Relating educational materials via extraction of their topics

Márcio de Carvalho Saraiva
supervised by Claudia Bauzer Medeiros
Institute of Computing
University of Campinas
13083-852
Campinas-SP / Brazil
marcio.saraiva@ic.unicamp.br

ABSTRACT

Digital educational documents are growing in size and variety, and scientists are facing difficulties to find their way through them. One of the initiatives that have emerged to solve this problem involves the use of automatic classification algorithms. However, it is difficult to analyze implicit relationships among topics of materials. This paper presents CIMAL, a framework for enabling flexible access to material stored in arbitrary repositories. CIMAL combines semantic classification, taxonomies and graphs to elicit relationships among topics of educational documents. We validated our work using materials from Coursera (courses offered by Johns Hopkins University and University of Michigan) and a Higher Education Institute, from Brazil.

1. INTRODUCTION

Usually, lecturers use educational material repositories to publish, store and share materials with their peers in academia and students. The access to those documents is usually open. Given such availability, how to find and choose the material(s) more suitable to study a given topic?

Sites such as the International Bank of Educational Objects, the ACM Learning Center and the ACM Techpack, the Coursera platform, MERLOT and SlideShare show that the access to collections of educational materials in different formats and the analysis of their contents are still done in a restricted way. Even simple queries through the interfaces of these repositories can result in a large number of items, making it difficult to understand them and select the relevant ones. Furthermore, none of these repositories offers means to analyze relationships among the stored objects, which would help select material.

This paper presents the design and implementation of CIMAL (Courseware Integration under Multiple relations to Assist Learning), abstractly presented in [10]. CIMAL is a framework to analyze educational documents repositories, allowing visualizations of relationships among materials' topics through the use of graph algorithms. This work was validated with data from Johns Hopkins University and University of Michigan provided at Coursera, which is one of the largest e-learning repositories at the moment, and a Higher Education Institute from São Paulo - Brazil. Our work expands the analysis options in educational material repositories. Moreover, our proposal improves the search among different material formats by standardizing topics they cover.

Proceedings of the VLDB 2018 Ph.D. Workshop, August 27, 2018. Rio de Janeiro, Brazil.

.

2. THEORETICAL FOUNDATION AND RELATED WORK

2.1 Educational Data Mining

According to Romero [9] EDM is concerned with "researching, developing, and applying computerized methods to detect patterns in collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist".

Typically, research towards helping users to select educational material can be roughly classified as (i) development of tools to analyze, access or store materials in repositories, (ii) mechanisms to integrate heterogeneous materials via user monitoring, and (iii) use of learning objects to encapsulate and standardize contents.

2.2 Components and Content from Educational Material

The strategy we adopted to extract and represent topics of educational material is inspired by a concept that we name *components of educational material*. Components are positional structures that highlight information of a given material in order to facilitate its understanding. Header, body, footer and numbering of slides are examples of components of slides; titles, subtitles and the progress bar are examples of components of videos. This information also can be used for analysis; in our work, we use these characteristics in classification, indexing, comparison and retrieval tasks.

Unlike other approaches in the literature that use the entire text of a document equally, we also extract information of components from different types of material to guide classification tasks. Our work presents a novel strategy for documents analysis, which considers the components present in the documents to facilitate the identification of topics in the documents.

2.3 Classification of topics

To classify educational materials, we use a technique called Explicit Semantic Analysis. In natural language processing and information retrieval, According to Egozi et al. [4], Explicit Semantic Analysis (ESA) is semantic representation of text (entire documents or individual words) that uses a document corpus as a knowledge base.

2.4 Recognition of relationships

According to Jiang et al. [5], extraction of relations is the task of detecting and characterizing the semantic relations between entities in texts. They affirm that current state-of-the-art methods use carefully designed features or kernels and standard classification to solve this problem.

Mining of metadata (e.g., number of accesses to data or identification of entities in the documentation of objects) is often used to derive relationships among data, such as the work of Pereira[8]. Relationships of educational materials are viewed as the connections or associations among materials considering educational aspects, such as the association on the contents or connection of lecturers schedules [7].

Another approach to recognize relationships is to use external taxonomies ([6]) or to build an architecture with hierarchies to organize objects in levels, so that these relationships among the objects become the relationships between the levels ([12]).

2.5 Analysis using graph databases

We can characterize a graph database through its data model that differentiates it from traditional relational databases [1]. A data model is a set of conceptual tools to manage and represent data, consisting of three components [3] : 1) data structure types, 2) collection of operators or inferencing rules, and 3) a collection of general integrity rules. Data in a graph database are stored and represented as nodes, edges, and properties.

