Use of graphs and taxonomic classifications to analyze content relationships among courseware

Márcio de Carvalho Saraiva[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – University of Campinas (UNICAMP)
Av. Albert Einstein, 1251 - Cidade Universitaria, Campinas/SP - Brazil, CEP 13083-852

{marcio.saraiva,cmbm}@ic.unicamp.br

Abstract. *The search for educational content in courseware repositories is laborious and time consuming. There is an abundance of such repositories, and research efforts to facilitate search, but access is guided by keywords and/or terms selected by courseware authors, thus lacking flexibility. The goal of this project is to design and develop a suite of tools to assist users to find, analyze and select pieces of educational content that are relevant to their learning goals. Contributions will be both at the algorithm and software design level, and at the user (application) level.*

1. Introduction and Motivation

Lecturers and students need access to various educational materials to understand a new topic or to update their knowledge. The viewing of relationships among topics facilitates this process. However, the increasing amount of educational material available in repositories on the Internet hampers finding the appropriate content.

Sites such as the International Bank of Educational Objects [1], the ACM Learning Center and the ACM Techpack [2], the Coursera platform [3], the ARIADNE Foundation [4], MERLOT [5] and SlideShare [6] show that the access to collections of educational materials in different formats and the analysis of content are still done in a limited way. Simple queries in those repositories can result in a large number of items, making it difficult to understand them and select relevant ones. The answer of traditional search engines results is a set (or disjunction) of potentially interesting documents, which may not be adapted to learning [Changuel et al. 2015]. Furthermore, none of these repositories offers means to analyze relationships among the stored objects, which would help select material.

The main goal of this research is to assist users in finding relevant courseware content in multimedia educational repositories. To achieve this goal, the project is defining and developing algorithms to extract hidden relationships among courseware content. The project concentrates on material presented in lectures, namely slides used during a lecture and videos of the lecture itself. These relationships will assist in the learning process and facilitate the handling of materials that are (indirectly) related to each other.

[1] http://objetoseducacionais2.mec.gov.br/

[2] http://learning.acm.org/, http://techpack.acm.org/cloud/

[3] https://www.coursera.org/

[4] http://www.ariadne-eu.org/

[5] http://www.merlot.org/

[6] http://www.slideshare.net/

To meet the goal, the research will deal with several computing challenges regarding courseware. One big challenge is the integration of different types of courseware that are not necessarily documented. Most lecturers do not publish their lessons using additional information that would help finding them (e.g. metadata). The identification of relationships is another big challenge and involves many issues - e.g, content classification, definition of data structures to store relationships among content, and visual representation os relationships. Still another issue that will be tackled here concerns search mechanisms - once relationships are established, how to search for and navigate along related content?

The three main expected contributions of this proposal are thus: (1) new algorithms to analyze courseware using graphs; (2) new methods to interconnect courseware from various sources, highlighting relationships among content; (3) construction of the CIMAL toolbox, through which lecturers and students from diverse scientific areas will be able to navigate across courseware, using the relationships that progressively emerge, to guide their studies.

2. Methodology and Related Work

To analyze the relationships among courseware we need to define what kind of relations that we would to consider. In this research, our **first hypothesis** is that the relationships among content are more useful to support the choice of appropriate courseware to learning goals than other kind of relations commonly present in educational data mining, e.g. relationships among authors or user profiles.

Since educational material in slides or video format does not have an ”official standard” for content structuring, researchers propose distinct methods to collect this information. Several studies use semantic annotations, tags, XML etc. to add information about the contents of different types of media. However, such an approach require extra effort of teachers having to incorporate an additional step in the production of teaching materials.

Research on automatic extraction of content does not even consider that educational materials may have some intrinsic characteristics, such as the order in which the texts appear and the number of words repetitions.

Often, a single lesson may contain more than one topic. For this reason, the extracted texts of the videos and slides will be organized according to various time intervals, represented via their start and end points. Thus, each lesson will be split and transformed into several text documents. This stage involves several open questions such as: split criteria, structures to store tags and multiple tags. Thus, we formulate our **second hypothesis**: intrinsic characteristics can aid in the extraction of important elements for identifying multiple topics in educational materials.

Since we select important elements of teaching materials, we can classify the content present in each material. Techniques such as unsupervised learning, named entity recognition and explicit semantic analysis (ESA) can be used in the classification task.

In natural language processing and information retrieval, explicit semantic analysis is a vector representation of text. Research like [Shirakawa et al. 2015, Gabrilovich and Markovitch 2007] use ESA algorithms to compute the percentage of similarity (relatedness) between two texts. The measures of similarity helps to distinguish the text from educational materials. This gives rise to our **third hypothesis**: an algorithm that uses ESA and taxonomies can classify educational materials content using extracted elements of these materials.

Relationships among the contents should be stored to be used to facilitate the search for educational materials. As reported by [Khan et al. 2012], a graph database can handle directly a wide range of queries that we are expecting in this work, e.g., queries to analyze relations among content, to compare and check the similarities between lessons and lecturers, or the use of algorithms on graphs, which would otherwise require deep join operations in normalized relational tables. In [Cavoto et al. 2015] authors argue that for analysis of data focusing on a network, complex connections or objects and their interactions, it is better to use graph databases than the relational model. The **fourth hypothesis** of this study is that the use of graph databases can support navigation through the content of educational materials highlighting the relationships among them.

