Universidade Estadual de Campinas
Instituto de Computação

Márcio de Carvalho Saraiva

Relationships among educational materials through the extraction of implicit topics

Relacionamentos entre materiais didáticos através da extração de tópicos implícitos

CAMPINAS
2019

Márcio de Carvalho Saraiva

Relationships among educational materials through the extraction of implicit topics

Relacionamentos entre materiais didáticos através da extração de tópicos implícitos

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientadora: Profa. Dra. Claudia Maria Bauzer Medeiros

Este exemplar corresponde à versão final da Tese defendida por Márcio de Carvalho Saraiva e orientada pela Profa. Dra. Claudia Maria Bauzer Medeiros.

CAMPINAS
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Saraiva, Márcio de Carvalho, 1988-
Sa71r Relationships among educational materials through the extraction of implicit topics / Márcio de Carvalho Saraiva. – Campinas, SP : [s.n.], 2019.

Orientador: Claudia Maria Bauzer Medeiros.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Material didático. 2. Educação - Inovações tecnológicas. 3. Mineração de dados (Computação). 4. Classificação. I. Medeiros, Claudia Maria Bauzer, 1954-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Relacionamentos entre materiais didáticos através da extração de tópicos implícitos
Palavras-chave em inglês:
Teaching materials
Education - Technological innovations
Data mining
Classification
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Claudia Maria Bauzer Medeiros [Orientador]
Kelly Rosa Braghetto
Claudia Lage Rebello da Motta
Julio Cesar dos Reis
Ariadne Maria Brito Rizzoni Carvalho
Data de defesa: 14-08-2019
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-2203-1265
- Currículo Lattes do autor: http://lattes.cnpq.br/3382719006017432

Universidade Estadual de Campinas
Instituto de Computação

INSTITUTO DE
COMPUTAÇÃO

Márcio de Carvalho Saraiva

Relationships among educational materials through the extraction of implicit topics

Relacionamentos entre materiais didáticos através da extração de tópicos implícitos

Banca Examinadora:

- Profa. Dra. Claudia Maria Bauzer Medeiros
Instituto de Computação (IC) - Universidade Estadual de Campinas (UNICAMP)
- Profa. Dra. Kelly Rosa Braghetto
Instituto de Matemática e Estatística (IME) - Universidade de São Paulo (USP)
- Profa. Dra. Claudia Lage Rebello da Motta
Núcleo de Computação Eletrônica (NCE)- Universidade Federal do Rio de Janeiro (UFRJ)
- Prof. Dr. Julio Cesar dos Reis
Instituto de Computação (IC) - Universidade Estadual de Campinas (UNICAMP)
- Profa. Dra. Ariadne Maria Brito Rizzoni Carvalho
Instituto de Computação (IC) - Universidade Estadual de Campinas (UNICAMP)

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 14 de agosto de 2019

Dedicatória

Dedico esta tese a minha família, que sempre me apoiou, motivou, confiou em mim e deu tudo o que podiam para me ajudar na realização deste sonho.

Bom não é ser IMPORTANTE
O importante é ser BOM.
(Roque Schneider)

Agradecimentos

Agradeço a Deus e a família de Nazaré por estar comigo em todas as horas boas e mais difíceis que passei durante o Doutorado. Sem a fé e a força que recebo Deles nada seria possível. À Bruna, por sempre estar ao meu lado. Pelo incentivo e pela compreensão. Por me ajudar a enfrentar todos os desafios, e por fazer de mim uma pessoa melhor a cada dia. À minha família, pelo apoio e pela confiança. Aos meus queridos irmãos, Fernanda e Eugênio, e pais, Márcio e Solange, meus sogros, Eveline e Paulo, e Tia Loura, pelo suporte e constante preocupação e atenção. À professora Claudia, por sempre me ajudar muito durante todo esse processo. Uma verdadeira professora, orientadora e até mãe. Pela sabedoria empregada em nosso trabalho, pelas críticas, paciência e carinho. Tenho muito orgulho de ser aluno dela. Aos amigos que estiveram sempre comigo ao longo dessa jornada. Em especial aos amigos do Laboratório de Sistemas de Informação - LIS, da Ctrl+Play, da Neoway, os professores e alunos do Instituto Federal de São Paulo - Hortolândia, professores do Instituto de Computação da Unicamp que encontrei durante ações, eventos e comissões e aos demais professores da Unicamp que encontrei em diversas experiências na universidade, (como na representação dos estudantes), pelas opiniões e discussões. Aos membros da banca, pelas sugestões e contribuições no trabalho. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Por isso, agradeço às agências de fomento CAPES (06/2014 - 02/2016/1406869), CNPq (141226/2016-1) e também aos projetos, FAPESP/CEPID in Computational Engineering and Sciences (2013/08293-7), INCT in Web Science pelo apoio direto ou indireto na realização deste trabalho.

Resumo

Os documentos educacionais digitais estão crescendo em tamanho e variedade, atendendo a um público cada vez mais heterogêneo. Como conseqüência, os alunos estão enfrentando dificuldades para encontrar documentos educativos para estudo. Vários cientistas criaram repositórios online para armazenar e facilitar o acesso a esses documentos. Infelizmente, na maioria desses repositórios, os documentos são armazenados de maneira aleatória. Isso dificulta a distinção entre o conteúdo desses materiais, bem como o acesso. Como conseqüência, os alunos interessados em acessar material educativo passam a utilizar mecanismos tradicionais de busca na Web. Na maioria dos casos, os resultados das buscas desses mecanismos de pesquisa são apresentados como um conjunto de documentos potencialmente interessantes, mas que necessitam ser analizados um a um pelos alunos. Algumas das iniciativas que surgiram para resolver esse problema envolvem o uso de algoritmos de classificação automática, por exemplo, Modelagem de Tópicos e Rotulagem de Tópicos. No entanto, permanece a dificuldade de analisar relações implícitas entre tópicos de materiais educativos de diferentes professores. Além disso, essas soluções não foram aplicadas a conjuntos de documentos educativos com diferentes formatos, como conjutos de slides e vídeos. As soluções existentes também não aproveitam as informações adicionais dos formatos dos documentos, como metadados, para extrair tópicos. Este trabalho apresenta o CIMAL, um framework para análise flexível de repositórios de documentos educacionais; O CIMAL combina classificação semântica, taxonomias e grafos para extrair tópicos e seus múltiplos relacionamentos. Validamos nossa proposta por meio de um protótipo que utiliza documentos reais da Coursera (Universidade Johns Hopkins e Universidade de Michigan) e de um Instituto de Ensino Superior, do Brasil. Até onde sabemos, é a primeira vez que os recursos de conjuntos de slides e vídeos educativos foram utilizados com técnicas para análise de texto, classificação de tópicos e descoberta de relacionamentos semanticos. A elicitação de tópicos abordados em vários documentos educacionais e de seus possíveis relacionamentos podem apoiar professores e alunos na realização de atividades acadêmicas mais dinâmicas que as convencionais - por exemplo, atividadades entre diferentes disciplinas e cursos. Isso também pode facilitar a pesquisa de documentos mais relevantes em repositórios educacionais para que uma turma de alunos possam aprender novos conceitos, aprimorando o desenvolvimento de novos cursos. Do ponto de vista computacional, esta pesquisa contribuiu para o aprimoramento de técnicas de manipulação de documentos não estruturados e de diferentes formatos.

Abstract

Digital educational documents are growing in size and variety, catering to an increasingly heterogeneous public. As a consequence, students are facing difficulties to find their way through such material. Several scientists have created online repositories to store and facilitate access to these documents. Unfortunately, in most such repositories documents are stored in a haphazard way. This hampers distinguishing among contents of these materials, as well as their retrieval. As a consequence, students interested in accessing relevant material revert to (traditional) Web search engines, or to browsing through one specific repository. In most cases, the results of invoking those search engines are presented as a set (or disjunction) of potentially interesting documents, which may not be adapted to the learning purpose. One of the initiatives that have emerged to solve this problem involves the use of automatic classification algorithms, e.g. Topic Modeling and Topic Labeling. However, them remains the difficulty to analyze implicit relationships among topics of materials and lecturers from different places, even within a single repository. Moreover, these solutions have not been applied to sets of documents with different formats, and do not take advantage of additional information - e.g., metadata to extract topics. This work presents CIMAL, a framework for flexible analysis of educational material repositories; CIMAL combines semantic classification, taxonomies and graph structures to extract topics and their multiple relationships. We validated our proposal through a prototype that uses real materials from Coursera (Johns Hopkins University and University of Michigan) and Higher Education Institute, from Brazil. As far as we known, this is the first time that both slide and video features guide text analysis, topic classification techniques and relationship discovery among documents. The elicitation of topics covered in various educational documents and of their potential relationships can support teachers and students in undertaking academic activities that are more dynamic than conventional ones – e.g., in which new relationships are found between different subjects from different sources. This can also make it easier to search the most appropriate items in educational repositories to learn new concepts, enhancing the development of new courses. From the computational point of view, this research contributes to the improvement of techniques for handling unstructured documents and documents of different formats.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

The avalanche of tools and shared environments for educational purposes has become a problem. The search for educational content on the internet or in courseware repositories is laborious and time consuming. Thus, choosing relevant documents has become a costly task.

In particular, lecturers and students need access to various educational materials to understand a new topic or to update their knowledge. Although there is an abundance of such repositories, and research efforts to facilitate search, the access is guided by keywords and/or terms selected by courseware authors, thus lacking flexibility.

Our goal is to assist lecturers and students in finding relevant courseware content in multimedia educational repositories and navigate through collections of courseware. The idea is to help handling course content, emphasizing relationships among topics therein and among distinct courses. Related work concerning integration and visualization of such content shows that there are still many challenges in the field, e.g., limitations to handle relationships, and query flexibility. Moreover, there is still need for detection of differences between content produced by distinct lecturers or even by a single lecturer, but at different points in time.

Moreover, to achieve our goal, we define and develop algorithms to elicit hidden relationships among courseware content. The research concentrates on material presented in lectures, namely slides used during a lecture and videos of the lecture itself. Any other support material (e.g. eBooks) is not being considered for now, but the solution proposed is extensible to different kinds of material. These relationships will assist in the learning process and facilitate the handling of materials that are (indirectly) related to each other. Our motivation came from interdisciplinary and multiviews research, where lecturers and students need many different views from the same set of materials and topics.

The research presented in this thesis concerns challenges in Educational Data Mining (EDM), in particular to overcome the problem of extract and correlate topic of heterogeneous educational material. The research deals 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 could 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 of 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?

To analyze the relationships among courseware we need to define the kinds of relations we should consider. In this research, our **first hypothesis** is that the relationships among content are more useful to support the choice of appropriate courseware for learning goals than other kinds of relations commonly present in EDM, 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 approach requires extra effort from lecturers who have 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 times each word repeated.

Often, a single lesson may contain more than one topic. For this reason, we extract texts of the videos and slides and organize these texts according to various time intervals, represented via their start and end points. Thus, each lesson is splitted 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 [67, 24] uses ESA algorithms to compute the percentage of similarity (relatedness) between two texts. The measures of similarity help 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 from these materials.

Relationships among the contents should be stored to be used to facilitate the search for educational materials. As reported in [33], a graph database can directly handle a wide range of queries that we expected 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 [13], 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 are ultimately evaluated via specification and implementation of a suite of tools - Courseware Integration under Multiple Relations to Assist Learning (CIMAL). 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.

1.2 Problem Statement and Research Problems

In 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] researchers created repositories for courseware. However, even 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 provided by traditional search engines is usually a set (or disjunction) of potentially interesting documents, which may not be adapted to learning [14]. Furthermore, none of these repositories offers means to analyze relationships among the stored objects, which would help select material. As remarked by [49] such relationships represent an important information.

An underlying overall issue is the heterogeneity inherent to such content, which is produced by many people, adopting a variety of teaching techniques and emphasizing topics in different ways. Moreover, the same lecturer will often change the way (s)he teaches a given subject, e.g., depending on the students' level, the lecturer's understanding of the subject, or the need to relate it to other subjects. In fact, the problem is not so much finding content, but selecting pieces that are relevant to one's learning goals.

This thesis concentrates on course material presented in lectures, namely slides used during a lecture and videos of the lecture itself. Any other support material (e.g. books) will not be considered, but the solution proposed in our work can be extended to different kinds of material. Our research question is the following: How to find and choose the most appropriate material(s) to study a particular subject?

Given this context, we can now refine our goal. The objective of this work is to enable the integration of educational materials (slides and videos of lectures) using relationships among content to assist in the learning process and facilitate the handling of materials that are related to each other.

Towards this goal, we have specified and implemented a suite of tools - Courseware Integration under Multiple Relations to Assist Learning (CIMAL). This is a software infrastructure that elicits and highlights the presence of relationships among the educational materials and the content covered by them.

This research concentrates on the following challenges regarding courseware.

[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/

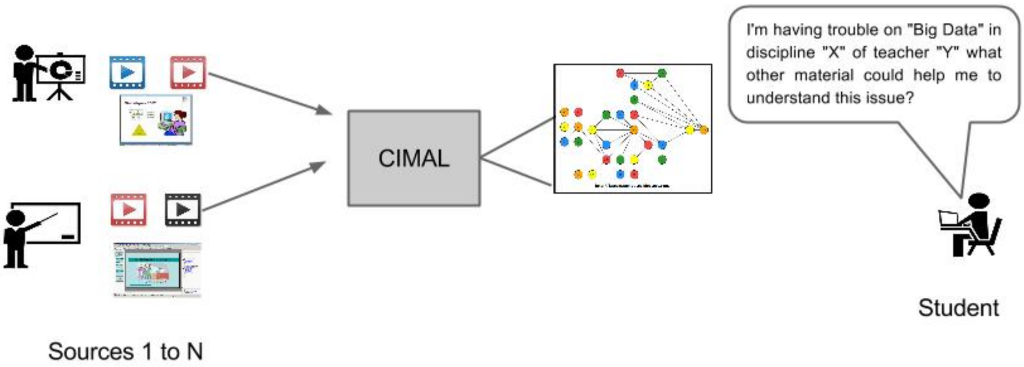

I'm having trouble on "Big Data" in discipline "X" of teacher "Y" what other material could help me to understand this issue?
CIMAL
Student
Sources 1 to N

Figure 1.1: Overview of the use of CIMAL

- **Challenge 1: Integration of different types of 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).

- Challenge 2: Identification of relationships. 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 of relationships. Another issue that was tackled here concerns search mechanisms - once relationships are established, how to search for and navigate along related content?

1.3 Research Overview

Figure 1.1 illustrates an overview of the use of CIMAL in this doctoral research. At first, teachers will make available lessons recorded on videos, and slides used ("Sources 1 to N"). These materials will be stored in courseware repositories, serving as input to CIMAL. At the opposite end, a user can perform queries to find materials of interest. As a result, the system will output the files containing educational materials which are related to the desired topics. We propose that the output be visualized as a graph in which nodes are courseware or topics and edges indicate relationships among course content. Relationships include "broader", "narrower" or "related" when they occur between topics, and "mentions" when the relationship is between topics and courseware. Relationships are labelled according to their semantics - e.g. course A "mentions" topic X, or topic Y "is broader than" topic Z.

Briefly, the three main contributions of this proposal are thus: (1) to reduce the effort to elicit relationships among various educational materials; (2) to specify and implement algorithms for integration of courseware metadata (videos and slides) from various sources; and (3) to enable students from diverse scientific areas to conduct search on videos and slides to guide their studies.

Figure 4.1, which also appears in Chapter 3 and 4 (publications), sums up our work and main contributions. It represents the grey box of Figure 1.1. As explained in Chapter 3, information flow is as follows: 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.

We point out that some of these modules were implemented by direct invocation of existing code, while others required construction of appropriate data structures and algorithms. For instance, the "Classifier" module essentially consists of invoking the ESA function within Lucene[7]. On the other hand, the "Combiner" module constructs a complex structure for subsequent topic elicitation. Details appear in each chapter.

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

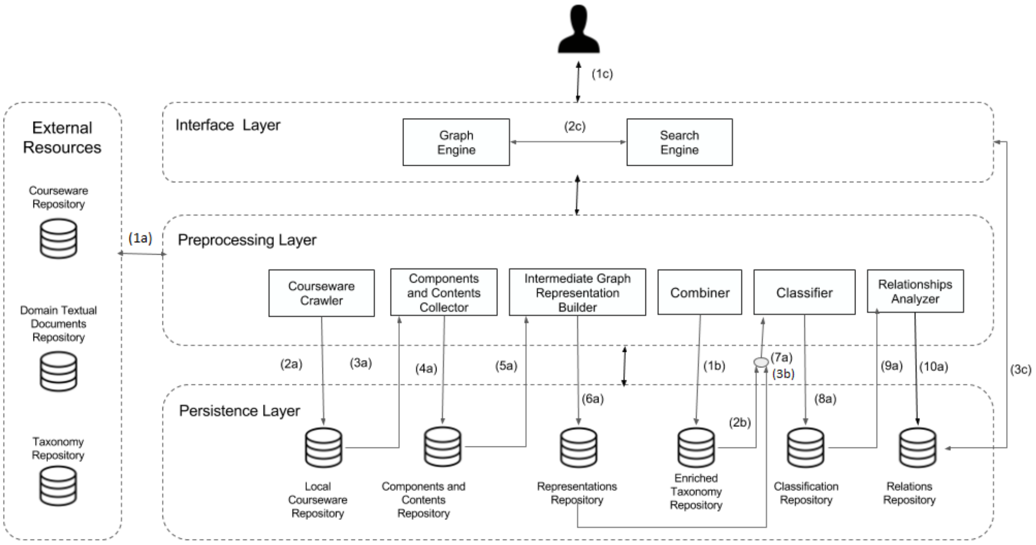


Figure 1.2: System Architecture for Analysis of Relationships among Educational Material Contents.