Each graph database management system has its own specialized graph query language, and there are many graph models. For example, many graph databases based on Resource Description Framework (RDF) use SPARQL (SPARQL Protocol and RDF Query Language), but Neo4J, a graph database widely used in research, uses the Cypher language. Finally, integrity rules in a graph database are based on its graph constraints.

Several researchers have adopted graph representations and graph database systems as a computational means to deal with situations where relationships are first-class citizens (e.g. [2]). They interpret scientific data using concepts of linked data, interactions with other data and topological properties about data organization.

3. THE CIMAL'S ARCHITECTURE

CIMAL's architecture is a novel design to support the analysis of relationships among educational material based on their implicit topics. This architecture combines multiple algorithms for content extraction and classification of topics given a suite of educational material repositories.

Figure 1 presents an overview of our architecture, which comprises three layers. The *Persistence Layer* is composed by six repositories: *Local Courseware, Components and Contents, Representations, Enriched Taxonomy, Classification and Relations.* The *Preprocessing Layer* prepares data from educational material for subsequent search. The latter provides all the services needed to look for materials using graph algorithms. These services can be accessed through the *User Interface* by lecturers and students.

The first step is to set up the repositories (actions represented by arrows with letters 'a' and 'b') before users can perform a search (arrows with letter 'c') . Preprocessing starts when the *Courseware Crawler* imports such materials from external resources (1a) and stores them in a *Local Courseware Repository* (2a). Next, the *Components and Contents Collector* extracts texts and the position of these texts from the materials in the Local Courseware Repository (3a). Extracted data are stored in the *Components and Contents Repository* (4a). Next, the *Intermediate Graph Representation Builder* creates a graph representation for each material from the repositories via the components and contents stored by the previous step (5a). These representations are stored in the *Representations Repository* (6a).

In parallel, the *Combiner*, also proposed in our research, imports an external taxonomy from a *Taxonomy Repository*, and a set of external expert texts from *Domain textual documents Repository* (1a). These data are unified in an Enhanced Taxonomy, in which each concept of the taxonomy has a reference to a text by experts, and stored in the *Enriched Taxonomy Repository* (1b).

Once representations and enriched taxonomy repositories are created, the *Classifier* is ready to define the topics covered in each of the materials (2b,3b,7a). This information is then stored in the *Classification Repository* (8a).

Lastly, the *Relationships Analyzer* looks for prespecified relationships among the items and their topics in the Classification Repository (9a), creating the *Relations Repository* (10a).

All preprocessing steps must be performed every time we add educational material, taxonomy or texts from a domain textual base.

After such preprocessing, lecturers and students can run queries through the *Interface Layer* (1c). It redirects the query to the *Graph Engine* and the *Search Engine* (2c). The latter accesses the *Relations Repository* (3c) to find relevant educational materials that are related to the user query.

4. IMPLEMENTATION

The CIMAL software is the first implementation of the architecture described in Section 3. We have developed the components of Interface and Preprocessing Layer using JAVA code, our texts come from Wikipedia, the taxonomy from ACM Computing Classification System, and methods of Apache Lucene, a high-performance full-featured text search engine library.

Since CIMAL uses graphs to perform relationships analysis, the Persistence Layer stores all data in a database with native support for graphs (Neo4j). With this approach, we are able to use already established technologies and solutions for processing graphs. We chose the Neo4j database system because it is the most popular graph database in big companies (e.g. eBay and Wallmart) and in research, according to the Db-Engines site, an initiative to collect and present information on 341 database management systems.

Our main implementation is divided in four steps: (Step A) Extraction of elements of interest; (Step B) Intermediate Representation Instantiation – based on the schema defined