The proposed methodology and the hypotheses will be ultimately evaluated via specification and implementation of a suite of tools - Courseware Integration under Multiple Relations to Assist Learning (CIMAL). In this research we will design and build a software infrastructure that will implement this suite of tools.

Important concepts for this research are: educational data mining to recognize intrinsic relationships and relationship analysis using graph databases. Work on browsing of multisource material usually focus in one kind of data, e.g. scientific papers and web pages, ([Mishra et al. 2010]) and/or focus in semantic annotations to fusion multiple data about the same real-world object in a single database record ([Santanchè et al. 2014, Mota and Medeiros 2013]).

According to [Jiang 2012], extraction of relations is the task of detecting and characterizing the semantic relations between entities in texts. Jiang et al. (2012) affirms that current state-of-the-art methods use carefully designed features or kernels and standard classification to solve this problem.

Educational data mining usually focus on the objects metadata available in log files or extracted by specific tools (e.g., number of accesses to data, identification of entities in the documentation of objects or features like colours and shapes from images or video) to derive relationships among objects, e.g. [Pereira 2014, Little et al. 2012, Ricarte and Junior 2011, Ouyang and Zhu 2007].

In the work of [Little et al. 2012] the authors look at the integration of multimedia search in the SocialLearn platform to assist users to build their own learning pathways by exploring and remixing content. The work emphasizes how content-based multimedia search technologies can be used to help lecturers and students to find new materials and learning pathways by identifying semantic relationships between educational resources in a social learning network. This proposal uses the visual similarity search (VSS) method, which does not work with the courseware present in the dataset of our research, because the latter does not have similar or standardized visual characteristics.

Other common approaches use external taxonomies ([Matos-Junior et al. 2012]) or building an architecture with hierarchies to organize objects in levels, so the relationships among the objects become the relationships between the levels

([Sathiyamurthy et al. 2012]). But these approaches do not take into account characteristics of courseware and still need a step to update the hierarchies created when new topics arise.

3. CIMAL Description

CIMAL encapsulates multiple algorithms that elicit hidden relationships across slides' and videos' contents, so that users can navigate across material produced by different lecturers for distinct subjects. Any other support material (e.g. eBooks) is not being considered for now, but the solution proposed is extensible to different kinds of material.

Figure 1 illustrates CIMAL's software framework. The input (fed continuously) are sets of videos ans slides used in lectures; the output is a graph in which nodes are pieces of courseware and edges indicate relationships that have been elicited among courses' contents. Each box implements one or more algorithms designed in this research.

Box 1 (Extractor) extracts textual elements from the courseware. Next, these textual elements are processed to compose an intermediate graph representation produced by the ”Intermediate Graph Representation Builder” (Box 2). In the present version, the following elements are extracted: name of courseware, author, date, titles, author's descriptions of courseware and texts from slides and subtitles of videos. This intermediate graph representation will be stored in a graph database; there is one graph for each course material.

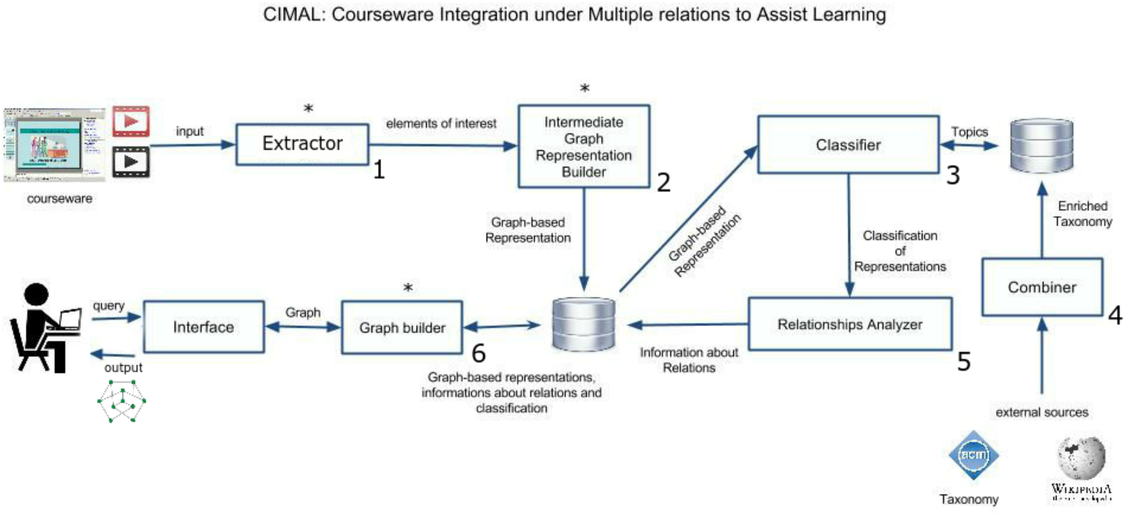
CIMAL: Courseware Integration under Multiple relations to Assist Learning
courseware
input
*
Extractor
1
elements of interest
*
Intermediate Graph Representation Builder
2
Classifier
3
Topics
Enriched Taxonomy
Combiner
4
external sources
Taxonomy
Graph-based Representation
Graph-based Representation
Classification of Representations
Relationships Analyzer
5
Information about Relations
*
Graph builder
6
Graph-based representations, informations about relations and classification
Graph
Interface
query
output

Figure 1. Internal architecture of CIMAL. Boxes marked ”*” are already under development.