1.4 Thesis Organization

This chapter presented the organization of this PhD thesis. The remainder of this text is organized as a collection of papers, as follows.

Chapter 2 corresponds to the paper "Finding out Topics in Educational Materials Using their Components", published in the The 47th Annual Frontiers in Education (FIE) Conference, 2017. This chapter discusses the area of Educational Data Mining, the important algorithms to classify documents, and presents a case study of real educational materials extracted from a course. We also show the most important topics in a course obtained using our approach.

Chapter 3 corresponds to the paper "Correlating Educational Documents from Different Sources Through Graphs and Taxonomies", published in the 33rd Brazilian Symposium on Databases (SBBD) 2018. This chapter presents the formalization and implementation of our solution, an architecture to view relationships between topics, which combines graphs, taxonomies and a Natural Language Process (NLP) algorithm, called Explicit Semantic Analysis (ESA). Moreover, in this chapter we describe more case studies that analyze real life examples of courses and educational material and how they can benefit from our proposal.

Chapter 4 corresponds to the paper "Analysis of Semantic Relationships among Educational Material via Graphs", submitted to the journal Multimedia Tools and Applications. This paper, currently under review, describes our efforts to propose a property graph data model with a set of operators and graph algorithms to manipulate data about relationships of courseware topics. Also, in Chapter 4, we define viewpoints in educational material graphs to support the need for multiple perspectives in interdisciplinary research. Finally, this chapter presents how classic algorithms of graphs can be used in the analysis of relationships of our study.

Chapter 5 contains our final conclusions and some directions for future work.

Appendices A, B, C, contain respectively questionnaires used in Case Studies mentioned in Chapter 3 and 4, and approval of the University's Ethics Committee. Appendices are in Portuguese.

Besides the papers in Chapters 2, 3 and 4, others were also published in the course of this thesis, directly related to this research. Following are all publications produced throughout this PhD research.

- Saraiva, Márcio de Carvalho; Medeiros, Claudia Bauzer. Use of graphs and taxonomic classifications to analyze content relationships among courseware (conference). Proceedings of the 31st Brazilian Symposium on Databases, Salvador, Bahia, Brazil, 2016.

- Saraiva, Márcio de Carvalho; Medeiros, Claudia Bauzer. Finding out Topics in Educational Materials Using their Components (conference). Proceedings of The 47th Annual Frontiers in Education (FIE) Conference, Indianapolis, Indiana, United States of America, 2017.

- Saraiva, Márcio de Carvalho; Medeiros, Claudia Bauzer. Correlating Educational Documents from Different Sources Through Graphs and Taxonomies (conference). Proceedings of the SBC 33rd Brazilian Symposium on Databases (SBBD), Rio de Janeiro, Rio de Janeiro, Brazil, 2018.

- Saraiva, Márcio de Carvalho; Medeiros, Claudia Bauzer. Relating educational materials via extraction of their topics (conference). Proceedings of the VLDB 2018 Ph.D. Workshop, Rio de Janeiro, Rio de Janeiro, Brazil, 2018.

- Saraiva, Márcio de Carvalho; Medeiros, Claudia Bauzer. Analysis of Semantic Relationships among Educational Material via Graphs submitted to journal Multimedia Tools and Aplications, 2019

Chapter 2

Finding out Topics in Educational Materials Using their Components

2.1 Introduction

Increasingly, lecturers create digital educational material to support their students. This material is commonly shared via the Internet, transforming the Web into a huge platform to share courseware [71, 42, 47]. Several scientists have created repositories to organize and facilitate access to these materials. However the producers of these materials often do not indicate all the topics covered in a given content or do not follow a standard protocol to indicate this information. This hampers distinguishing among such contents, as well as their retrieval.

In most cases, when someone (e.g. a student) is looking for educational contents or a specific subject, the results of traditional search engines are presented as a set (or disjunction) of potentially interesting documents, which may not be adapted to learning purposes [14]. A technique called Topic Modeling was developed to reduce this problem; it is used to discover, extract and collate large collections of thematic structures of documents [11, 74]. Topic modeling is a set of algorithms capable of discovering and extracting topics from the structure of a documents corpus, aiming at the identification of this collection and facilitating the subsequent analysis of them for e-learning [64].

Topic Modeling is generally used in conjunction with labeling techniques. Topic Labeling is a technique that allows users to view topics semantically more consistent, decreasing dependence on specialized knowledge (on the domain or collection) necessary for the interpretation of such topics.

However, these and other solutions commonly 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. Moreover, they require extra tasks in addition to collecting the documents (such as [58]). In addition, these solutions have not been applied to sets with different formats of material and do not use other information from these materials to aid in the classification of topics.

This paper presents our strategy to solve this gap. Our method is mainly based on exploiting what we name "components of educational material". It will be illustrated via an example of its application. Though our work is general purpose, it is being tested against slides and videos from Coursera[1], a web platform that provides access to online educational material and courses from several organizations and universities.

The elicitation of topics covered in various educational materials could support teachers and students to undertake study activities in a dynamic way. As will be seen, our proposal lets each person customize the connections (relationships) across courseware from different sources, thus creating a personalized set of materials according to a person's interests and goals. It can also make it easier to search the most appropriate items in educational repositories to learn some new concept, enhancing classes.

From the computational point of view, this research contributes to the improvement of techniques for handling unstructured data, with different formats. To the best of our knowledge, our is the first proposal in which slide and video features will guide text analysis and topic classification techniques.

2.2 Concepts and Related Work

2.2.1 Educational data mining

Our work involves a recent research area called Educational data mining (EDM). 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. According to Romero and Ventura [57] research in EDM is formed by intersection of the areas: Data mining and machine learning, Computer-based education and Learning analytics.

2.2.2 Components of educational materials

The strategy presented here to represent courseware content is inspired by a concept created in this research: *components of educational material.* Components are positional structures that highlight information of some material in order to facilitate the understanding of these materials. 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 tasks, indexing, comparison and retrieval.

Unlike other approaches in the literature that use the entire text of a document equally, we will also extract information of components from different types of material to guide classification tasks.

2.2.3 Document Analysis

Currently, document analysis is concentrated in three main strategies to deal with large volumes of complex and heterogeneous documents[46]: 1) to convert the original document

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

into a specific format; 2) to use interoperable standards (e.g., XML) to extract information from the documents; 3) to use only user-provided metadata, requiring user assistance.

The first strategy generally presents an ad hoc conversion methodology for a document type and needs to be changed if different types of documents are used concomitantly. In strategy two, the main difficulty is to handle format diversity, since interoperable formats and predefined schemes are a prerequisite - e.g. other studies that use the same documents, will be limited to use XML. On the other hand, approach three deals very well with file format diversity, but adds an extra step in document production, making the production process even more tedious and laborious.

Our work presents a novel strategy to documents analysis, which considers the components present in the documents to facilitate the identification of topics in the documents.

2.2.4 Topic Modeling

Topic Modeling is based on a set of unsupervised techniques that assume that documents are composed of a mixture of topics. Thus, documents are represented as the set of topics. Topics can be regarded as a probability distribution over the vocabulary; they are learned in an unsupervised manner, that distribution indicates semantic coherence between words[1, 73] .

Probabilistic topic models allow work such as [8, 7], to represent and handle documents at a higher level (topics rather than words). On the other hand, those work is limited to the document vocabulary, hence documents of authors with very different vocabularies may not be composed by the same mixture of topics.

Both unsupervised and probabilistic approaches are highly dependent on the vocabulary a lecture used in a given document. This makes it difficult to analyze educational materials from different sources and hinders the choice of the best material for study. Our strategy uses an external authoritative source to standardize the topics extracted from courseware, and thus decreases the problem of manipulating various documents with different vocabularies.

2.2.5 Topic Labeling

Topic labeling is an activity whose goal is to choose few phrases that sufficiently explain the meaning of the topic. According to Allahyari and Kochut[2], this task can be labor intensive particularly when dealing with hundreds of topics, attracting considerable attention to this area.

Most research in topic models uses the distribution over words to represent the knowledge of a topic (e.g. [59, 11]). However, some authors (e.g [19]) claim that these approaches demand some familiarity with the domain and the document collection. Users without this knowledge will not be able to elicit concepts from a set of words, to identify the main subject or to compare different themes.

Studies, such as [38, 72, 19], use phrases or words extraction methods to group and classify documents. These approaches focus only in corpus analysis and do not consider any other information from the document. We believe that some extra information from

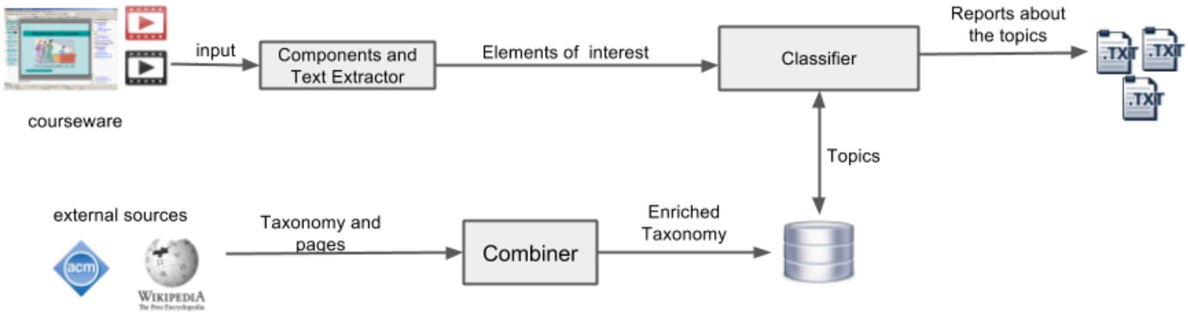


Figure 2.1: Overview of our proposal.

the document (e.g our components) can support classification tasks.

As will be seen, we also perform topic labeling. However, to define the topics present in educational material, we use the components of these materials and external bases to standardize the labels used in the classification.

Lau et al.[34] also used external databases to generate labels for topic models, but the authors limit themselves to a single label to classify the topics for the whole document, even when a document might address a variety of issues (something very recurrent in educational materials).

The components used in our work to classify the topics in a courseware will also guide a method to divide the material when the topic changes in the text.

2.3 Extracting topics from educational material

Figure 2.1 shows the stages of our methodology to find out topics in educational materials topics using their components. The first module "Components and Text Extraction" extract components under the assumption that they are good descriptions of that document. Components extracted include author, date, header, body, footer, numbering of slides and title, subtitle and progress bar of videos. At the end of this module, the text from each of these components are extracted to compose a set of elements of interest.

Next, elements of interest will be used as input for the next stage, the "Classifier". Here, we access a database which stores an enriched taxonomy created in the "Combiner" module. The "Combiner" accesses external sources of knowledge (such as Wikipedia[2] and the ACM Computing Classification System) to create a new structure, called "enriched taxonomy", which helps topic classification, e.g. topic "Graph-based database models" from ACM is linked to the Wikipedia page with title "Graph database". This structure uses the ACM taxonomy as a basis, and links each taxonomy term to one Wikipedia page.

Using an Explicit Semantic Analysis (ESA) algorithm, defined by [24], we calculated the similarity of elements of interest in each courseware to the set of pages of Wikipedia present in the "enriched taxonomy" created by the "Combiner". Thus, we can recognize each topic covered in a educational material and create a hash table that associates

[2] https://www.wikipedia.org /

material to topic labels regarding the classification of a topic, e.g. topic "Database".

Lecturers often teach a given set of subjects in a course. For this reason, we search for every topic mentioned by a given lecturer in an educational material, "slicing" the material by time (for video) or placing "markers" in slides (such as changes in the titles of the slides).

In the last module, the "Classifier" generates text reports that indicate all the topics found in each of the documents. These reports can be used to conduct analyzes of educational materials more easily and quickly than examining the content of each material separately.

2.4 Proposal shortcomings

Although our proposal is general purpose, it may have some shortcomings whether the educational material has no identifiable components. In this case, we would be limited to just using the corpus of the educational material, like other strategies reported in the literature. Also, we can not "slice" or "mark" the material to indicate divisions of topics.

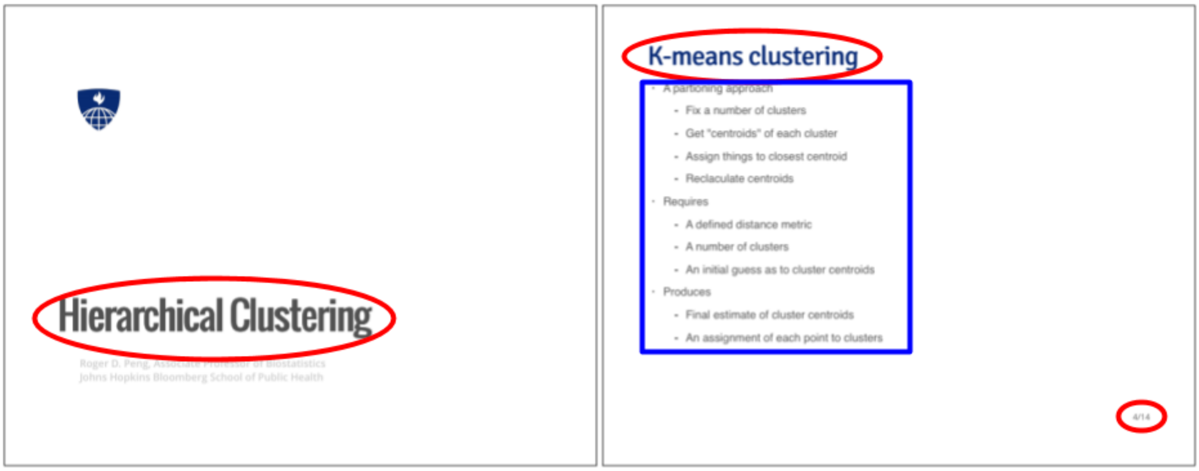
Another obstacle happens whether educational material has many images. Some lecturers usually fill their slides and videos with pictures to illustrate concepts, draw audience attention or even graphics that require interpretation. Since our solution does not present any technique for processing image contents, these contents are not taken into account for the recognition of topics covered in the materials. To solve this problem, a module for extracting content of visual components should be added in the Component and Text Extraction stage.

Our methodology will not work whether topics covered in slides and videos are implicit, specific, or new. For example, imagine that a lecturer wants to tell a personal or common sense story to illustrate the importance of a particular concept for that subject. Probably, the topics found by our methodology will be inadequate to classify the material used by that lecturer. Another example is that the lecturer prepares a material on a recent subject-matter such as "Big Data" or "Internet of Things". Though our proposal uses ACM's latest taxonomy (until this moment, it was last changed in 2012), such terms are not present.

Our approach treat these two examples as any other cases. The components and texts of this material are related to the most similar Wikipedia page (even though the similarity is close to zero), and the latter is related to the ACM taxonomy node, which will be defined as the topic addressed in that material. This can lead to topics that are not appropriate for the materials in these two examples. To solve this problem, in future work new external sources must be proposed, such as constantly updated technical books and ontologies to improve the algorithms of topics classification used in our research.

2.5 Case Study

To show the applicability of our approach, we performed each step described above in educational materials from Coursera, a web platform that provides universal access to education material and courses online from universities and organizations around the world. We collected 97 documents in the slide format and 97 videos from the Specialization course in Data Science, offered by Johns Hopkins University[3], to be used as a case study. The following is an example of our approach applied to a file in slide format and to another in video format. In Figure 2.2 and 2.3 we can observe the components and texts, respectively highlighted through ellipses and rectangles that will be used for classification.



Hierarchical Clustering

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

K-means clustering

- A partioning approach
  - Fix a number of clusters
  - Get "centroids" of each cluster
  - Assign things to closest centroid
  - Reclaculate centroids
- Requires
  - A defined distance metric
  - A number of clusters
  - An initial guess as to cluster centroids
- Produces
  - Final estimate of cluster centroids
  - An assignment of each point to clusters

4/14

Figure 2.2: Components and text extracted from slides.

The texts from header and number of slides were extracted as components of each slide. In addition, the texts present on the body of slides were also extracted.

