Correlating educational documents from different sources through graphs and taxonomies

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

[1]Institute of Computing – University of Campinas
Caixa Postal 13083-852 – Campinas – SP – Brazil

{marcio.saraiva,cmbm}@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 [1], the ACM Learning Center and the ACM Techpack [2], the Coursera platform [3], MERLOT [4] and SlideShare [5] 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. On the other hand, Web search engines return a set of potentially interesting documents, which may not be adapted to learning [Changuel et al. 2015].

Indeed, there has been a lack of solutions to identify topics in these materials and how they relate to others. Nevertheless, some efforts have emerged to help solving this problem, such as [Blei 2012, Rossi et al. 2015, Zhuang 2017] that try to discover, extract and collate large collections of thematic structures of documents. However, these and other solutions found in the literature have been conceived to classify documents based on training sets and annotations, strongly coupling the methods to a set of examples.

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

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

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

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

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

Moreover, these solutions require extra tasks in addition to collecting the documents. Last but not least, such solutions have not been applied to sets with different formats of material and do not take advantage of other information from these materials to aid in the classification of topics.

Our proposal is a step towards helping people choose materials of interest from educational repositories. The problem handled in this paper is the elicitation and analysis of relations among different digital educational materials. Unlike related work, which concentrates only on textual sources, our methods process both slides and videos, extracts relevant topic and correlates them. In solving this problem we present the following contributions: (1) to reduce the effort to elicit relationships among various materials; (2) to specify and implement algorithms for correlation of educational material data (videos and slides) from different lecturers; (3) to enable users to conduct search on videos and slides to guide their studies.

This paper presents the design and implementation of CIMAL (Courseware Integration under Multiple relations to Assist Learning), abstractly presented in [Saraiva and Medeiros 2016]. 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.

2. Theoretical Foundation and Related Work

2.1. Educational Data Mining

According to Romero [Romero and Ventura 2013] Educational data mining 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.

An example of (i), Ricarte et al. [Ricarte and Junior 2011] present a methodology to process data collected from educational environments to provide feedback to lecturers about the usage of the content they offer and to students about their behavior inside the environment. However, their work only provides information about access to a particular set of materials, and nothing is said about the content of these resources, the relationships between disciplines, teaching materials and topics mentioned.

An example of (ii) is the work of Little et al. [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.

Finally, we can say that the set of slides and videos used in our research make up groups of learning objects, an example of (iii). According to Sathiyamurthy et. al [Sathiyamurthy et al. 2012] and the Institute of Electrical and Electronics Engineers (IEEE)[Learning Technology Standards Committee of the IEEE 2002] the notion of learning objects (LO) is recurrent in the context of research in EDM.

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. [Egozi et al. 2011], Explicit Semantic Analysis (ESA) is semantic representation of text (entire documents or individual words) that uses a document corpus as a knowledge base. As described by [Gabrilovich and Markovitch 2009], ESA uses an association-based method that interprets a text segment by the strength of its association with concepts that are described in domain documents.

ESA assumes the availability of a vector of basic concepts, [C1, . . . , Cn], and represents each text fragment t by a vector of weights, [w1, . . . , wn], where wi represents the strength of association between t and Ci. Thus, the set of basic concepts can be viewed as a canonical n-dimensional semantic space, and the semantics of each text segment corresponds to a point in this space. This weighted vector is the semantic interpretation vector of t.

Such a canonical representation is very powerful, as it effectively allows us to estimate semantic relatedness of text fragments by their distance in this space.

2.4. Recognition of relationships

According to Jiang et al. [Jiang 2012], 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[Pereira 2014]. 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 [Ouyang and Zhu 2007].

Another approach to recognize relationships is to use external taxonomies ([Matos-Junior et al. 2012]) 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 ([Sathiyamurthy et al. 2012]).