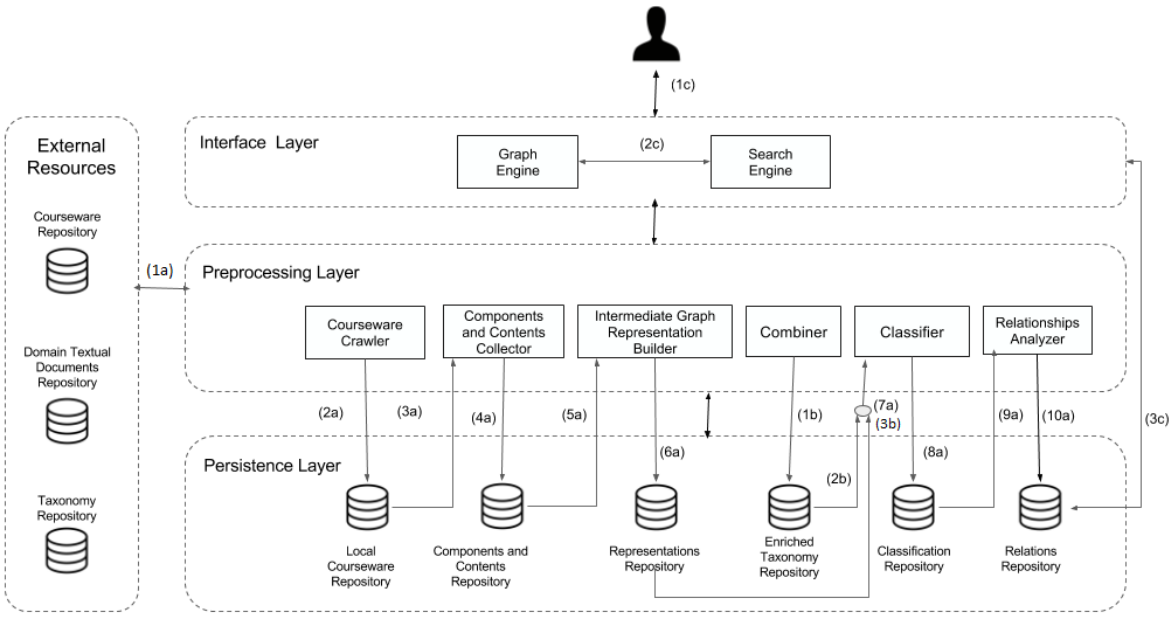

Figure 1: System Architecture for Analysis of Relationships among Educational Material Contents.

in our research; (Step C) Intermediate Representation Analysis; (Step D) Interaction with users.

4.1 Step A - Extraction of elements of interest

At Step A, the Components and Contents Collector extracts components from material based on a Java Framework called DDEx and several APIs for document handling. It scans educational material based on a set of positional rules defined by users and identifies the desired components. Each identified component is encapsulated in a standard representation and forwarded to Step B.

The texts from header and body, and number of slides were extracted automatically using DDEX as components of each slide. In addition, the texts present on the body of slides were also extracted. Through the subtitle file, available for each of the videos, the texts and the time stamps of each of the lecturers' statements were extracted.

4.2 Step B - Intermediate Representation Instantiation

Step B creates the Intermediate Graph Representation and stores this representation in a repository. The use of this representation enables the manipulation of parts of educational material without interfering with the material themselves.

The components and contents of a material are transformed into a graph where the nodes represent the elements of interest that are used in our work. These elements differ according to the kind of material, for example in a video we would like to extract the subtitles and in a slide we extract sections.

4.3 Step C - Intermediate Representation Analysis

Step C has three software modules we implemented: The first module ("Combiner" tool) is concerned with creation and storage of an enriched taxonomy. The second (Classifier tool) recognizes the topics of each Intermediate Representation according to the taxonomy and creates a document about the "Classification of Representations". In our studies, we defined that the words present in the components of the slides or that are among the five most repeated in videos subtitles should be 3 times more important in the classification than the words in the rest of the documents. The third module (Relationship Analyzer tool) concerns the production of information about relations, based on the "Classification of Representations".

The Combiner tool adds one page of Wikipedia to each node of the Taxonomy, thus producing an Enriched Taxonomy. Next, the Classifier tool calculates the similarity of each text of Intermediate Graph Representation (related a each educational material) for each pages of the Enriched Taxonomy.

4.4 Step D - Interaction with users

At last, in Step D users can perform queries to find relevant content. Here we implemented in Java programs and 2graph the Interface layer tools. 2graph is a java-based API to perform Extract, Transform and Load (ETL) resources to graph structures/databases, to handle the information produced by CIMAL and interact with users.

5. RESEARCH CHALLENGES

To achieve the objective of this research the following obstacles have been faced:

1) Although widespread, the idea of sharing teaching materials still faces resistance from lecturers. In order to perform classification tests and also to verify relationships between the topics, it is necessary to find different materials but with similar approaches to explain topics. The solution found was to use materials from the same repository (Coursera) and from the Computing area, in which the idea of electronic sharing is more popular.