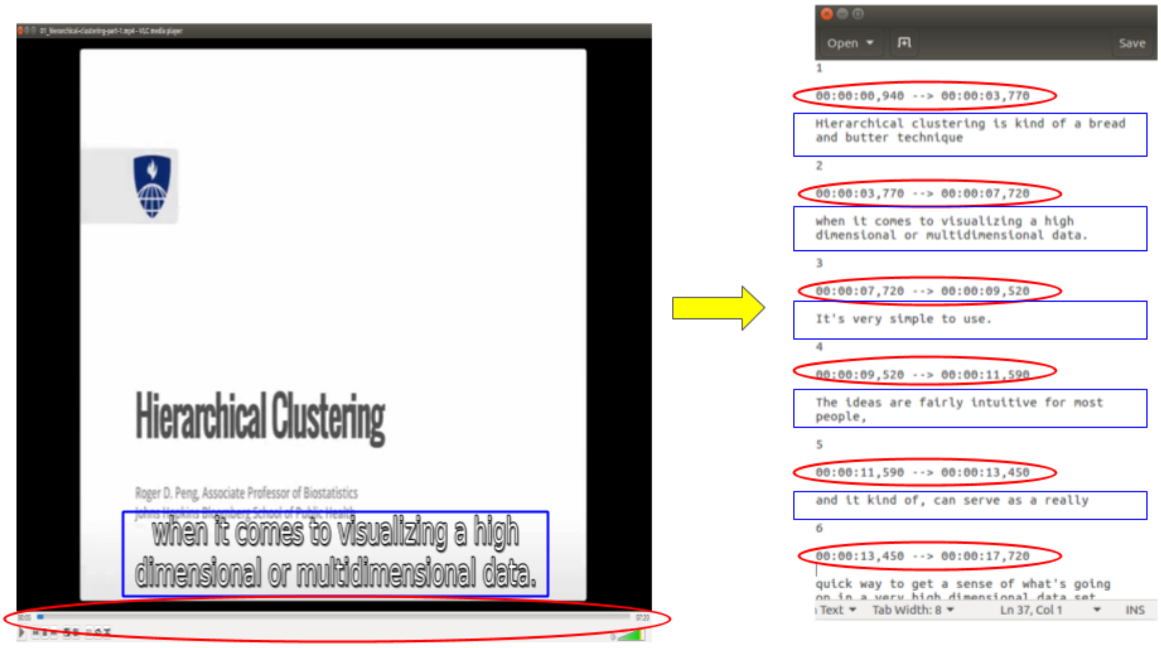
Through the subtitle file, available for each of the videos, the texts and the time stamps of each of the lecturers' statements were extracted.

This information was then used to classify each of the educational materials in the case study collection. Finally, the similarities of the texts of the slides and videos were compared with a set of 900 Wikipedia pages, selected according to the ACM taxonomy. In this case of study, the words that appear in the headers are twice the weight of the words that appeared in the body to slides classification. We have created this difference between word weights as we believe that headings are more important in determining the topics present in a lecture. The Wikipedia names of the most similar pages for each educational material were used to represent the topics of each material.

The time of each speech assisted in the detection of topic changes throughout each video, allowing to verify that a given video could address more than one subject. To accomplish this analysis, the subtitle text was divided into five-minute "time windows", so a set of subtitle extracted in a 30 minute period became six new subtitles, allowing each to be classified separately. In an analogous way were made tests with slides separating its contents every 5 slides.

This case study is our first step in the direction towards applying our proposal to produce a system for aggregating resources across multiple online teaching resources. Despite this, the case study is comprehensive enough to show the effectiveness of the

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



Hierarchical Clustering

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

when it comes to visualizing a high
dimensional or multidimensional data.

```
1
00:00:00,940 --> 00:00:03,770
Hierarchical clustering is kind of a bread
and butter technique
2
00:00:03,770 --> 00:00:07,720
when it comes to visualizing a high
dimensional or multidimensional data.
3
00:00:07,720 --> 00:00:09,520
It's very simple to use.
4
00:00:09,520 --> 00:00:11,590
The ideas are fairly intuitive for most
people,
5
00:00:11,590 --> 00:00:13,450
and it kind of, can serve as a really
6
00:00:13,450 --> 00:00:17,720
quick way to get a sense of what's going
on in a very high dimensional data set
```

Figure 2.3: Components and text extracted from video subtitles.

proposal for finding out topics on educational slides and videos.

2.6 Results discussion

At the end of the case study, we are able to discover the topics covered throughout the specialization course without the need for notes or other extra tasks for teachers. Because of these findings, some analysis/research questions were possible, for example: "What are the n topics most frequently covered during the course?"

To answer this question, we extracted the five main topics from each of the slides in a lecture and videos. Then we computed the frequency of each topic in slides and videos, and proceeded to compute the frequency of each topic in the course (set of slides and of videos). We chose the "top five" in the classification of topics, as this concept is commonly used in the literature. However, using our approach, we can verify the presence of more than hundreds of topics present in the ACM Computing Classification System. If we used a very large number, the relevance of the topics found would decrease. In fact, if n is not provided, the system returns hundreds of topics with insignificant frequency. Figure 2.4 shows the answer to our question.

Thus, we can conclude that Regression Analysis is the most recurrent topic during the Specialization course in Data Science at Johns Hopkins University, present in 24.74% of classes. It is followed by Robust Regression (20.62%), SQL (16.49%), Relational Database Model (15.46%) and Linked Lists (12.37%). These topics could be briefly presented as requirements or even in a short course that would be offered to all students before enrolling in the specialization course.

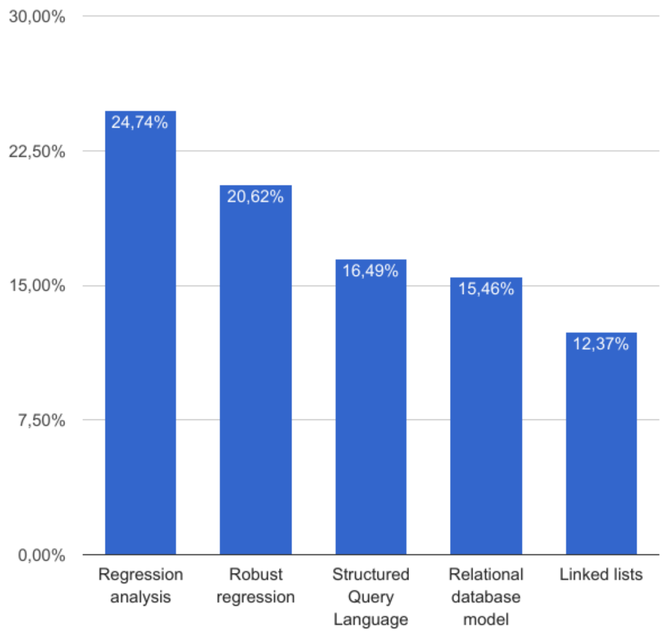


Figure 2.4: Top 5 topics covered in the Specialization course in Data Science at Coursera.

2.7 Ongoing work

We are currently investigating ways to analyze the possible relationships among the topics elicited from the educational materials. Relationships among the contents should be stored to be used to facilitate the search for educational materials.

According to Khan et al.[33], using a graph database we can handle directly a wide range of queries that we are expecting that students and lecturers would make on a platform for access to educational material, 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.

At this stage, our hypothesis is that the use of graph databases can support navigation through the content of educational materials highlighting the relationships among them.

2.8 Conclusions and future work

This text presented our research towards designing a new approach to discover topics in educational materials using their components. We extract component from these materials (slides and videos) and input them to a classification algorithm. Our classification algorithm combines ESA algorithms, ACM Computing Classification System and Wikipedia. Our solution was tested against slides and videos from Coursera and showed that the placement of text on slides and videos can be used to text classification and topic extraction of these materials.

In future, we would like to extend our work to incorporate relationships among subjects themselves (and not just relate material and topics). Moreover, it would be interesting to add more information about the documents structure to facilitate understanding the results. In addition, our research could be applied in educational institutions to propose new multidisciplinary activities that can be proposed combining our methodology with recommendation algorithms. Through information collected about experiences of students and teachers in courses, we could recommend activities.

Researchers could apply our methodology to other domains or other media, such as audio recordings, books and figures. 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.

A Recommender System could be developed to improve the choice of slides and videos; However, it would be necessary to collect data from user access 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.

Chapter 3

Correlating educational documents from different sources through graphs and taxonomies

3.1 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 [14].

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 [11, 58, 74] 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. 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

[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/

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 [61]. CIMAL is a framework to analyze educational document 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.

3.2 Theoretical Foundation and Related Work

3.2.1 Educational Data Mining

According to Romero and Ventura, [57] 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. [55] present a methodology to process data collected from educational environments to provide feedback to lecturers about the usage of the content they offer 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. [39]. 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 [64] and the Institute of Electrical and Electronics Engineers (IEEE)[35] the notion of learning objects (LO) is recurrent in the context of research in EDM.

3.2.2 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.

3.2.3 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. [21], Explicit Semantic Analysis (ESA) is a semantic representation of text (entire documents or individual words) that uses a document corpus as a knowledge base. As described by [25], 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.

3.2.4 Recognition of relationships

According to Jiang et al. [31], 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[51]. 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 [49].

Another approach to recognize relationships is to use external taxonomies ([44]) 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 ([64]).

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 [24]. 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.

3.2.5 Analysis using graph databases

We can characterize a graph database through its data model that differentiates it from traditional relational databases [3]. A data model is a set of conceptual tools to manage and represent data, consisting of three components [16] : 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. [13]). 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. [33], 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. [13] 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, 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. [27] represented process models as graphs to reduce the problem of process matching to a 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

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

[7] http://neo4j.com/

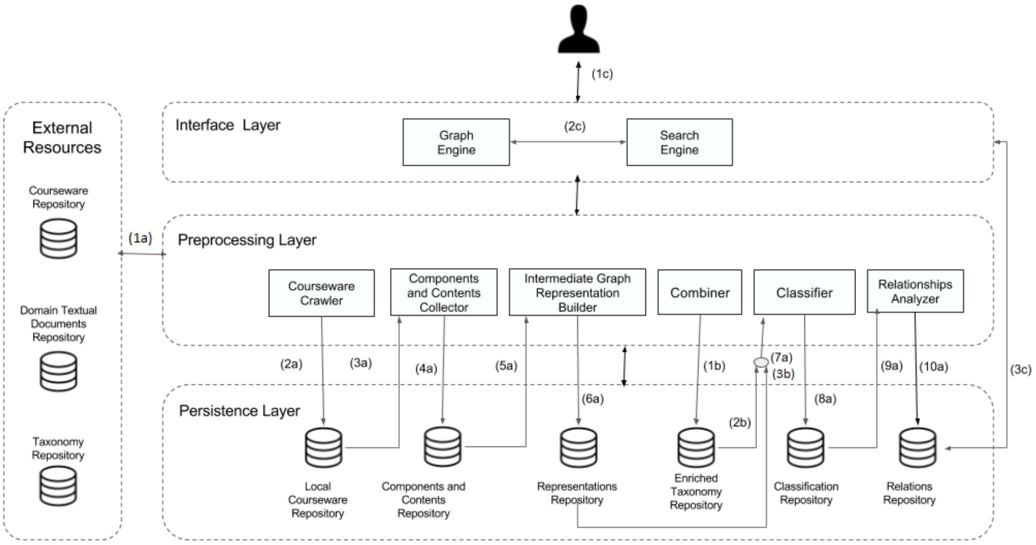


Figure 3.1: System Architecture for Analysis of Relationships among Educational Material Contents.

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.

3.2.6 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, ([45, 68]) and/or exploit metadata to fusion multiple data about the same real-world object in a single database record ([60, 9]). 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.

3.3 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 3.1 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.

3.4 Implementation

The CIMAL software is the first implementation of the architecture described in Section 3.3. 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[8], and methods of Apache Lucene[9], a high-performance full-featured text search engine library.

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

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

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.

3.4.1 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 3.2 and 3.3 show the components and texts, respectively highlighted through ellipses and rectangles, that will be used for classification.

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.

3.4.2 Step B - Intermediate Representation Instantiation

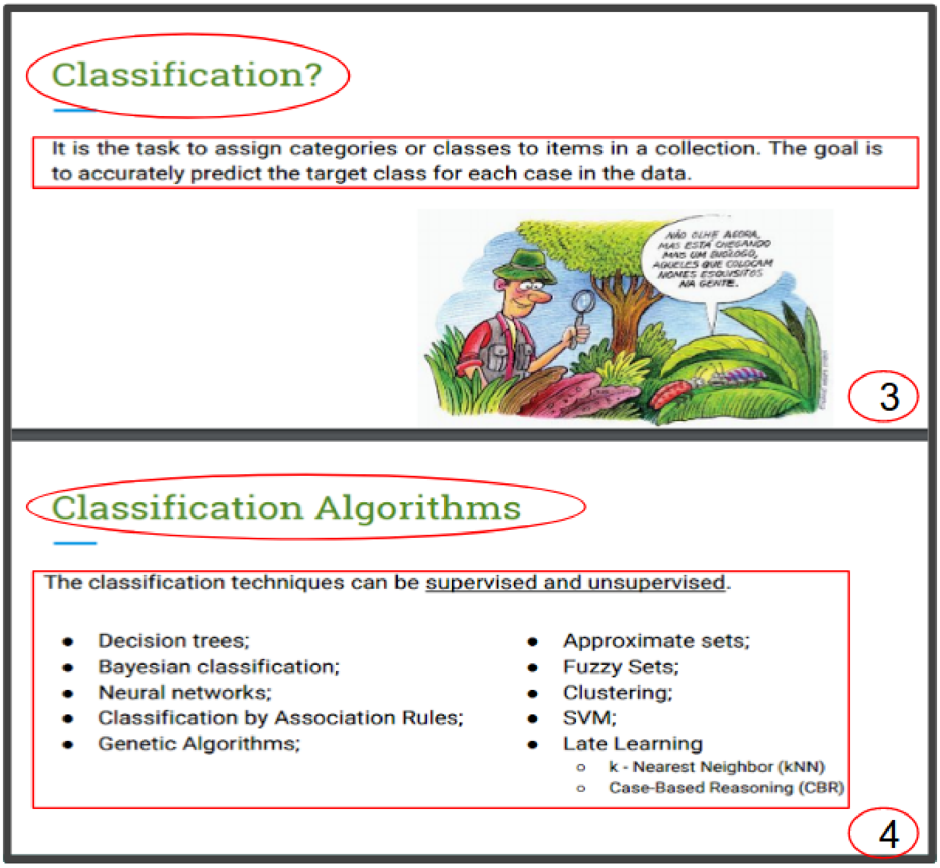
Step B creates the Intermediate Graph Representation adapting the concept of shadows [46] and stores this representation in a repository. The use of shadows enables the manipulation of parts of educational material without interfering with the material themselves. In the original work, shadows were implemented using XML files, but in our research we implement shadows in a graph format by the reasons already explained in Section 3.2.5.

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.

[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



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.

Classification Algorithms

The classification techniques can be supervised and unsupervised.

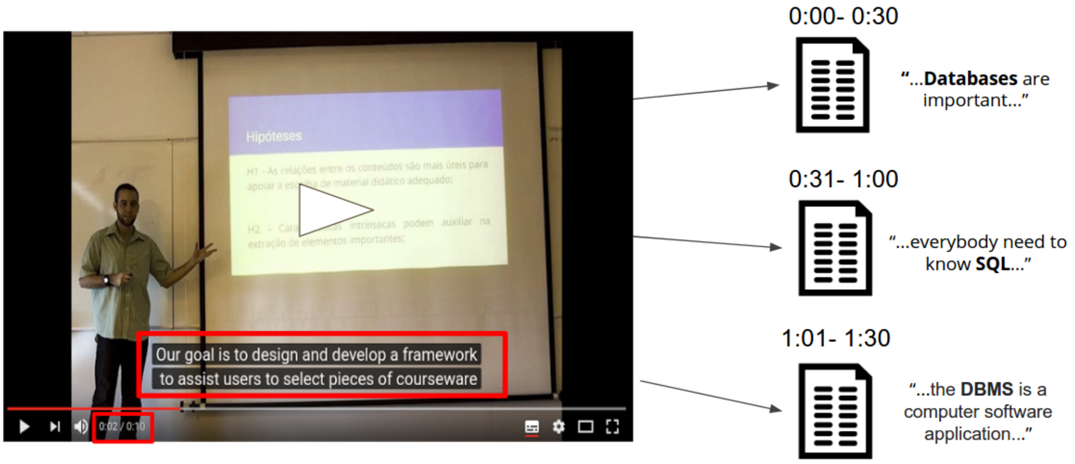
- Decision trees;
- Bayesian classification;
- Neural networks;
- Classification by Association Rules;
- Genetic Algorithms;
- Approximate sets;
- Fuzzy Sets;
- Clustering;
- SVM;
- Late Learning
  - k - Nearest Neighbor (kNN)
  - Case-Based Reasoning (CBR)

Figure 3.2: Components and text extracted from slides.

3.4.3 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.



Our goal is to design and develop a framework
to assist users to select pieces of courseware

Hipóteses
0:02 / 0:10
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 3.3: Components and text extracted from video subtitles.

3.4.4 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.

3.5 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.

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

3.6 Case Studies

3.6.1 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 [62].

3.6.2 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. Through this study, we were able to find similarities among different courses, thereby highlighting possible intersections, thus revealing potential multi-course activities.

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.

3.6.3 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)

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

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.

3.7 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.

