87-103. <http://dx.doi.org/10.1007/s11412-007-9006-4>
- [4] F. Xhafa, C. Paniagua, L. Barolli, and S. Caballe. "A parallel grid-based implementation for real-time processing of event log data of collaborative applications". *International Journal on Web and Grid Services*, vol. 6, no. 2, 2010, pp. 124-140. <http://dx.doi.org/10.1504/IJWGS.2010.033788>
- [5] G. Lavigne, G. Gutiérrez, L. McAnally-Salas and J. Organista, "Log Analysis in a Virtual Learning Environment for Engineering Students", *International Journal of Educational Technology in Higher Education*, vol. 12, no. 3, 2015, pp. 113-128 <http://dx.doi.org/10.7238/rusc.v12i3.2162>
- [6] O. R. Zaiane and J. Luo, "Towards evaluating learners' behaviour in a Web-based distance learning environment," *Proceeding IEEE International Conference on Advanced Learning Technologies*, 2001, pp. 357-360. <http://dx.doi.org/10.1109/ICALT.2001.943944>
- [7] C. C. Lin and C. H. Chiu, "Correlation between Course Tracking Variables and Academic Performance in Blended Online Courses," *2013 IEEE 13th International Conference on Advanced Learning Technologies*, 2013, pp. 184-188. <http://dx.doi.org/10.1109/icalt.2013.57>
- [8] R. Ferguson, "Learning analytics: drivers, developments and challenges", *International Journal of Technology Enhanced Learning*, vol 4, no. 5, 2012, pp. 304-317. <http://dx.doi.org/10.1504/IJTEL.2012.051816>
- [9] I. Guitart, J. Conesa, L. Villarejo, A. Lapedriza, D. Masip, A. Pérez and E. Planas. "Opinion Mining on Educational Resources at the Open University of Catalonia" *International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, 2013,, pp. 385-390. <http://dx.doi.org/10.1109/cisis.2013.70>
- [10] Y. Park, and I.H. Jo, "Development of the learning analytics dashboard to support students' learning performance", *Journal of Universal Computer Science*, vol. 21, no. 1, 2015,, pp. 110-133.
- [11] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J.L. Santos, "Learning analytics dashboard applications", *American Behavioral Scientist*, 2013, pp. 1500-1509. <http://dx.doi.org/10.1177/0002764213479363>
- [12] D. Conrad, "Building knowledge through portfolio learning in prior learning assessment and recognition." *Quarterly Review of Distance Education*, vol 9, no 2, 2008, pp 139-150.
- [13] Chen, Sherry Y., and Robert D. Macredie. "Cognitive styles and hypermedia navigation: Development of a learning model." *Journal of the American society for information science and technology*, vol. 53, no. 1, 2002, pp 3-15. <http://dx.doi.org/10.1002/asi.10023>
- [14] N. Capuano, M. Gaeta, A. Micarelli, and E. Sangineto, "Automatic student personalization in preferred learning categories." *3rd International Conference on Universal Access in Human-Computer Interaction*. 2005.
- [15] T. Hofmann. "Probabilistic latent semantic indexing". In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, 1999, pp. 50-57. <http://dx.doi.org/10.1145/312624.312649>
- [16] D. M. Blei, Y. N. Andrew and I. J. Michael, "Latent dirichlet allocation.". *The Journal of machine Learning research*, vol. 3 2003, pp. 993-1022.
- [17] S. T. Dumais, "Latent semantic analysis". *Annual review of information science and technology*, vol. 38 no. 1, 2004, pp. 188-230. <http://dx.doi.org/10.1002/aris.1440380105>
- [18] R. Feldman and J. Sanger. "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data". Cambridge University Press. 2006 <http://dx.doi.org/10.1017/CBO9780511546914>
- [19] S. Grimes, "An Introduction to Text Analytics: Social, Online and Enterprise," *Text and Social Analytics Summit, 2013*; www.slideshare.net/SethGrimes/an-introduction-to-text-analytics-2013-workshop-presentation. 2013 (last visited: May 2016).
- [20] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet., "Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications". Elsevier, 2012.
- [21] M. Dascalu, "Analyzing Discourse and Text Complexity for Learning and Collaborating - A Cognitive Approach Based on Natural Language Processing". *Studies in Computational Intelligence*, Springer 2014, pp. 1-228
- [22] R. Ferguson, Z. Wei, Y. He, and S. Buckingham Shum. "An evaluation of learning analytics to identify exploratory dialogue in online discussions". In *Proceedings of LAK*, 2013, pp 85-93. <http://dx.doi.org/10.1145/2460296.2460313>
- [23] M. Wen, D. Yang, and C. P. Rosé, "Sentiment Analysis in MOOC Discussion Forums: What does it tell us?" In *Proceedings of EDM*, 2014, pp. 130-137.
- [24] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. "Understanding MOOC discussion forums using seeded LDA". In *Proceedings of the 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2014. <http://dx.doi.org/10.3115/v1/W14-1804>
- [25] Aysu Ezen-Can, Kristy Elizabeth Boyer, Shaun Kellogg, and Sherry Booth. "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach". In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15)*, 2015, pp.146-150 <http://dx.doi.org/10.1145/2723576.2723589>
- [26] Lucene score: https://lucene.apache.org/core/3_5_0/scoring.html (last visited: May 2016).
- [27] X. Carreras, I. Chao, L. Padró, and M. Padró, "FreeLing: An Open-Source Suite of Language Analyzers". *Proceedings of the*

4th International Conference on Language Resources and Evaluation (LREC'04), 2004.

- [28] T. Joachims. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization". In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, 1997, pp. 143-151.
- [29] Titan Database: <http://thinkaurelius.github.io/titan/> (last visited: May 2016).
- [30] J. Moré, J. Conesa, D. Baneres, and M. Junyent, "A Semi-Automated System for Recognizing Prior Knowledge". *International Journal of Emerging Technologies in Learning (IJET)*, vol 10, no. 7, 2015, pp.23-30. <http://dx.doi.org/10.3991/ijet.v10i7.4610>

AUTHORS

I. Guitart, is with IT, Multimedia and Telecommunication Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: iguitarth@uoc.edu).

J. Conesa, is with IT, Multimedia and Telecommunication Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: jconesac@uoc.edu).

D. Baneres, is with IT, Multimedia and Telecommunication Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: dbaneres@uoc.edu).

J. Moré, is with Technology and Infrastructure Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: jmore@uoc.edu).

J. Duran, is with Technology and Infrastructure Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: jduran-cal@uoc.edu).

D. Gañan, is with IT, Multimedia and Telecommunication Department at the Open University of Catalonia, Rambla del Poblenou 156, Barcelona 08018, Spain. (e-mail: dganan@uoc.edu).

This work was partly funded by the eLearn Center from the UOC, the SmartLEARN and SOM research groups, and the Spanish Government through the project TIN2013-45303-P "ICT-FLAG" (Enhancing ICT education through Formative assessment, Learning Analytics and Gamification). Submitted 10 June 2016. Published as resubmitted by the authors 13 August 2016.