We do not assume that authors of educational material create metadata, but absence of metadata complicates the use of techniques that need this information. Therefore, we will use an approach similar to Explicit Semantic Analysis (ESA) presented in [Gabrilovich and Markovitch 2007]. The latter used a list of concepts to relate texts with Wikipedia articles. As will be seen in our case studies, we relate educational materials using text extracted from these materials, articles from Wikipedia and a taxonomy from an external authoritative source.

2.5. Analysis using graph databases

We can characterize a graph database through its data model that differentiates it from traditional relational databases [Angles and Gutierrez 2008]. A data model is a set of conceptual tools to manage and represent data, consisting of three components [Codd 1980] : 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[6] (SPARQL Protocol and RDF Query Language), but Neo4J [7], 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. [Cavoto et al. 2015]). They interpret scientific data using concepts of linked data, interactions with other data and topological properties about data organization.

As reported by Khan et al. [Khan et al. 2012], a graph database can handle directly a wide range of queries such as those expected in our work and which would otherwise require deep join operations in normalized relational tables. Cavoto et al. [Cavoto et al. 2015] argues 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, considering it is usually necessary to create complex and/or inefficient SQL queries to derive the relationships.

Trying to solve the problem of finding similarities, Gater et al. [Gater et al. 2011] represented process models as graphs to reduce the problem of process matching to a

[6] https://www.w3.org/TR/rdf-sparql-query/

[7] http://neo4j.com/

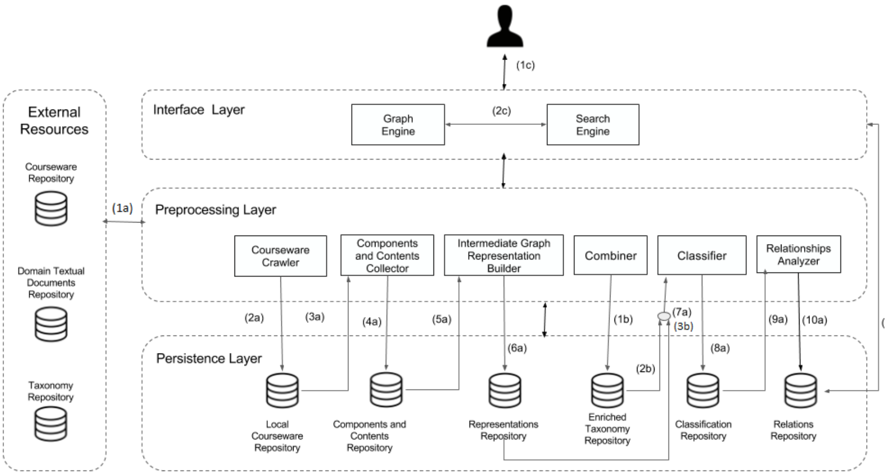


Figure 1. System Architecture for Analysis of Relationships among Educational Material Contents.

graph matching problem. Our research is inspired by the same concepts. We use graph databases to store relationships to take advantage of pattern matching algorithms. Also, using a graph database will help to analyze relations among content, to compare and check the similarities between lessons and lecturers. Algorithms such as Minimax, Betweenness Centrality and Clique may be used and thus facilitate the analysis of the topics extracted from educational materials.

There are many kinds of graph data structures. We chose to model data via *property graphs*, because this allows to create descriptive properties attached to nodes and edges. In our case, the nodes will be the educational materials, and the properties inserted into the edges will describe the relationships among the contents of the nodes. As far as we know, this is the first proposal to use graph databases with information about relationships among contents of educational materials connected to edges.

2.6. Integration of multimedia data

Work that performs the integration of multimedia data from various sources usually focus in one kind of multimedia data, e.g. web pages, ([Mishra et al. 2010, Silva and Santanchè 2009]) and/or exploit metadata to fusion multiple data about the same real-world object in a single database record ([Santanchè et al. 2014, Beneventano et al. 2011]). Examples of metadata used are: author's name, file creation date, labels.

In these proposals, search is performed among different media by searching the metadata describing the stored objects. It is also necessary to implement various different functions to perform similarity search. In our research, we do not consider metadata; rather, we seek to use the contents of educational material and external sources to integrate multimedia data.