Chapter 4

Analysis of Semantic Relationships among Educational Material via Graphs

4.1 Introduction

Educational material is produced in all scales and shapes, with different quality levels. Furthermore, the volume of such material easily outpaces the speed with which it can be analyzed and subsequently understood and correlated. This can be observed in many material-related repositories and sites such as the International Bank of Educational Objects[1], the ACM Learning Center and the ACM Techpack[2], the Coursera platform[3] or SlideShare[4]. Each repository offers users several options and some kind of search mechanism. However, none of these repositories offers means to analyze how topics among the stored objects are related. Furthermore, correlation of materials across repositories is not possible because, among other reasons, each of them stores and organizes contents in different ways. Thus, lecturers and students looking for materials across repositories (or even within one repository) usually have to rely on each repository's interface.

Moreover, each person has her/his own way of organizing contents. Thus, there is a need for what we call "analysis of different viewpoints" - namely, to support distinct kinds of grouping of contents. This can also help identify, e.g., how courses can differ across institutions.

In previous work [63], we proposed CIMAL (Courseware Integration under Multiple relations to Assist Learning) a computational framework capable of extracting and classifying the topics covered in several educational materials available in a variety of repositories. Students and lecturers can use this information to understand what each material addresses without having to manually access and read the entire contents of each of these materials. Moreover, our tools help create metadata for contents, thereby helping authors in rendering their materials findable. From now on, we will use the term

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

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

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

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

"authors" to denote the person(s) who produce and deposit their materials on repositories and "users" to refer to anyone (including the authors themselves) who want to find material that meets their educational requirements.

That work was important step towards helping users in finding "right" materials. Many issues remained, such as the need for understanding relations across materials, and catering to multiple perspective. This paper extends our framework, adding tools to assist users to find and analyze relationships among courseware contents in multimedia educational repositories, furthermore allowing analysis of different viewpoints. It is expected that these relationships will assist in the learning process and facilitate the handling of materials that are (indirectly) related to each other.

This paper describes how we dealt with some of the challenges that need to be faced regarding courseware management and retrieval. The first one is how to handle courseware that are not necessarily documented via metadata. Most authors do not publish their lessons using additional information that would help organizing them, which hampers search mechanisms.

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 of relationships. Yet another challenge concerns providing flexible search mechanisms, so that users can search for and navigate across contents, under multiple viewpoints.

As will be seen, our approach to deal with these challenges considers the following steps:

- Extract topics taking advantage of the intrinsic characteristics of educational materials (here summarized, for completeness take covered in the original CIMAL[63] paper);

- Derive relationships across the extracted topics with help of ontologies, and create graphs to correlate topics and courses, storing all this information in a graph database. The use of graph databases helps to progressively extend the graph to cover materials that are offered arbitrarily, and takes advantage of a graph DBMS' native support for querying and independence from underlying implementation details;

- Apply classic graph algorithms to correlate, search, navigate and extract multiple viewpoints.

The main contributions of our research are therefore: (i) a suite of algorithms to mine relationships across topics addressed in educational materials; (ii) creation of solutions for visualization of relationships of material in educational datasets to support the need for viewing multiple viewpoints; (iii) extension of the CIMAL framework, described in [63], to support different viewpoints over courseware; (iv) revisiting classical graph algorithms to analyze educational materials. Our proposal is showcased using material from distinct courses. In addition, we point out that, unlike related research, we process and correlate text from slides and video captions.

The rest of this paper is organized as follows: Section 4.2 presents the basics concepts behind our research; Section 4.3 presents an overview of CIMAL's architecture, and how it can be used to extract topics in materials; Section 4.4 presents details about how we construct graphs to relate educational material; Section 4.5 exemplifies how traditional graph algorithms can be used to help users find contents that meet their requirements; Section 4.6 describes the data repositories that we used in our case studies; Section 4.7 discusses the methodology used to implement the case studies and their validation; Section 4.8 presents related work and finally, Section 4.9 presents conclusions and future work.

4.2 Theoretical Foundations

4.2.1 Basic vocabulary

Following the terminology used in ACM's curricular guidelines 2013, we adopt the following:

- Class - a period of time in which students and lecturers are taught something;
- Topic - a matter dealt with in a class. A class may cover many topics;
- Course - we mean an institutionally recognized unit of study. Depending on local circumstance, full-time students will take several "courses" at one time, typically several per academic year. While "course" is a common term at some institutions, others will use other names, for example "module" or "paper".
- Program - is a set of courses that will eventually lead to a degree or a certification;
- Courseware /Educational Material - digital education material designed for use in an educational or training course, e.g. sets of slides and videos;

To illustrate these concepts, let us consider one of our case studies, from Coursera. It concerns using courseware from the program "Python for Everybody Specialization", provided by University of Michigan, and comprises slides and videos. It is divided in 5 individuals units (courses): "Programming for Everybody (Getting Started with Python)", "Python Data Structures", "Using Python to Access Web Data", "Using Databases with Python" and "Capstone: Retrieving, Processing, and Visualizing Data with Python". Each course is divided in some weeks (classes), for example, the first course has 5 classes. In these 5 classes, students will study various topics, including: "Python programming language", "Install Python and write programs", "Use variables to store, retrieve and calculate information", "functions and loops".

4.2.2 Graph Databases

We chose to take advantage of graph databases as persistent storage for all the materials, topics and relationships studied in this research. First, databases are a good choice in a situation such as ours, in which we want to support progressive growth of knowledge extracted from materials, and have support for low-level storage management and query products. The second reason for choosing graph databases is our need to extract and navigate across relationships, which are natively provided by such databases. In other database models (e.g. relational) relationships need to be constructed via queries. Graph databases, on the other hand, explicitly store these relationships, which can grow arbitrarily, providing smooth extension - an important requirement in a case such as ours, in which we desire to progressively add new courseware to our database. Graph databases allow to represent information about the connectivity of unstructured data.

In more detail, the graph data management paradigm is characterized by using graphs as data models and graph-based operations to express data manipulation. Graph databases are usually adopted to represent data sets where relations among data and the data itself are at the same importance level [3]. Therefore, this type of database is recommended for our research, since we want to store and study the relationships across (topics of) educational materials. This being said, there is a wide variety of such databases, each of which supports representation of graphs using different structures.

The formal foundation of all graph data structures is based on the mathematical definition of graphs. On top of this basic layer, several graph data structures were proposed [56, 43]. One of the most popular structures supported by many graph database systems is the *property graph*. It tries to arrange all the features that these graph types express in a single and flexible structure through key-value pairs to describe nodes and edge characteristics, such as type, label or direction [18].

Both nodes and edges can have any number of properties associated with them. This data structure represents a multi-relation graph, since two given nodes can be connected via many edges, each of which represents a distinct relationship between the nodes. This flexibility is one of the advantages of property graphs as opposed to other graph structures, which either do not support multiple edges or have little flexibility assigning properties to edges and nodes. For this reason, we have chosen a graph database management system that natively supports property graphs to model and store courseware information, the Neo4J[5] that are widely used in research.

4.2.3 Viewpoints

The notion of viewpoints of study appears in many educational studies [30, 54]. Each user has his/her own perception of a topic and its relevance to an educational goal. In other words, courseware may have several viewpoints, each of which representing how it can be analyzed, visualized and interpreted.

By the same token, a data item can be interpreted in distinct ways. Two videos about Programming Concepts, for example, could be recommended for a study that needs low level details about "functions and variables" or "programming paradigms". From a macro viewpoint, these two videos may also be studied to check differences between materials produced by distinct authors.

In our research, we will only analyze viewpoints referring to the topics present in educational materials. Thus, a viewpoint can be understood as the creation of information

[5] https://neo4j.com/

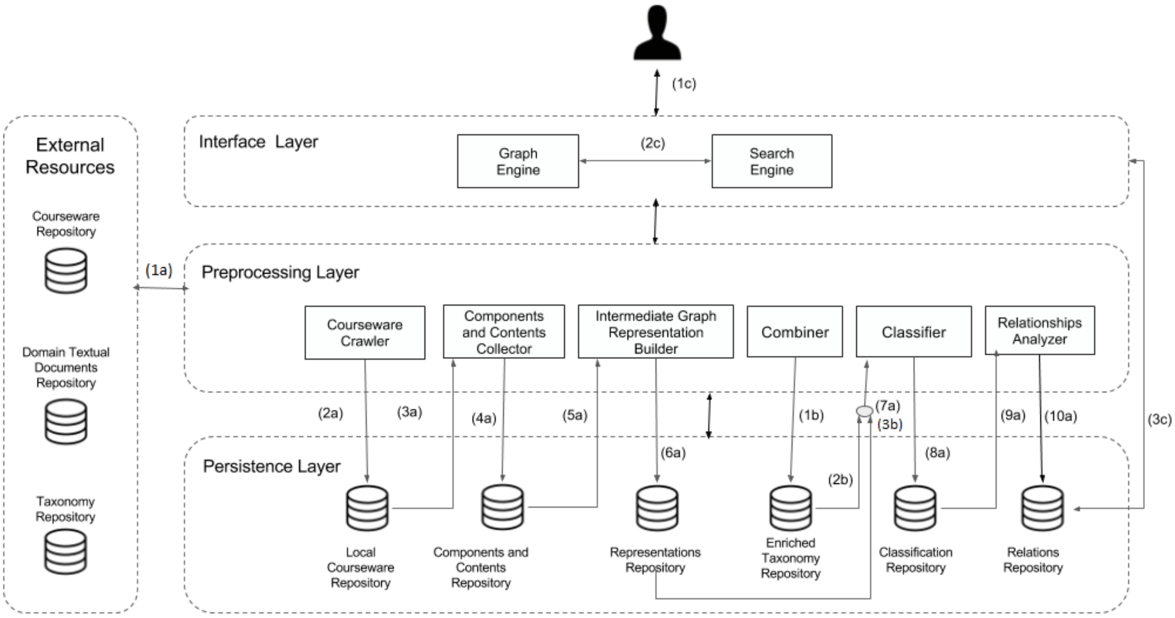


Figure 4.1: System Architecture for Analysis of Relationships among Educational Material Contents.

derived from rearrangements of topics extracted from such material.

Related work has also been concerned with allowing analysis of different points of view among digital contents, such as [66, 44] . However, they need to build hierarchical structures and have not developed solutions for static (slides) and dinamic (videos), as is our case. The next section presents an overview of CIMAL, explaining its main functionalities.

4.3 Extraction of topics from educational material

This section gives an overview of our previous work, [63] in which we describe our Courseware Integration under Multiple relations to Assist Learning (CIMAL) framework. CIMAL assists identification of relevant educational materials by extracting and ranking topics mined from such materials. Figure 4.1, copied from [63], illustrates the main features in the framework.

CIMAL 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 external sources for subsequent search. It provides all the services needed to look for materials using topics. These services can be accessed through the *Interface layer* by users (lecturers and students).

The repositories need to be prepared (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 based on 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 texts produced by experts from the *Domain textual documents Repository* (1a). Taxonomy and texts are unified in an *Enriched Taxonomy Repository* (1b). In the Enriched Taxonomy, each concept from the original taxonomy has a reference to a text by experts. Users can determine the appropriate taxonomy-document "pairs", thereby allowing personalization and anonymization of materials.

Once representations and enriched taxonomies are created, the *Classifier* is ready to extract the topics covered in each of the materials (2b,3b,7a). This information is then stored in the *Classification Repository* (8a). We point out that (7a) plays an important role in CIMAL , since it creates the final graph structure for subsequent topic classification.

Lastly, the *Relationships Analyzer* looks for relationships among the items and their topics in the Classification Repository (9a), and stores these relationships in the *Relations Repository* (10a). All preprocessing steps must be performed whenever we add educational material, taxonomy or texts from a domain textual base.

After preprocessing steps 1 through 9, users 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.

In this paper we go a step ahead. We start from the topics extracted and classified, and construct courseware graphs. Next, we show how we construct multiple viewpoints and relationships across the corresponding materials. We used the methodology described in [63] to classify the topics of slides and videos.

4.4 Creating a courseware graph

This section describes how we construct courseware graphs based on topics extracted from such courseware using CIMAL, as explained in section 4.3. Our graph has two kinds of nodes, to represent materials and topics - there is one node for each material, and one node for each topic. Graph edges link each material to the topics that was extracted from these materials. This means that there may be more than one edge associated with a topic node (occasionally more than one material refers to that topic). Edges linking materials and their topics are labeled with the "mentions" property.

Next, we link topic nodes to each other. In [63], to analyzing computing courseware we classified topics using the poly-hierarchical ontology from the Classification System of the Association for Computing Machinery (ACM CSS 2012). Now, through this ontology, we created edges with "broader", "related" and "narrower" properties that link topic nodes. These properties provide information about which topics are related. Using this

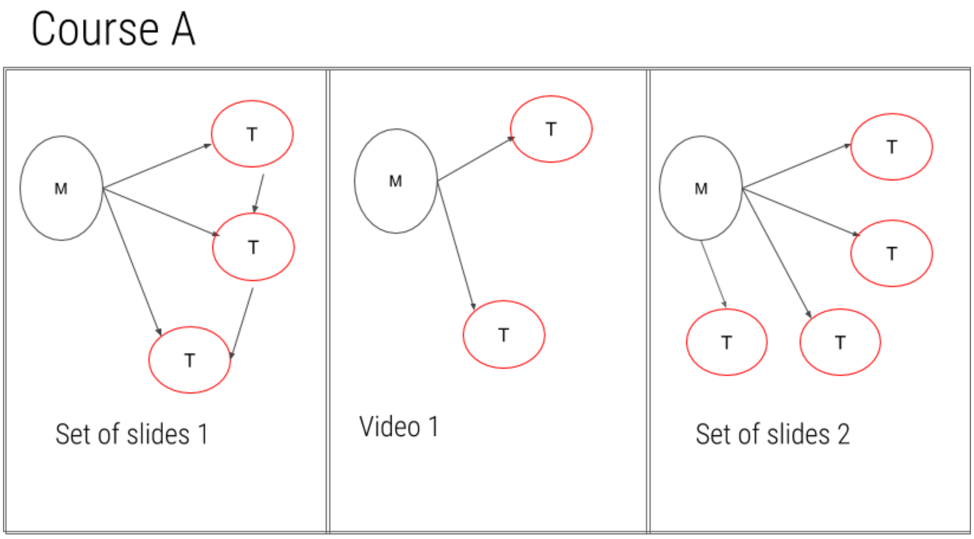


Figure 4.2: Materials of a Course A. M - name of a material, T - topics extracted. Also, each edge is labeled with a property.

information it is possible to perform many kinds of search operations - e.g, search for a particular topic or to browse more generic or specific topics for a better understanding of a course. We stress that topics are elicited from textual contents of the materials, and not from eventual metadata provided by authors.

Figure 4.2 illustrates how a set of graphs is created for a given course - one graph per set of slides or video. Figure 4.3 depicts the final graph constructed correlating the distinct materials. It shows materials from different courses that may have topic relationships.

Figure 4.4 shows part of a property graph that illustrates build from a hypothetical database courses showing how classes and topics are related (within one course) through edges with properties "broader", "narrower", "mentions" and "related". For instance, the figure shows that the courseware "Introduction to Databases" mentions topic "SQL", which in turn is related to topic "Databases History" and narrower than topic "DBMS".

These properties also allow analysis of many kinds of relationships. Examples include linking materials whose topics are not directly related, or grouping materials by generic themes. Thus, one can produce many different viewpoints on the same data set.

4.5 Graph Algorithms

Given a courseware graph, we can now run graph algorithms such as "Clique", "Centrality", "PageRank" or "Short-Path" to observe new kinds of correlation among the corresponding materials, and to create multiple viewpoints.

This section shows how each of these algorithms can be used to perform these analyses. To illustrate the application of the algorithms, a graph was created with materials from the program "Systems Analysis and Development" (Figure 4.5) described in [63].

The materials were grouped by courses[6], constituting 14 different courses represented by the 14 nodes in blue; the labels of the edges between the nodes indicate the similari-

[6] As defined in section 4.2

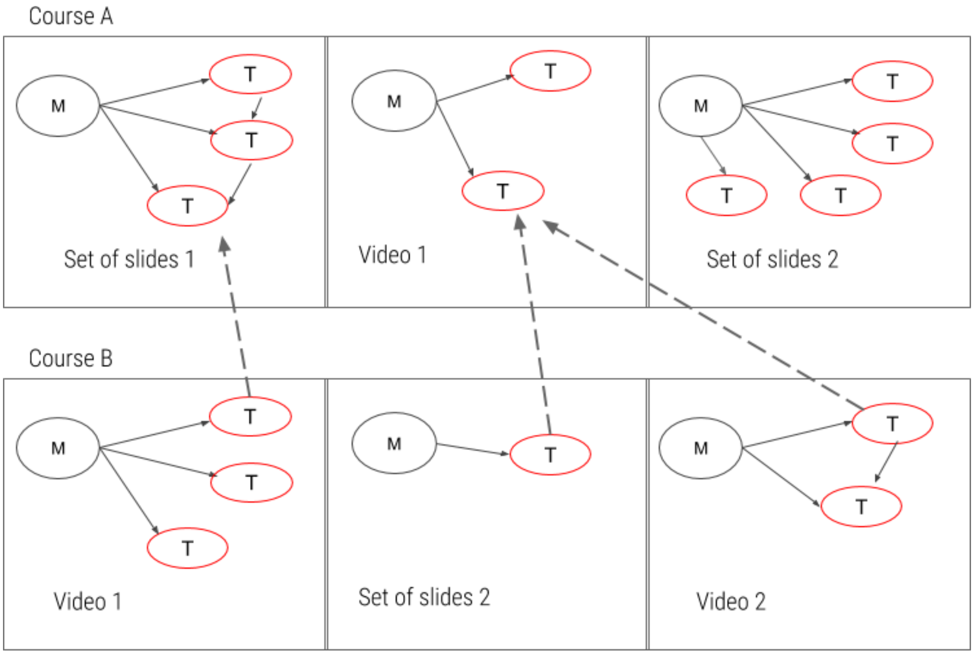


Figure 4.3: Materials of different courses can be related by topics. Edges linking courses (e.g., A and B) also have labels.