2) Most of the lesson videos are produced for a specific audience. Consequently, many lectures only explain concepts in a specific language, and do not produce subtitles for other audiences. Automatic transcription of captions is still a research problem. Therefore, we have selected only videos that had their subtitle produced manually, which drastically reduced the amount of educational videos available in educational repositories that could be used. Thus, we used videos from the Coursera platform, which follow a standard of subtitle production, thereby making the analysis of video content more adequate.

3) The use of graphs for analysis of relationships is very common in many research domains, but this practice is not yet widespread in the educational field. In our work we only use volunteers with knowledge in graphs to analyze the contributions of this research.

6. CASE STUDIES

6.1 Analysis of important topics in a Specialization Course from Coursera

We collected 97 sets of slides and 97 videos from the Specialization course in Data Science, offered by Johns Hopkins University, to be used as a case study. Using our system, we are able to discover the topics covered throughout the specialization course without requiring annotations or other extra tasks for teachers. We point out that CIMAL can thus also be used by lecturers to annotate and classify their materials. More details on this case study can be found at [11].

6.2 Proposed new multidisciplinary activities in an educational institution

A second case study was conducted at an educational institution in the state of São Paulo, Brazil. We show how we find similarities among different courses, thereby highlighting possible intersections, thus revealing potential multicourse activities.

We were able to extract the contents and topics covered in each of the documents that regulated the courses of this institution and relate each of their contents through graphs. Documents with many relations revealed possible interactions between their respective courses.

6.3 Standardizing validation

To finalize our study, we designed a questionnaire to evaluate the classification of topics extracted from 6 materials (randomly chosen for the questionnaire does not get too long) from the "Python for Everybody Specialization", provided by University of Michigan. Thirty volunteers of different levels of education and specialties in sub-areas of Computer Science gave opinions for each of five topics extracted using the CIMAL implementation. After this activity, we can see that CIMAL classifies the materials using pertinent topics, since 64% of the topics indicated by the framework were evaluated "Some related (16,5%)", "Related (15%)" or "Closely related (32,5%)" by the volunteers.

7. CONCLUSIONS AND FUTURE WORK

This paper presented the design and implementation of CIMAL, which allows searching content from educational material, and eliciting relationships among topics. This framework contributes to helping lecturers and students navigate through collections of materials. Our implementation is validated on slides and videos from case studies and showed that the components on slides and videos can be used to classify text and relate topic of these materials.

One particular question is of interest to us: "Can the history of courses taken by students influence the topics that the students are looking for in educational material repositories?"

To answer this question, it is necessary to collect data of user accesses to these materials. For example, data on the last courses that a student held in Coursera could be used to construct a personalized study guide on subjects that would be interesting for this student; the recommendation system could also recommend more Coursera courses.

8. REFERENCES

[1] R. Angles and C. Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39, Feb. 2008.

[2] P. Cavoto, V. Cardoso, R. Vignes Lebbe, and A. Santanchè. FishGraph: A Network-Driven Data Analysis. In *11th IEEE Int. Conf. on eScience*, Germany, 2015.

[3] E. F. Codd. Data models in database management. *SIGPLAN Not.*, 16(1):112–114, June 1980.

[4] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29(2):8:1–8:34, Apr. 2011.

[5] J. Jiang. Information extraction from text. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 11–41. Springer US, 2012.

[6] O. Matos-Junior, N. Ziviani, F. C. Botelho, M. Cristo, A. Lacerda, and A. S. da Silva. Using taxonomies for product recommendation. *JIDM*, 3(2):pages 85–100, 2012.

[7] Y. Ouyang and M. Zhu. eLORM: Learning object relationship mining based repository. *Proc. - IEEE Int. Conf. on E-Commerce Technology and CEC/EEE*, pages 691–698, 2007.

[8] B. Pereira. Entity Linking with Multiple Knowledge Bases: An Ontology Modularization Approach. In *ISWC*, pages 513–520. Springer, 2014.

[9] C. Romero and S. Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[10] M. C. Saraiva and C. B. Medeiros. Use of graphs and taxonomic classifications to analyze content relationships among courseware. In *SBBD 2016, Salvador, Bahia, Brazil*, pages 265–270, 2016.

[11] M. C. Saraiva and C. B. Medeiros. Finding out topics in educational materials using their components. In *47th Annual IEEE FIE, Indianapolis, IN, USA, pp. 1-7*, 2017.

[12] K. Sathiyamurthy, T. V. Geetha, and M. Senthilvelan. An approach towards dynamic assembling of learning objects. In *ICACCI*, pages 1193–1198. ACM, 2012.