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 fromWikipedia, the taxonomy from ACM Computing Classification System[8], and methods of Apache Lucene[9], a high-performance full-featured text search

[8] https://www.acm.org/publications/class-2012

[9] https://lucene.apache.org/

engine library.

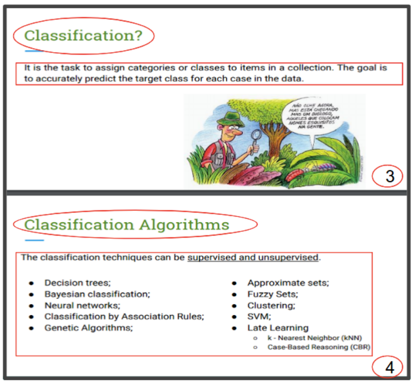
Since CIMAL uses graphs to perform relationships analysis, the Persistence Layer stores all data in a database with native support for graphs (Neo4j[10]). With this approach, we are able to use already established technologies and solutions for processing graphs. We chose the Neo4j database system because it is the most popular graph database in big companies (e.g. eBay and Wallmart) and in research, according to the Db-Engines site[11], 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 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[12] 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 following is an example of Step A applied to a file in slide format and to another in video format. Figures 2 and 3 show the components and texts, respectively highlighted through ellipses and rectangles, that will be used for classification.



Classification?

It is the task to assign categories or classes to items in a collection. The goal is to accurately predict the target class for each case in the data.

NÃO OLHE AGORA, MAS ESTÁ CHEGANDO MAIS UM BIÓLOGO, AQUELES QUE COLOCAM NOMES ESQUISITOS NA GENTE.
3

Classification Algorithms

The classification techniques can be supervised and unsupervised.

- Decision trees;
- Bayesian classification;
- Neural networks;
- Classification by Association Rules;
- Genetic Algorithms;
- Approximate sets;
- Fuzzy Sets;
- Clustering;
- SVM;
- Late Learning
  - k - Nearest Neighbor (kNN)
  - Case-Based Reasoning (CBR)

Figure 2. Components and text extracted from slides.

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. The bold words in the figure represent the terms that were most frequent in the observed time interval.

[10] https://neo4j.com/

[11] http://db-engines.com/en/ranking/graph+dbms

[12] Open Source Project available at http://code.google.com/p/ddex

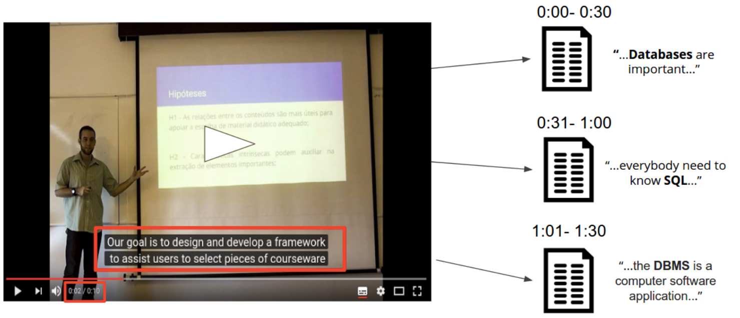

Our goal is to design and develop a framework
to assist users to select pieces of courseware
0:00- 0:30
"...Databases are important..."
0:31- 1:00
"...everybody need to know SQL..."
1:01- 1:30
"...the DBMS is a computer software application..."

Figure 3. Components and text extracted from video subtitles.

4.2. Step B - Intermediate Representation Instantiation

Step B creates the Intermediate Graph Representation adapting the concept of shadows [Mota and Medeiros 2013] and stores this representation in a repository. The use of shadows enables the manipulation of parts of educational material without interfering with the material themselves. In the original work, shadows were implemented using XML files, but in our research we implement shadows in a graph format by the reasons already explained in Section 2.5.

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”. We developed all these tools using Java code and Apache Lucene to search documents based on text similarity.

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 and 2graph[13] 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.