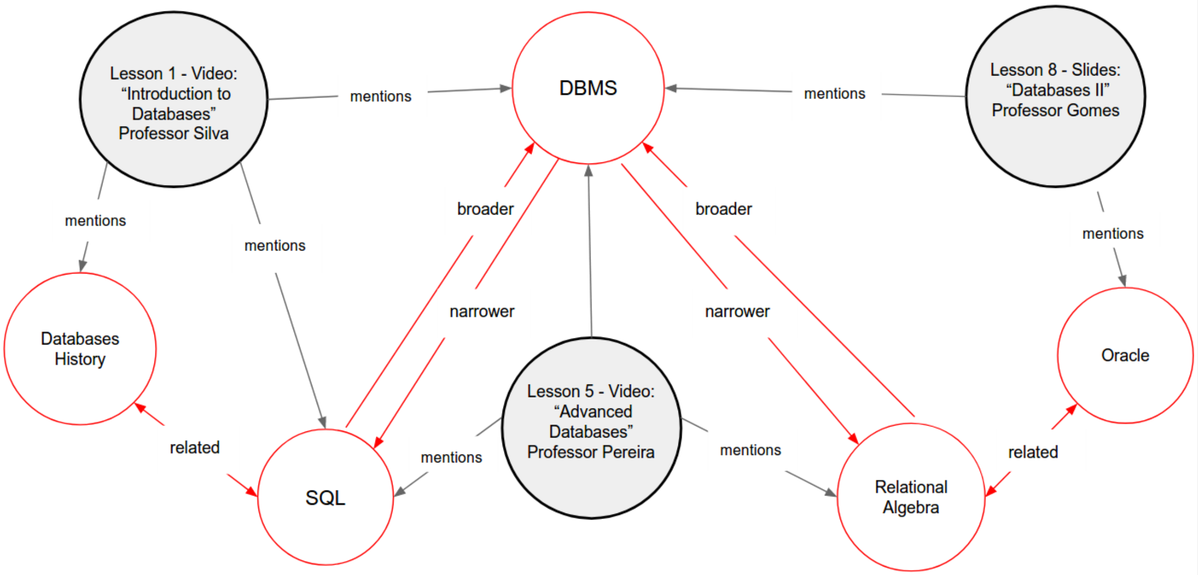

Lesson 1 - Video: "Introduction to Databases" Professor Silva
mentions
DBMS
mentions
Lesson 8 - Slides: "Databases II" Professor Gomes
mentions
mentions
broader
broader
mentions
narrower
narrower
Databases History
Oracle
Lesson 5 - Video: "Advanced Databases" Professor Pereira
related
mentions
mentions
related
SQL
Relational Algebra

Figure 4.4: Adding edges to allow various viewpoints on topics and materials

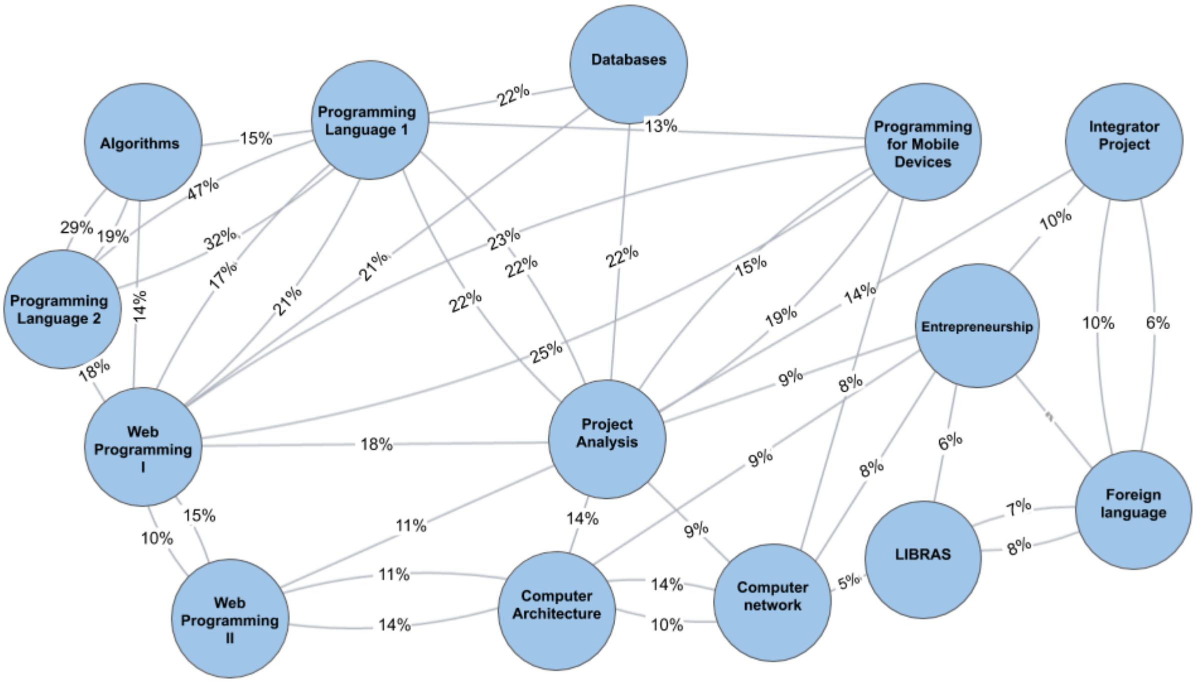


Figure 4.5: Graph to illustrate the Systems Analysis and Development program

ties between the topics covered in a course and the topics covered in other courses. For example, the topics in course "Programming Language 1" are 47% similar to the topics in course "Programming Language 2" (arrow from "Programming Language 1" to "Programming Language 2"), while the topics in "Programming Language 2" are only 32% similar to those in "Programming Language 1" (opposite arrow).

4.5.1 Clique

A clique, C, in an undirected graph G = (N, E) is a subset of the nodes, C contained in V, such that every two distinct nodes are adjacent [40]. This is equivalent to the condition that the induced subgraph of G induced by C is a complete graph.

By applying a clique algorithm, we can define a set of educational materials that have relationships with each other and can thus be seen from a more abstract viewpoint as important materials for a given course or program.

Figure 4.6 illustrates a clique found among the courses analyzed. It shows that courses "Programming Language I", "Programming Language II", "Algorithms", "Web Programming I" form a clique.

4.5.2 Centrality

Centrality algorithms are used in many different fields to identify influential concepts. For instance, in social networks it is possible to identify the most influential person(s), or sources of dissemination of diseases, or the key infrastructure nodes in urban networks. Here we use centrality to identify the "most important" educational material in a set. The

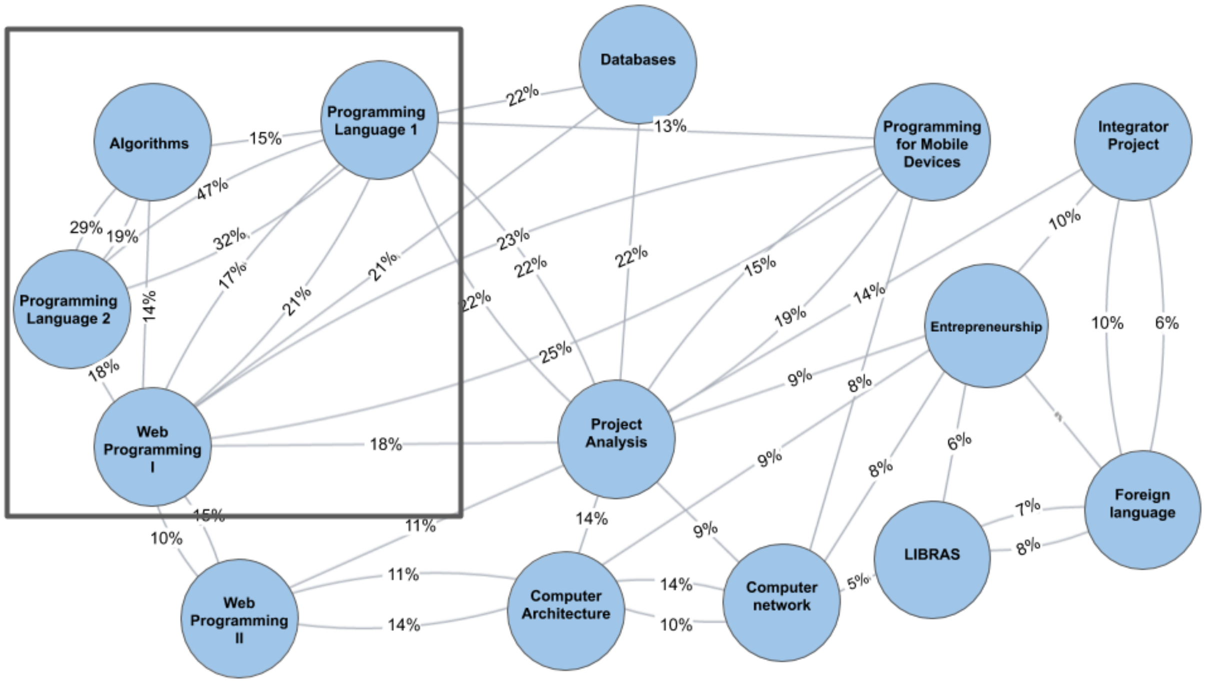


Figure 4.6: A clique - four courses with at least one topic in common between every pair, based on Figure 4.5.

underlying notion is that the most important material is the one that creates a "bridge" between the largest number of topics.

There are several algorithms to obtain the centrality of a graph. In this study, we chose the Betweenness centrality [23], because through it we can quantify the number of times a material acts as a bridge along the shortest path between two other materials. According to Freeman [23], the nodes that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness. Hence, we use betweenness to visualize the degree of importance of a material for the understanding of a course or even a program. In proposing the use of this algorithm, we can moreover find "bottlenecks" in curriculum plans. In Figure4.7, "Project Analysis" is the "central course" in the program "Systems Analysis and Development".

4.5.3 PageRank

PageRank, described in [12], is an algorithm that is used to know the individual importance of an element in a graph. Here, we use PageRank to detect the relative importance of all educational materials in a course. In our research, PageRank works by counting the number of graph links to a material to estimate how important the material is in a course. The underlying assumption is that more important materials are likely to receive more links from others. Usually a material that covers more topics and more fundamental issues of an area is connected with more materials. The PageRank algorithm outputs a probability distribution used to represent the likelihood that a user randomly clicking on links will arrive at any particular educational material.

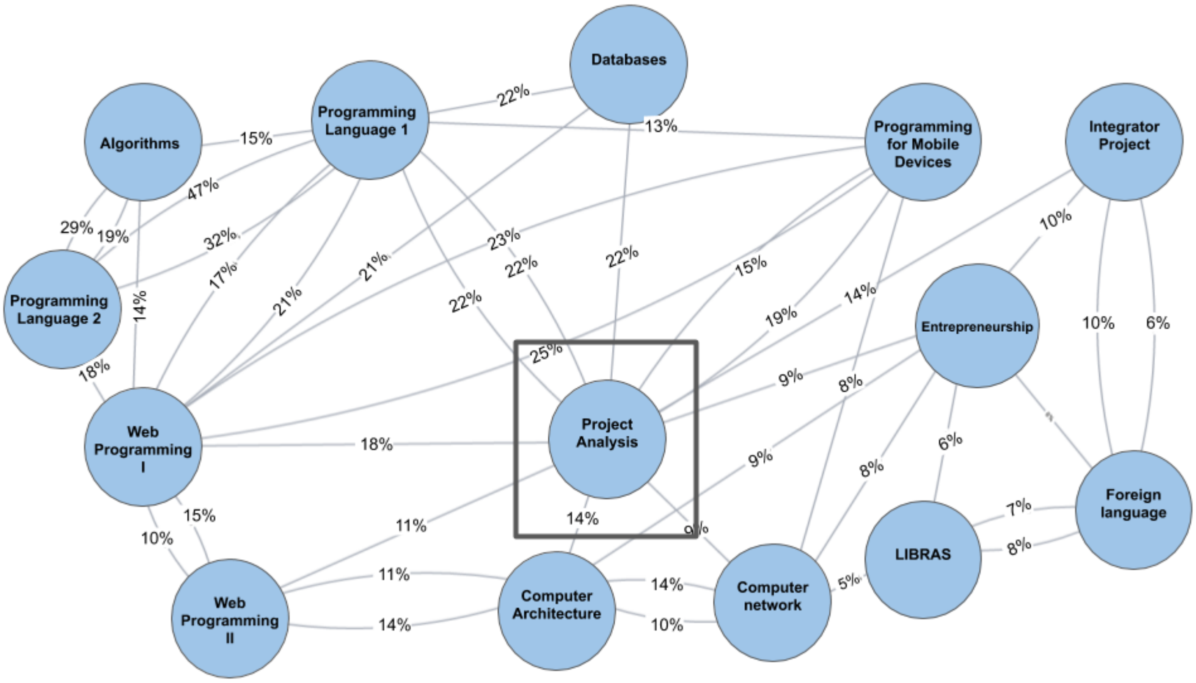


Figure 4.7: The course highlighted by the centrality algorithm.

Figure 4.8 shows the three main courses (groups of materials) of the program, according to PageRank algorithm. The number one course of the program is "Programming Language I", the second is "Project Analysis" and the third is "Web Programming I", because they have more topics referenced by other courses.

4.5.4 Shortest-Path

We have observed that people often try to choose the least amount of courses they should study to understand a certain course. Given an educational material E, what is the smallest set of materials that should be studied to understand a certain topic in E?

We map this to the problem of finding the shortest path in a graph from one node to another. To us, the "shortest" is the path with the least number of edges, which results in fewer materials to study. Since the edges do not have weights, we can adapt the famous Dijkstra algorithm to indicate the best path between two materials.

Figure 4.9 illustrates the shortest-path (smallest number of different topics) between the courses "Web Programming I" and "Programming for Mobile Devices". According to this path, After "Web Programming I", a person should understand the topics of "Programming Language I", then "Project Analysis" and finally learn the topics on "Programming for Mobile Devices". Running this algorithm, users can create their study plans more efficiently.