These representations are used as input for the ”Classifier” (Box 3). It classifies the elements of the graph in topics/subjects with help of an enriched taxonomy. This is a special structure, created by the ”Combiner” algorithms (Box 4). It combines existing consensual taxonomies with Wikipedia pages that refer to taxonomic elements. Presently, we are processing computing courseware and using the ACM taxonomy

Using an Explicit Semantic Analysis (ESA) algoritm, the ”Classifier” calculates the similarity of text elements of each intermediate graph representation to the set of

Wikipedia pages in the ”enriched taxonomy”. Thus, the ”Classifier” recognizes each topic covered in a course and creates the ”Classification of Representations”, a hash table that associates representations to codes regarding the classification of a topic. Lecturers often teach a given set of subjects in a course. Therefore, the ”Classification” will search for every topic mentioned by a given lecturer, creating ”markers” in videos and slides that, e.g., delimit topics.

The ”Relationship Analyzer” (Box 5) uses the ”Classification of Representations” to produce information about the relationships among the topics of each courseware and stores this information into the database. At present we are using ACM’s relations (”broader”, ”narrower”, ”related”). Finally, The ”Graph builder” (Box 6) creates and stores a graph to represent the knowledge in the database. User queries are processed here, to return subgraphs of interest.

Though our work is general purpose, it is being tested against WebLectures[7], an open courseware repository at UNICAMP, with over 340 two hours lectures on Computer Science, Physics and Mathematics, and hundreds of videos and slides.

4. Preliminary Conclusions

This text presented our research towards designing a suite of tools to integrate courseware by highlighting the presence of relationships among the content covered by these courseware. We use graph databases as the basis of data management and to navigate among the data. The approach to classify the educational material combines ESA algorithms and the ACM Computing Classification System. Our solution was explained via a preliminary sketch of CIMAL’s architecture, using a graph example depicting the result of a query. We have also pointed out some of computing challenges to be faced.

Researchers could use other media, such as audio recordings, books and figures, in future work. Also, a module for viewing maps can be implemented to support analysis of educational materials from different education institutes around the world. An atlas of educational materials could be useful for implementing space-time queries that could enrich research in Education and Computer Science.

5. Acknowledgment

Work partially financed by CAPES, FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), FAPESP-PRONEX (eScience project), INCT in Web Science (CNPq 557.128/2009-9), and individual grants from CAPES and CNPq.

[7] http://lampiao.ic.unicamp.br/weblectures

References

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

Changuel, S., Labroche, N., and Bouchon-Meunier, B. (2015). Resources sequencing using automatic prerequisite–outcome annotation. *ACM Trans. Intell. Syst. Technol.*, 6(1):pages 6:1–6:30.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611, CA, USA. Morgan Kaufmann Publishers Inc.

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

Khan, A., Wu, Y., and Yan, X. (2012). Emerging graph queries in linked data. In Kementsietsidis, A. and Salles, M. A. V., editors, *ICDE*, pages 1218–1221. IEEE Computer Society.

Little, S., Ferguson, R., and Rüger, S. (2012). Finding and reusing learning materials with multimedia similarity search and social networks. *Technology, Pedagogy and Education*, 21(2):pages 255–271.

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

Mishra, S., Gorai, A., Oberoi, T., and Ghosh, H. (2010). Efficient Visualization of Content and Contextual Information of an Online Multimedia Digital Library for Effective Browsing. *WI-IAT2010*, pages 257–260.

Mota, M. S. and Medeiros, C. B. (2013). Introducing shadows: Flexible document representation and annotation on the web. *ICDE Workshops*, pages 13–18.

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

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

Ricarte, I. L. M. and Junior, G. R. F. (2011). A methodology for mining data from computer-supported learning environments. *Informática na educação: teoria & prática*, 14(2).

Santanchè, A., Longo, J. S. C., Jomier, G., Zam, M., and Medeiros, C. B. (2014). Multifocus research and geospatial data - anthropocentric concerns. *JIDM*, 5(2):pages 146–160.

Sathiyamurthy, K., Geetha, T. V., and Senthilvelan, M. (2012). An approach towards dynamic assembling of learning objects. In *ICACCI*, pages 1193–1198, NY, USA. ACM.

Shirakawa, M., Nakayama, K., Hara, T., and Nishio, S. (2015). Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes. *IEEE Trans. Emerging Topics Comput.*, 3(2):pages 205–219.