[13] Available at http://www.lis.ic.unicamp.br/ matheus/projects/2graph

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

Coursera is a web platform that provides universal access to educational material and courses online from universities and organizations around the world. However like other producers of educational material, Coursera often does not indicate all the topics covered in a given content. This hampers distinguishing among courses.

We collected 97 sets of slides and 97 videos from the Specialization course in Data Science, offered by Johns Hopkins University [14], to be used as a case study. For this study, our enriched taxonomy was based on ACM Computing Classification System.

Using our system, we are able to discover the topics covered throughout the specialization course without requiring annotations or other extra tasks for teachers. These topics can then be briefly presented as requirements or even in a short course that would be offered to all students before enrolling in the specialization course.

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 [Saraiva and Medeiros 2017].

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 multi-course activities.

[14] https://www.coursera.org/specializations/jhu-data-science

This educational institution seeks to promote interdisciplinary activities to prepare students for the increasingly complex labor market, which requires diversity of knowledge. However, there are many courses that make it difficult to see their relationships.

Using our architecture, 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.

The results of this case study were presented to the faculty of the Institute, who through a questionnaire evaluated if the information obtained could be used to elaborate activities involving courses. In total, 20 lecturers from different courses answered the questionnaire, and 75% answered that it was possible to use the information obtained to propose new interdisciplinary activities between 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 (2 undergraduate student, 3 undergraduate degree, 3 specialists, 6 Master in progress, 4 Master’s degree, 8 PhD in progress, 4 PhD completed) gave opinions for each of five topics extracted using the CIMAL implementation. Since the course was about ”Python programming language” and in the ACM taxonomy these terms are not present, we added manually in our database the Wikipedia page about this topic.

We analyze 900 answers, in each of them a volunteer indicated if he had knowledge about the topic that is being asked. Only answers from volunteers who reported having knowledge about the topic were considered (747 answers). 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. Acknowledgment

We thank Prof. R. D. Peng from Johns Hopkins University and Prof. C. R. Severance, who made slides, videos and subtitles available to be used in a case study. Also we would like to thank the faculty members of Federal Institute of Education, Science and Technology of São Paulo - Campus Hortolândia, who participated in the evaluation and analysis of the results. Also we thank all the members of the Laboratory of Information System - Institute of Computing/Unicamp in which this research was conducted, who assisted several times with suggestions.

Our work partially financed by FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), INCT in Web Science (CNPq 557.128/2009-9), and individual grants from CAPES and CNPq.

References

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

Beneventano, D., Gennaro, C., Bergamaschi, S., and Rabitti, F. (2011). A mediator-based approach for integrating heterogeneous multimedia sources. *Multimedia Tools and Applications*, 62(2):427–450.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

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.

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

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

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.

Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34(1):443–498.

Gater, A., Grigori, D., and Bouzeghoub, M. (2011). A graph-based approach for semantic process model discovery. *Graph Data Management*, pages 438–462.

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 *ICDE*, pages 1218–1221. IEEE.

Learning Technology Standards Committee of the IEEE (2002). Draft standard for learning technology - learning object metadata. Technical report, IEEE Standards Department, New York.

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).

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

Rossi, R. G., Rezende, S. O., and Lopes, A. A. (2015). Term network approach for transductive classification. volume 9042, pages 497–515. Springer International Publishing.

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.

Saraiva, M. C. and Medeiros, C. B. (2016). Use of graphs and taxonomic classifications to analyze content relationships among courseware. In *Brazilian Symposium on Databases, SBBD 2016, Salvador, Bahia, Brazil*, pages 265–270.

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

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

Silva, L. M. D. and Santanchè, A. (2009). ARARA: Autoria de Objetos Digitais Complexos Baseada em Documentos. *Simpósio Brasileiro de Informática na Educação*, (2009):10.

Zhuang, Y. (2017). Bag-of-discriminative-words (bodw) representation via topic modeling. *IEEE Transactions on Knowledge and Data Engineering*, 29(5):977–990.