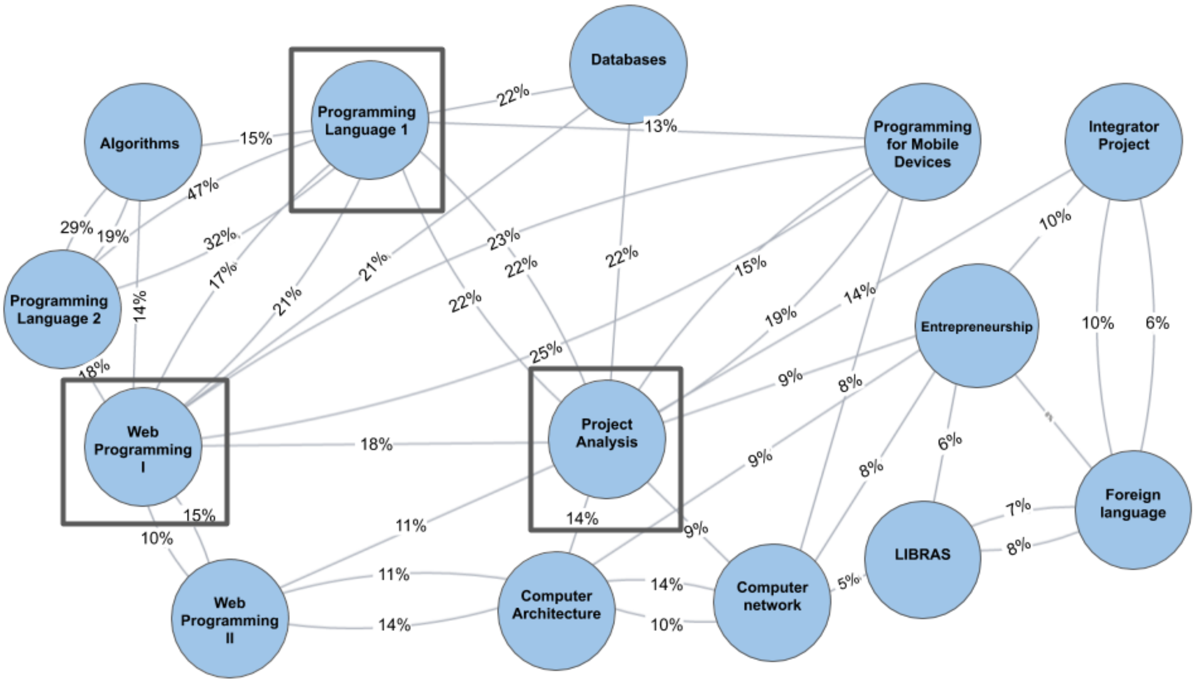


Figure 4.8: The top-3 of all courses in Figure 4.5

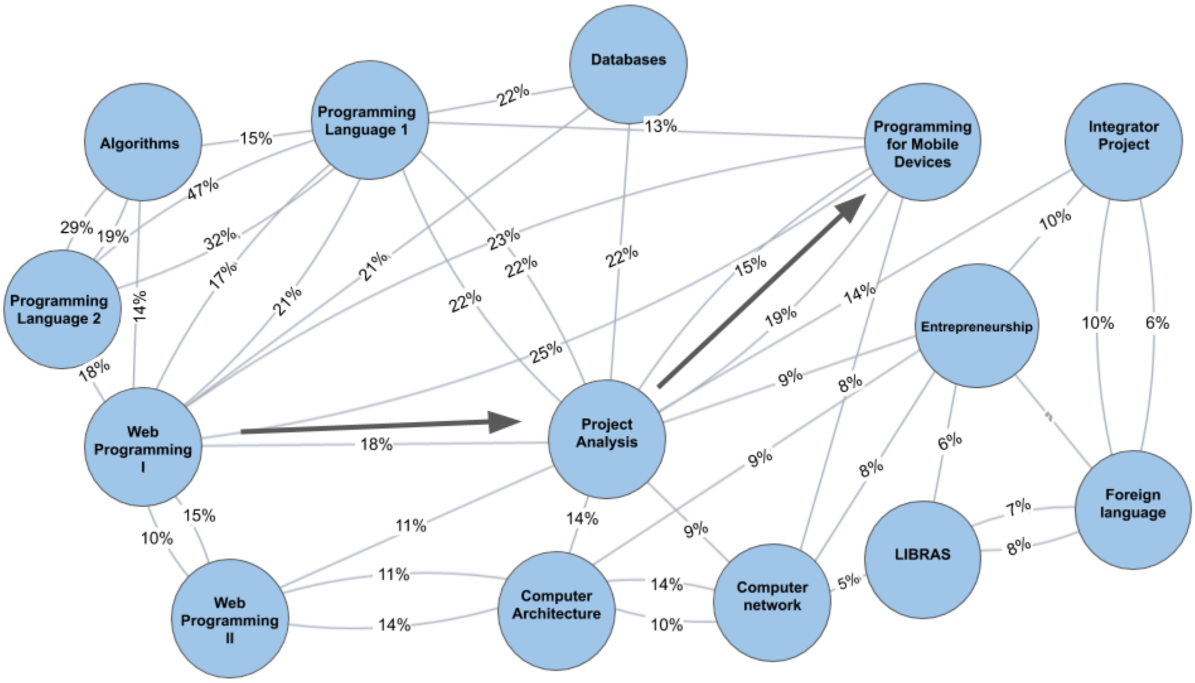


Figure 4.9: Shortest-Path between the courses Web Programming I and Programming for Mobile Devices.

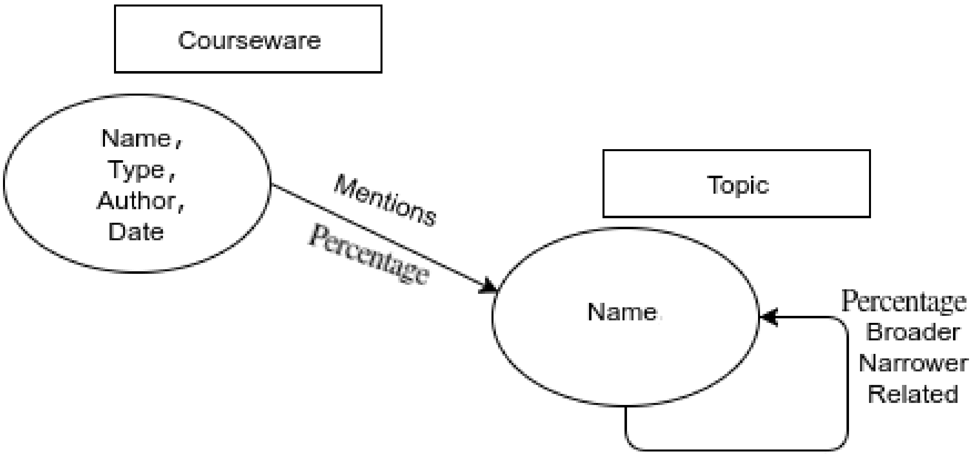


Figure 4.10: Schema of graph database.

4.6 Experimental Set Up

This section presents details of the data repositories used to showcase our work. Although graph databases have no explicit schema, there is an implicit schema - e.g., in related work we find the terms "reference graph" [10] or "metamodel" [32]. Based on this, we define a schema of a graph database for courseware. Figure 4.10 depicts the overall schema for a courseware graph.

Our schema describes the semantic organization of all modeled information, specifying entities, which relations are valid for each entity, and what kinds of properties are relevant. Here, nodes are educational materials or topics and edges are labeled with properties "mentions", "broader", "narrower" and "related". Table **??** describes the properties of nodes (Courseware and Topic) and Edges.

Figures 4.5 through 4.9 are actual screen copies of running Neo4j/CIMAL queries for our case study. They were used to help faculty understand how to interact and improve them educational materials. For legibility sake, we increased node size and emphasized arrows.

4.7 Designing the methodology to validate our work

As presented in section 4.3, the first step of our approach is the extraction of topics from educational materials. To check whether the generated graphs and operations are appropriate, we first need to verify that CIMAL adequately extracts topics and classifies materials. We thus carried out the evaluation of our solution by specialists in Computer Science.

The validation experiments were performed between the months of February and April 2018. We point out that there was no validation of CIMAL. Rather, experiments were conducted on the use of CIMAL for specific applications.

Using social networks (namely, Facebook) and personal contacts, we selected 28 people who had at least an undergraduate degree in Computer Science or related fields. We then had these people to recruit acquaintances with the same kind of qualification, using the so-

Graph entity	Properties	Description
Courseware	Name	name given by the producer of the material.
	Type	digital format of the material, in this research all material are "sets of slides" or "videos".
	Author	developer of the material.
Topic	Name	name of topic according to ACM CSS 2012.
Edge	Percentage	percentage of the topic covered by the material.
	Mentions	relationship between a material and a topic.
	Broader	Topic "A" have a broader relation with topic "B" if this is a topic more generic than "A".
	Narrower	Topic "A" have a narrower relation with topic "B" if this a topic more specific than "A".
	Related	Topic "A" have a related relation with topic "B" if both topics have a broader relationship in common.

Table 4.1: Properties of nodes and edges.

called snowball sampling [48]. From a total of 50 individuals, thus selected, 37 completed the questionnaire created to evaluate CIMAL.

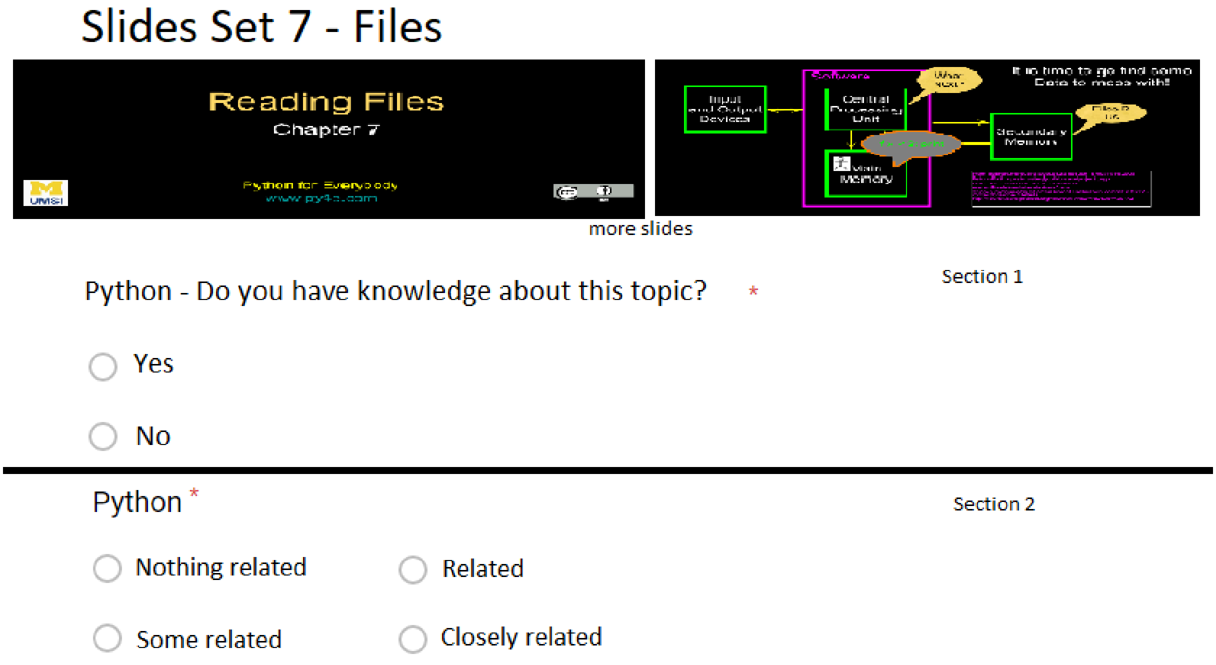
Validation process. After selection of the experts, an invitation was sent by e-mail, explaining the research objectives. Upon acceptance of the invitation, each expert was emailed a web link to two questionnaire: one to characterize his/her educational background and another questionnaire to evaluate our solution. Participants were asked to answer the questionnaires within 7 days.

The second questionnaire contained questions to analyze the relationships and topics highlighted by CIMAL from the "Python for Everybody Specialization", provided by University of Michigan, and to evaluated the degree of relevance of topics and relationships. Participants were also able to send suggestions for modifications.

4.7.1 Questionnaires

The first questionnaire contained questions related to name, level of English comprehension, educational level, area of study / work, university where studied. The second questionnaire, the CIMAL Analysis Protocol, was organized in two sections: the first was related to the topic content of each educational material and the second section referred to the topic relationships highlighted among all the material.

The first section contained two parts: the first was related to the participant's level of knowledge about a topic and the second asked the participant to evaluate whether this topic was related to a given educational material as indicated by CIMAL. In the second section, the questions dealt with relevance of relationships. Questionnaire multiple choice answers were based in Likert scale [37] (where DT = totally disagree, D = disagree, C= agree; and CT = totally agree). In section 1, the items indicate whether the topic is "Nothing related", "Some related", "Related" or "Closely related" to the topic addressed in the material. Figure 4.11 shows a screen copy of part of questionnaire where in section 1 it is asked if the person has knowledge about the topic "Python", and section 2 is asking if the material "Slides Set 7 - Files" has a relationship with the topic "Python".



Slides Set 7 - Files

more slides

Python - Do you have knowledge about this topic? *

Section 1

○ Yes

○ No

Python *

Section 2

○ Nothing related ○ Related

○ Some related ○ Closely related

Figure 4.11: Screen copy of part of questionnaire

4.8 Related work

Several educational studies use data mining techniques for knowledge discovery, decision-making, and recommendation [22, 70]. These studies gave rise to the educational data mining (EDM) research field [5]. EDM emerges as a paradigm oriented to design models, tasks, methods, and algorithms for exploring data from educational settings [50].

According to Pena [50], research in EDM can be divided in 7 categories: Student Modeling; Student behavior modeling; Student performance modeling; Assessment; Student support and feedback; Curriculum, domain knowledge, sequencing, and teachers support; Tools. Our research can be classified in three of these 7 categories.

a) Tools: We designed, developed and tested tools to assist with tasks associated with selection of materials for learning.

b) Curriculum, domain knowledge, sequencing, and teachers support: We have created means that support the decision making of lecturers. For example, they can analyze which topics are covered in materials used in courses and so plan course updates.

c) Student support and feedback: We have designed and implemented a system to allow students to easily find materials that address a particular course and their relationships with other materials. For example, a student who has difficulty with a topic X can study other materials with Y and Z topics related to X to decrease the difficulty s/he has to understand X.

To facilitate our study of related work, we classify how each of them uses EDM techniques to assist teachers and/or students. The result of this classification appears in table 4.2. The columns on the table indicate the main goals of related work. Column 1, "Create tools" corresponds to research in which computational tools are created to

Research	Create tools	Solutions to teachers	Solutions to students	"Learning paths"	Organize information
[15]			X	X	
[41]	X	X			X
[29]		X	X		
[26]	X		X	X	
[17]		X	X		
[65]			X	X	X
[69]	X		X	X	
[4]			X		
[6]	X				X
[28]		X			
[36]		X			X
[20]	X	X			X
CIMAL	X	X	X	X	X

Table 4.2: Related work

support education. A related work can focus the solution to assist teachers (Column 2) or students (Column 3). Finally, solutions can indicate a sequence of actions to create a plan or study strategies (Column 4), or just organize information into computational structures (Column 5).

Research that creates tools (such as [41, 26, 69, 6, 20]) does not focus on the analysis of the educational materials used in courses. On the other hand, research that focus on the support to the teachers most often use classification techniques, clustering and mining techniques to facilitate analysis (such as [41, 29, 17, 28, 36, 20]). However, no work was found that seeks to use graphs and perspectives on topics for analysis of materials.

To assist and give feedback to students, the research found in the literature seeks to classify, group students and create association rules through probabilistic algorithms and machine learning with the purpose to express: suggestions, requests, and evaluations (such as [15, 29, 26, 17, 69, 4, 65, 36])

Research involving "learning paths" generally use Artificial Intelligence algorithms [15, 26, 69, 65], and need access to student access logs, something that we did not have access to in this study. Indeed, the focus is different. Related work tries to find learning paths, whereas we suggest them.

It is also very common to find a "training phase" in these search mechanisms, which must be run again every time new data is inserted into the database. Our solution is "training free", that is, an already classified material need not be re-evaluated, unless the ontology used is completely altered.

Ours is not the first time that ontologies or data from courseware are used to organize information in Education Data Mining [41, 65, 6, 36, 20]. However, in these studies, this information are used to model the students' study profile or materials, but nothing is done to organize educational materials of different formats as in our work.

4.9 Conclusions and Future directions

This paper presented a solution to enable students and lecturers to analyze multi-viewpoints for topics from educational materials. Our work also showed new uses for classical graph algorithms and validated our classification of material presented in slides and videos. Validation also showed that the classification performed is useful for relationship analysis.

Among future directions, 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 about user accesses to these materials. For example, data on the last courses that a student held could be used to construct a personalized study guide on courses that would be interesting for this student; the recommendation system could also recommend more courses (as discussed in [52]). It would also be interesting to compose recommendation paths to the path actually taken by people to study courseware.

Another extension is the inclusion of other media in CIMAL's topic extraction algorithms, such as audio recordings, books and figures. Also, a module for viewing maps can be implemented to support analysis of 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.

Chapter 5

Final Conclusions and Extensions

5.1 Overview

The research presented in this thesis concerns challenges in Educational Data Mining, in particular to overcome the problem of extracting and correlating topics of heterogeneous educational material.

Our motivation came from interdisciplinary and multiviews research, where lecturers and students need many different views from the same set of materials and topics. This scenario usually requires complex algorithms or extra tasks from lecturers (e.g., annotations). To help them in these tasks, our work combines NLP techniques, taxonomies and graph databases to a handle a wide range of demands for managing heterogeneous contents and formats of educational material.

The semi-structured nature of data, such as slides and videos, and the high level of importance of data connections to visualize relationships across topics, led us to adopt the graph data management paradigm.

Based on this approach, we specified and implemented a prototype of a framework that elicits and manages relationships among topics of educational material, named CIMAL. The specification of our operators and framework are as generic as possible and can be implemented in different graph database engines and programming languages.

Also, different real world datasets of educational material were analyzed to validate our research. The case studies can clearly benefit from our framework to analyze materials. We also believe that our solution can be extended and adopted by other kinds of application domains with similar management and analysis requirements.

Our design and implementation considered a few key factors. For instance, our classification schema (Classifier module) were based build in in invoking the Explicit Semantic Analysis (ESA) component of Lucene. We chose ESA because it is a well-known algorithm to perform semantic analysis in Natural Language Processing. We had to tune it with the appropriate parameters, which we defined based on our analysis of the specific case studies. For instance, we set the same weight to all words, for lack of more information.

The Combiner creates an enriched taxonomy by directly linking each taxonomy to the relevant document. We point out that this is generic, and thus other taxonomies and domain documents would support other domains.

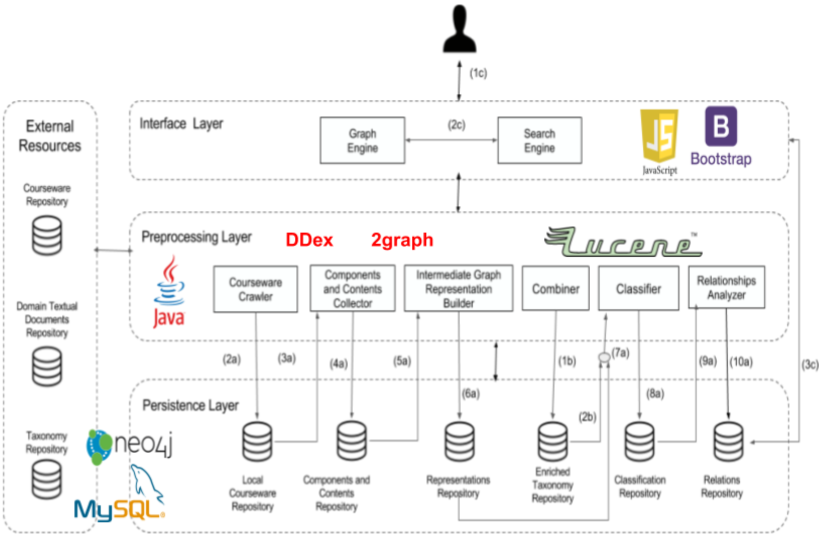


Figure 5.1: Tools used to implement CIMAL.

We extracted components using DDEX[1]. It must be pointed out that this may require making such adaptations when this identification is not automatically feasible.

Figure 5.1 revisits our System Architecture indicating the tools used to implement CIMAL.

5.2 Contributions from each chapter

The thesis contains contributions both to Computer Science and to users who need to find relevant material. Computational contributions include the choice of modules to implement CIMAL, the architecture, and construction of additional structures, in particular via the Combiner. Contributions for educational purposes include building a solution that does not require training sets and annotations, enabling the use of heterogeneous materials (slides and videos), facilitating material documentation. Furthermore, as shown in our second case study through CIMAL we were able to help lecturers in sharing materials and collaborative activities, and support vocabulary standardization.

Our first contribution, presented in Chapter 2, was to present the area of Educational Data Mining, and algorithms, e.g. Topic Modeling and Topic Labeling, that were not used for semi-structured materials such as slides and videos. This chapter introduced our first ideas to create an approach for extracting slide and video topics using only the data present in those materials. This contribution is related to Challenge 1 (described in Chapter 1), because our study uses slides and videos, different types of courseware.

The second contribution, introduced in Chapter 3, is the formalization and implementation of an architecture to view relationships between topics, which combines graphs, taxonomies and a Natural Language Process (NLP) algorithm, called Explicit Semantic Analysis (ESA). Using these techniques and graphs, we face Challenge 2 (described in Chapter 1), i.e., identification of relationships.

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

The third contribution of this thesis, presented in Chapter 4, is to propose a property graph data model with a set of operators and graph algorithms to manipulate data about relationships of courseware topics. This is the main contribution of our thesis, proposed to fill the gap of the lack of means to understand how topics of educational materials can be related and thus facilitate the process of learning and updating teachers and students.

The fourth contribution, formalized in Chapter 4, is to define viewpoints for the graph in educational materials and topic to support the need for multiple perspectives in interdisciplinary research. We show how viewpoints can be specified through classical graph algorithms.

Our last contribution, presented in Chapters 2, 3 and 4, is to analyze real life examples of courses and educational material and how they can benefit from our proposal. Using the case studies presented in these chapters, we show how relationships among topics can be explored by experts, lecturers and students using graph databases, pointing out the advantages of this approach.

Last but not least, let us revisit our challenges under the light of our work:

- **Challenge 1: Integration of different types of courseware.** In our case studies we show that CIMAL can be used to integrate slides and videos. Since we adopted graph structures to store courseware information, our solution can be extended to support other types of courseware.

- **Challenge 2: Identification of relationships.** We elicit the topics of each courseware and create graphs to analyze relationships among these topics. Also we show how to search for and navigate along related content using graph algorithms.

5.3 Extensions

This research can be extended to different practical/implementation and theoretical aspects. Some possibilities are:

- Investigate how to perform adaptations in our framework to improve performance, i.e., to use less computational resources or to reduce execution time.

- Develop a graphical user interface for the CIMAL framework.

The graph data structure is often better understood in a visual way. Analogously, to provide a graphical interface for view definition and exploration would improve our prototype, make it more interesting to stakeholders.

- Experiment with other NLP algorithms.

ESA has extensions that could be useful for this research, such as Cross-language explicit semantic analysis (CL-ESA) [53], a multilingual generalization of ESA, which would allow the visualization of relationships between materials produced in different languages.

- Extend CIMAL to include visual analysis of videos and images.

Some educational materials use plenty of illustrations. In particular, lecturers film classes in which they make drawings on the blackboard or use objects in the classroom. All this information was not used in our research and could assist in the extraction of topics in set of slides and videos.

- Gather new requirements from other study domains

Due to the complexity involving heterogeneous datasets of educational material, our thesis delimits a scope and a list of research problems to deal with. Indeed, many other requirements can arise outside this initial scope, for example, the need to understand perceptions of topic in areas such as Health or Arts. We limited our work to Computer Science courses. For this, among others, it would be necessary to create new taxonomies that involve such areas. As well, this would require validation and interaction with users from these domains.

- Design adaptations in our framework to explore geographic aspects of data.

Other relationships can also be constructed, e.g., express some kind of spatial correlation about data records. Thus, geographic coordinates in sources of educational material can be used, for instance, for analyzing educational institutions and recommending cooperation between them. Another related extension would be to include temporal predicates, to analyze the evolution of a set of topics in courses over the years.

- Extend relationships to other situations.

We implemented the relationships proposed by the ACM Computing Classification System[2]. Other domains may require distinct relationships which may need additional algorithms for their extraction. Also, additional relationships may be included, thereby sophisticating the subsequent analyses. This would require domain knowledge.

- Explicit collective intelligence.

Yet another improvement would be to take advantage of user feedback to adjust and improve relationships and topic extraction. This would help, e.g., to improve the taxonomy. This would require including, a new module into CIMAL.

- Adapt CIMAL for recommendation

Many possibilities exist to include modules in CIMAL for, e.g., recomendation of learning paths. Additionally, by collecting user profiles, our proposal would play a role in predicting user profiles and requirements. In this, again, user feedback would need to be recorded.

[2] http://www.acm.org/about/class/

Bibliography

[1] Parvin Ahmadi, Mahmoud Tabandeh, and Iman Gholampour. Persian text classification based on topic models. In *ICEE'16 24th Iranian Conference on Electrical Engineering*, pages 86–91, 2016.

[2] Mehdi Allahyari and Krys Kochut. Automatic Topic Labeling using Ontology-based Topic Models. In *ICMLA'15 IEEE 14th International Conference on Machine Learning and Applications*, pages 259–264, 2015.

[3] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Computing Surveys (CSUR)*, pages 1:1–1:39, 2008.

[4] AniGrubisic, Slavomir Stankov, and Ivan Peraić. Ontology based approach to bayesian student model design. *Expert Systems with Applications*, pages 5363 – 5371, 2013.

[5] Anjo Anjewierden, Bas Kollöffel, and Casper Hulshof. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *ADML'07 International Workshop on Applying Data Mining in e-Learning*, page 23, 2007.

[6] Ngamnij Arch-int and Somjit Arch-int. Semantic ontology mapping for interoperability of learning resource systems using a rule-based reasoning approach. *Expert Systems with Applications*, pages 7428 – 7443, 2013.

[7] Ehsaneddin Asgari and Jean-Cédric Chappelier. Linguistic resources and topic models for the analysis of persian poems. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 23–31, 2013.

[8] Ehsaneddin Asgari, Marzyeh Ghassemi, and Mark Alan Finlayson. Confirming the themes and interpretive unity of ghazal poetry using topic models. In *NIPS Workshop for Topic Models*, 2013.

[9] Domenico Beneventano, Claudio Gennaro, Sonia Bergamaschi, and Fausto Rabitti. A mediator-based approach for integrating heterogeneous multimedia sources. *Multimedia Tools and Applications*, 62(2):427–450, October 2011.

[10] Thomas Beyhl and Holger Giese. *Efficient and scalable graph view maintenance for deductive graph databases based on generalized discrimination networks*, volume 99. Universitätsverlag Potsdam, 2016.

[11] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[12] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[13] Patrícia Cavoto, Victor Cardoso, Régine Vignes Lebbe, and André Santanchè. Fish-Graph: A Network-Driven Data Analysis. In *11th IEEE Int. Conf. on eScience*, Germany, 2015.

[14] Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. Resources sequencing using automatic prerequisite–outcome annotation. *ACM Trans. Intell. Syst. Technol.*, 6(1):pages 6:1–6:30, March 2015.

[15] Beulah Christalin Latha Christudas, E. Kirubakaran, and P. Ranjit Jeba Thangaiah. An evolutionary approach for personalization of content delivery in e-learning systems based on learner behavior forcing compatibility of learning materials. *Telematics and Informatics*, 35(3):520 – 533, 2018. SI: EduWebofData.

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

[17] Ricardo Conejo, Eduardo Guzmán, Jose-Luis Perez de-la Cruz, and Beatriz Barros. An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, 41(2):594 – 606, 2014.

[18] J. Daltio and C. B. Medeiros. Handling multiple foci in graph databases. In Springer International Publishing Switzerland, editor, *LNBI - Proc. of 10th Int. Conf. on Data Integration in the Life Sciences*, volume 8574, pages 58–65, Lisboa, Portugal, 2014.

[19] Jonice Oliveira Diogo Nolasco. Detecting knowledge innovation through automatic topic labeling on scholar data. volume 00, pages 358–367, Los Alamitos, CA, USA, 2016. IEEE Computer Society.

[20] Anna Lea Dyckhoff, Dennis Zielke, Mohamed Amine Chatti, and Ulrik Schroeder. eLAT: An Exploratory Learning Analytics Tool for Reflection and Iterative Improvement of Technology Enhanced Learning. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 355–356. TU/e printservice, 2011.

[21] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29(2):8:1–8:34, April 2011.

[22] Fathi Essalmi, Leila Jemni Ben Ayed, Mohamed Jemni, Sabine Graf, and Kinshuk. Generalized metrics for the analysis of e-learning personalization strategies. *Computers in Human Behavior*, 48:310 – 322, 2015.

[23] Linton Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, 40(35):35–41, 1977.

[24] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

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

[26] Antonio Garrido, Lluvia Morales, and Ivan Serina. On the use of case-based planning for e-learning personalization. *Expert Syst. Appl.*, 60(C):1–15, October 2016.

[27] Ahmed Gater, Daniela Grigori, and Mokrane Bouzeghoub. A graph-based approach for semantic process model discovery. *Graph Data Management*, pages 438–462, 2011.

[28] Elena Gaudioso, Miguel Montero, and Felix Hernandez-Del-Olmo. Supporting teachers in adaptive educational systems through predictive models: A proof of concept. *Expert Syst. Appl.*, 39(1):621–625, January 2012.

[29] Sadaf Hina and P. D. D. Dominic. Gauging the school-based acceptability of web 2.0 collaborative tools. *Int. J. Bus. Inf. Syst.*, 21(3):321–341, February 2016.

[30] N. N. Meguru Ito and A. Imazawa. Differences in between technical colleges and institute of technology in japan, from the viewpoint of their educational objectives. In *2013 IEEE 5th Conference on Engineering Education (ICEED)*, pages 93–98, Dec 2013.

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

[32] Frédéric Jouault and Jean Bézivin. Km3: a dsl for metamodel specification. In *International Conference on Formal Methods for Open Object-Based Distributed Systems*, pages 171–185. Springer, 2006.

[33] Arijit Khan, Yinghui Wu, and Xifeng Yan. Emerging graph queries in linked data. In *ICDE*, pages 1218–1221. IEEE, 2012.

[34] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1536–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[35] Learning Technology Standards Committee of the IEEE. Draft standard for learning technology - learning object metadata. Technical report, IEEE Standards Department, New York, July 2002.

[36] Chee Kian Leong, Yew Haur Lee, and Wai Keong Mak. Mining sentiments in sms texts for teaching evaluation. *Expert Syst. Appl.*, 39:2584–2589, 2012.

[37] R. Likert. *A Technique for the Measurement of Attitudes.* Number N° 136-165 in A Technique for the Measurement of Attitudes. publisher not identified, 1932.

[38] Robert V. Lindsey, William P. Headden, III, and Michael J. Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 214–222, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[39] Suzanne Little, Rebecca Ferguson, and Stefan Rüger. Finding and reusing learning materials with multimedia similarity search and social networks. *Technology, Pedagogy and Education*, 21(2):pages 255–271, 2012.

[40] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[41] J. M. Luna, C. Castro, and C. Romero. Mdm tool: A data mining framework integrated into moodle. *Comput. Appl. Eng. Educ.*, 25(1):90–102, January 2017.

[42] Renata Machova, Jitka Komarkova, and Martin Lnenicka. Processing of big educational data in the cloud using apache hadoop. *Technical Co-Sponsored by IEEE UK/RI Computer Chapter*, page 46, 2016.

[43] Rodriguez Marko and Neubauer Peter. Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology*, 36(6):35–41, 2010.

[44] Osvaldo Matos-Junior, Nivio Ziviani, Fabiano C. Botelho, Marco Cristo, Anísio Lacerda, and Altigran Soares da Silva. Using taxonomies for product recommendation. *JIDM*, 3(2):pages 85–100, 2012.

[45] Surjeet Mishra, Amarendra Gorai, Tavleen Oberoi, and Hiranmay Ghosh. Efficient Visualization of Content and Contextual Information of an Online Multimedia Digital Library for Effective Browsing. *WI-IAT2010*, pages 257–260, August 2010.

[46] Matheus Silva Mota and Claudia Bauzer Medeiros. Introducing shadows: Flexible document representation and annotation on the web. *ICDE Workshops*, pages 13–18, 2013.

[47] L. Nadia. Design and implementation of information retrieval system based ontology. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, pages 500–505, April 2014.

[48] Sirinya On-at, C. Marie-Françoise Canut, André Péninou, and Florence Sèdes. Deriving user's profile from sparse egocentric networks: Using snowball sampling and link prediction. In *ICDIM*, pages 80–85. IEEE, 2014.

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

[50] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

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

[52] Ivens Portugal, Paulo Alencar, and Donald Cowan. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97:205–227, 2018.

[53] Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In *European conference on information retrieval*, pages 522–530. Springer, 2008.

[54] R. K. Raj, J. J. Ekstrom, J. Impagliazzo, S. Lingafelt, A. Parrish, H. Reif, and E. Sobiesk. Perspectives on the future of cybersecurity education. In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–2, Oct 2017.

[55] Ivan Luiz Marques Ricarte and Geraldo Ramos Falci Junior. A methodology for mining data from computer-supported learning environments. *Informática na educação: teoria & prática*, 14(2), 2011.

[56] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases.* O'Reilly Media, Inc., 2013.

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

[58] Rafael Geraldeli Rossi, Solange Oliveira Rezende, and Alneu Andrade Lopes. Term network approach for transductive classification. volume 9042, pages 497–515. Springer International Publishing, 2015.

[59] Sameendra Samarawickrama, Shanika Karunasekera, and Aaron Harwood. Finding high-level topics and tweet labeling using topic models. In *ICPADS*, pages 242–249, 2015.

[60] André Santanchè, João Sávio C. Longo, Geneviève Jomier, Michel Zam, and Claudia Bauzer Medeiros. Multi-focus research and geospatial data - anthropocentric concerns. *JIDM*, 5(2):pages 146–160, 2014.

[61] M. C. Saraiva and C. B. Medeiros. 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, 2016.

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

[63] M. C. Saraiva and C. B. Medeiros. Correlating educational documents from different sources through graphs and taxonomies. In *Brazilian Symposium on Databases, SBBD 2018, Rio de Janeiro, RJ, Brazil*, 2018.

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

[65] Yassine Zaoui Seghroucheni, Mohammed AL Achhab, et al. An approach to create multiple versions of the same learning object. *International Journal of Emerging Technologies in Learning (iJET)*, 9(5):17–21, 2014.

[66] Rodrigo Dias Arruda Senra and Claudia Bauzer Medeiros. Evaluate, Reorganize and Share: An Approach to Dynamically Organize Digital Hierarchies. *Journal on Data Semantics*, pages 225–236, 2014.

[67] Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes. *IEEE Trans. Emerging Topics Comput.*, 3(2):pages 205–219, 2015.

[68] La Matos Da Silva and André Santanchè. ARARA: Autoria de Objetos Digitais Complexos Baseada em Documentos. *Simpósio Brasileiro de Informática na Educação*, (2009):10, 2009.

[69] Setsuo Tsuruta, Rainer Knauf, Shinichi Dohi, Takashi Kawabe, and Yoshitaka Sakurai. *An Intelligent System for Modeling and Supporting Academic Educational Processes*, pages 469–496. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[70] César Vialardi, Jorge Chue, Juan Pablo Peche, Gustavo Alvarado, Bruno Vinatea, Jhonny Estrella, and Álvaro Ortigosa. A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction*, 21(1):217–248, 2011.

[71] Radovan Vrana. Open science, open access and open educational resources: Challenges and opportunities. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, pages 886–890. IEEE, 2015.

[72] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 437–445, New York, NY, USA, 2013. ACM.

[73] Yu Zhou, Ruifeng Xu, and Lin Gui. A sequence level latent topic modeling method for sentiment analysis via CNN based diversified restrict boltzmann machine. In *International Conference on Machine Learning and Cybernetics, ICMLC 2016, Jeju Island, South Korea, July 10-13, 2016*, pages 356–361, 2016.

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

Appendix A

Questionnaire of Case Study: Standardizing validation

Encontrando temas relacionados aos slides utilizados em aulas

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

CIMAL: Courseware Integration under Multiple relations to Assist Learning

Pesquisador: Márcio de Carvalho Saraiva
Orientadora: Claudia Bauzer Medeiros

Você está sendo convidado a participar como voluntário de uma pesquisa. Este documento, chamado Termo de Consentimento Livre e Esclarecido, visa assegurar seus direitos como participante. Por favor, leia com atenção e calma, aproveitando para esclarecer suas dúvidas. Se houver perguntas antes ou mesmo depois de assiná-lo, você poderá esclarecê-las com o pesquisador. Se preferir, pode levar este Termo para casa e consultar seus familiares ou outras pessoas antes de decidir participar. Se você não quiser participar ou retirar sua autorização, a qualquer momento, não haverá nenhum tipo de penalização ou prejuízo.

Justificativa e objetivos:
Professores e alunos precisam ter acesso a diversos materiais educacionais para entender um novo assunto ou atualizar seus conhecimentos. No entanto, o aumento da quantidade de material educativo disponível na internet faz com que essa tarefa seja bastante laboriosa e exija um grande dispêndio de tempo. Nosso projeto visa criar e desenvolver um conjunto de ferramentas computacionais para ajudar professores e os alunos a navegar através de coleções de material didático. A ideia é ajudar a lidar com o conteúdo de cursos, enfatizando as relações entre seus assuntos. O objetivo deste trabalho é permitir a integração de materiais educativos usando relações entre o conteúdo para auxiliar no processo de aprendizagem e facilitar a manipulação de materiais que estão relacionados entre si. Este projeto irá concentrar-se em materiais no formato de slides e vídeos. Qualquer outro formato de material (por exemplo, livros e áudios) não serão considerados por enquanto, mas a solução proposta no nosso trabalho pode ser estendida para outros tipos de material. Nesta pesquisa, vamos projetar e construir uma infra-estrutura de software que irá implementar um conjunto de ferramentas; chamamos essa infra-estrutura de CIMAL (em inglês: Courseware Integration under Multiple relations to Assist Learning). Para verificarmos que nossa meta foi atingida, faremos a avaliação do CIMAL com os materiais didáticos reais por meio de questionários e experimentos; além da análise de precisão da solução.
Os três principais contribuições esperadas da presente proposta são assim: (1) novos algoritmos para analisar material didático utilizando grafos; (2) novos métodos para interligar materiais didáticos de formatos diferentes (vídeos e slides) de diversas fontes, destacando-se as relações entre os conteúdos; (3) Construção da infra-estrutura CIMAL, através da qual professores e estudantes de diversas áreas serão capazes de navegar
através de diversos materiais didáticos, usando as relações que emergem progressivamente entre os assuntos, para orientar seus estudos.

Procedimentos:
Participando do estudo você está sendo convidado a:
● Avaliar o nível do relacionamento de temas sugeridos pelo sistema implementado nessa pesquisa à slides utilizados em sala de aula de cursos na área da Computação.
● Toda a atividade (experimento e entrevista) não deverá exigir mais do que 60 minutos.

Desconfortos e riscos:
Você não deve participar deste estudo se não estiver de acordo com os termos e procedimentos deste estudo. Esse estudo não traz riscos previsíveis aos voluntários, pois durante os testes só utilizarão um dispositivo com os qual já estão familiarizados, isto é, computadores, em atividades de estudo.
Em caso de dano decorrente da pesquisa, está garantida a assistência integral e imediata, de forma gratuita, pelo tempo que for necessário. Os voluntários também tem direito a indenização em caso de danos.

Benefícios:
É previsto que os voluntários realizem atividades de estudo mais dinâmicas que as convencionais, além de facilitar a busca, em grandes repositórios de material didático, do item mais adequado para aprendizagem de algum conceito novo, tornando seu aproveitamento em disciplinas potencialmente mais interessante.

Acompanhamento e assistência:
Durante a realização de atividades da pesquisa os pesquisadores estarão disponíveis, por meio de e-mail ou telefone, para ajudar e responder a quaisquer dúvidas dos voluntários em relação às atividades ou à tecnologia utilizada.

Sigilo e privacidade:
Você tem a garantia de que sua identidade será mantida em sigilo e nenhuma informação será dada a outras pessoas que não façam parte da equipe de pesquisadores. Na divulgação dos resultados desse estudo, seu nome não será citado. Todo o material coletado na pesquisa, que inclui questionários, é destinado para ajudar a projetar e avaliar a ferramenta produzida nessa pesquisa, para ser usada em diferentes atividades e por pessoas com diferentes habilidades. Dados e materiais obtidos dos sujeitos em interação com a tecnologia serão tornados anônimos. Os dados coletados serão armazenados por um período de cinco anos a partir do encerramento da pesquisa em um computador ao qual somente os pesquisadores têm acesso. Os dados coletados não serão utilizados em outros projetos, servindo exclusivamente para a pesquisa em questão.

Ressarcimento:
Este estudo não prevê nenhum tipo de ressarcimento, reembolso ou premiação financeira ou de qualquer outra natureza. O experimento será via internet, no horário conveniente aos voluntários, assim estes não terão custos adicionais para participar da pesquisa.

Contato:
Em caso de dúvidas sobre o estudo, você poderá entrar em contato com os pesquisadores:

Márcio de Carvalho Saraiva
Telefone: (19) 9 98100-4438
E-mail: marcio.saraiva@ic.unicamp.br

Claudia Bauzer Medeiros
Telefone: (19) 3521-5855
E-mail: cmbm@ic.unicamp.br
No Instituto de Computação da UNICAMP, Departamento Sistemas de Informação (Av. Albert Einstein 1251 - Cidade Universitária - Campinas/SP - Brasil CEP 13083-852)

Em caso de denúncias ou reclamações sobre sua participação e sobre questões éticas do estudo, você pode entrar em contato com a secretaria do Comitê de Ética em Pesquisa (CEP) da UNICAMP das 08:30hs às 13:30hs e das 13:00hs as 17:00hs na Rua: Tessália Vieira de Camargo, 126; CEP 13083-887 Campinas – SP; telefone (19) 3521-8936; fax (19) 3521-7187; email: cep@fcm.unicamp.br

Responsabilidade do Pesquisador:
Asseguro ter cumprido as exigências da resolução 466/2012 CNS/MS e complementares na elaboração do protocolo e na obtenção deste Termo de Consentimento Livre e Esclarecido. Asseguro, também, ter explicado e fornecido uma via deste documento ao participante. Informo que o estudo foi aprovado pelo CEP perante o qual o projeto foi apresentado. Comprometo-me a utilizar o material e os dados obtidos nesta pesquisa exclusivamente para as finalidades previstas neste documento ou conforme o consentimento dado pelo participante.

Consentimento livre e esclarecido:
Após ter recebido esclarecimentos sobre a natureza da pesquisa, seus objetivos, métodos, benefícios previstos, potenciais riscos e o incômodo que esta possa acarretar, aceito participar.

LInk para download do TCLE: https://goo.gl/G6Zzp2

*Obrigatório

1. **Concordo e aceito participar desta pesquisa** *
Marcar apenas uma oval.

○ Sim

○ Não *Pare de preencher este formulário.*

Algumas informações

POR FAVOR LEIA O TEXTO A SEGUIR:

Primeiramente, muito obrigado por ajudar em nossa pesquisa. Durante o doutorado, desenvolvi um sistema que indica quais são os temas mais relacionados a um conjunto de slide utilizado em aulas. Agora preciso da ajuda de pessoas da área da Computação para avaliar se os temas indicados pelo nosso sistema são relevantes, uma vez que os materiais educativos utilizados na pesquisa são todos dessa área.

Antes de avaliar a saída do nosso sistema, gostaríamos de saber seu nível de inglês (os slides estão inglês), escolaridade, área do curso, última universidade que passou (esta pesquisa esta sendo realizada em várias universidades) e seu e-mail (para possível contato). Sua identidade será mantida em sigilo, e na divulgação dos resultados deste estudo seu nome não será citado. Estamos interessados somente em dados relativos a avaliação da saída do nosso sistema. Portanto, todos os dados pessoais, que porventura forem registrados, serão completamente descartados ao final do experimento, assim como as respostas individuais do questionário, depois de analisado. Nenhum dado pessoal sobre os voluntários será mantido em qualquer formato ou meio.

2. **Área de estudo** *

3. **Nível de leitura em Inglês** *
Marcar apenas uma oval.

○ Básico
○ Intermediário
○ Avançado

4. **Escolaridade** *
Marcar apenas uma oval.

○ Graduação em andamento
○ Graduação finalizada
○ Mestrado em andamento
○ Mestrado finalizado
○ Doutorado em andamento
○ Doutorado finalizado
○ Outro:

5. **Atual Instituição de ensino ou instituição que obteve o último título** *

6. **E-mail** *

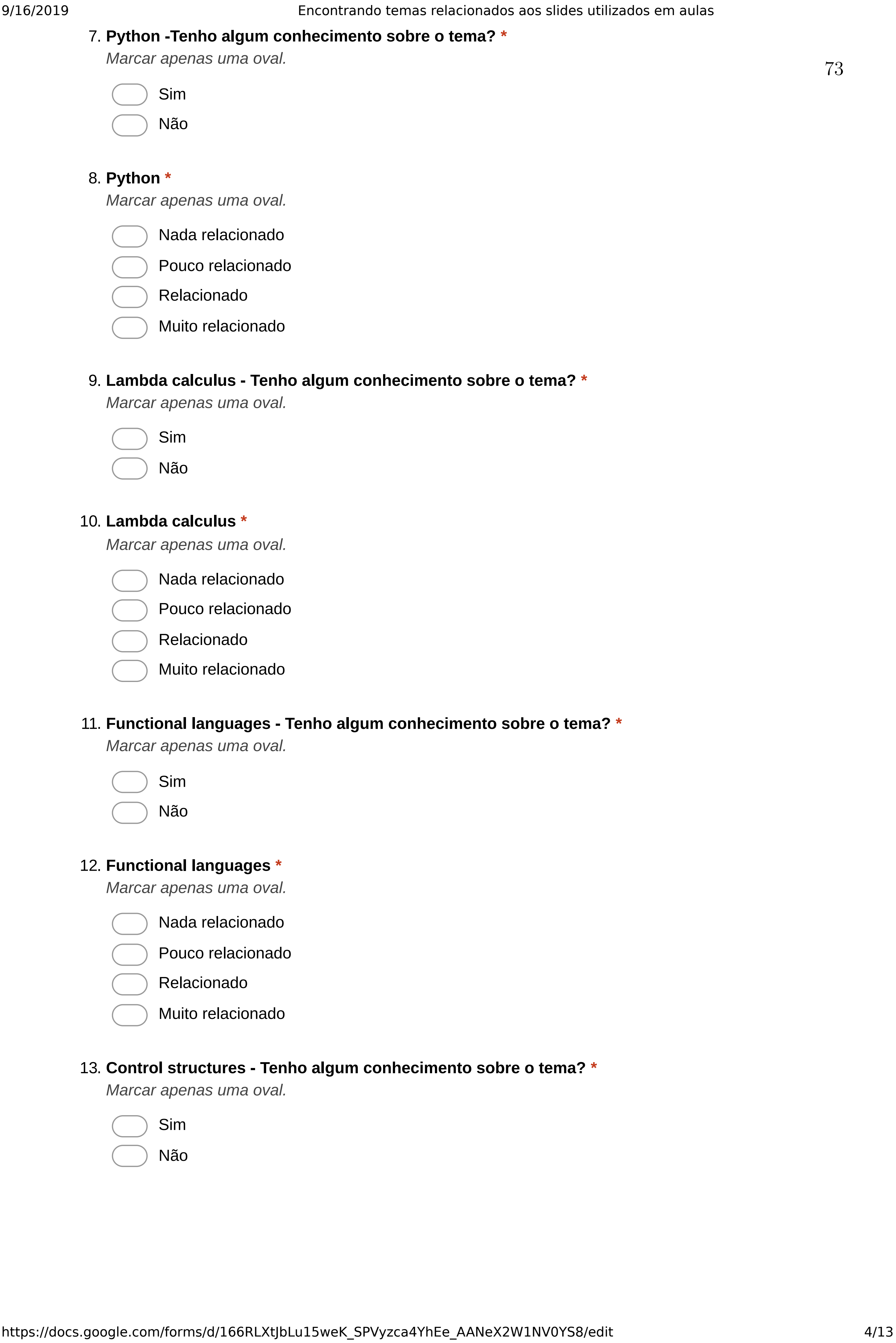
POR FAVOR LEIA O TEXTO A SEGUIR:

Clique no link abaixo do nome do conjunto de slides, analise os slides e escolha a alternativa que melhor representa se o tema tem relação ou não com os slides. Além disso, você pode sugerir um tema que você acredita que seja relacionado ao conjunto de slides, mas que não foi indicado pelo sistema.

Desde já muito obrigado por toda ajuda.

Conjunto de slides 4 - Functions

Link: https://goo.gl/iagU8f



7. **Python -Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

8. **Python** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

9. **Lambda calculus - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

10. **Lambda calculus** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

11. **Functional languages - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

12. **Functional languages** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

13. **Control structures - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

14. **Control structures** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

15. **Pattern matching - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

16. **Pattern matching** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

17. **Sugestão de tema**

____

Conjunto de slides 7 - Files

Link: https://goo.gl/9rdcBJ

18. **Control structures - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

19. **Control structures** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

20. **Text editing - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

21. **Text editing** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

22. **Python - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

23. **Python** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

24. **Structured Query Language - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

25. **Structured Query Language** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

26. **Linked lists - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

27. **Linked lists** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

28. **Sugestão de tema**

____

Conjunto de slides 8 - Lists

Link: https://goo.gl/6eR9cU

29. **Python - Tenho algum conhecimento sobre o tema? ***

Marcar apenas uma oval.

○ Sim
○ Não

30. **Python ***

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

31. **Linked lists ***

Marcar apenas uma oval.

○ Sim
○ Não

32. **Linked lists ***

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

33. **Control structures - Tenho algum conhecimento sobre o tema? ***

Marcar apenas uma oval.

○ Sim
○ Não

34. **Control structures ***

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

35. **Online social networks - Tenho algum conhecimento sobre o tema? ***

Marcar apenas uma oval.

○ Sim
○ Não

36. **Online social networks** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

37. **Social networking sites - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

38. **Social networking sites** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

39. **Sugestão de tema**

Conjunto de slides 9 - Dictionaries

Link: https://goo.gl/8zRv39

40. **Python - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

41. **Python** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

42. **Dictionaries - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

43. **Dictionaries** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

44. **Control structures - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

45. **Control structures** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

46. **B-trees - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

47. **B-trees** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

48. **Linked lists - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

49. **Linked lists** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

50. **Sugestão de tema**

Conjunto de slides 10 - Tuples

Link: https://goo.gl/aMV9xZ

51. **Relational database model - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

52. **Relational database model** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

53. **Python - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

54. **Python** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

55. **Structured Query Language - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

56. **Structured Query Language** *

Marcar apenas uma oval.

○ Nada relacionado

○ Pouco relacionado

○ Relacionado

○ Muito relacionado

57. **Linked lists - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim

○ Não

58. **Linked lists** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

59. **Interval arithmetic - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

60. **Interval arithmetic** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

61. **Sugestão de tema**

Conjunto de slides 16 - Retrieving and Visualizing Data

Link: https://goo.gl/adibqL

62. **Web crawling - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

63. **Web crawling** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

64. **Web search engines - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

65. **Web search engines** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

66. **Search engine indexing - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

67. **Search engine indexing** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

68. **Deep web - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

69. **Deep web** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

70. **Python - Tenho algum conhecimento sobre o tema?** *

Marcar apenas uma oval.

○ Sim
○ Não

71. **Python** *

Marcar apenas uma oval.

○ Nada relacionado
○ Pouco relacionado
○ Relacionado
○ Muito relacionado

72. **Sugestão de tema**

Appendix B

Questionnaire of Case Study: Proposed new multidisciplinary activities in an educational institution

Avaliação do resultado da análise de semelhanças entre disciplinas

Caros colegas.

Gostaria de contar com sua ajuda para avaliar o resultado do método que utilizei para verificar semelhanças e possíveis colaborações entre as disciplinas.

Por favor, respondam este pequeno formulário que irá contribuir com minha pesquisa de doutorado.

Atenciosamente,

Márcio de Carvalho Saraiva
Professor de Informática

*Obrigatório

1. **Nome:** *
   Seu nome não será aparecerá em lugar algum, só será utilizado para organizar as respostas deste formulário. Após a pesquisa, essa informação será apagada.

   ____

2. **Professor de qual área?** *
   Marcar apenas uma oval.

   ○ Automação *Ir para a pergunta 3.*
   ○ Informática *Ir para a pergunta 8.*
   ○ Mecânica *Ir para a pergunta 13.*
   ○ Núcleo Comum *Ir para a pergunta 18.*

Para os professores de Automação

De acordo com os resultados apresentados nos links abaixo:

1 - https://goo.gl/nhEqlm
2 - https://goo.gl/F61iZt

Dê sua opinião sobre a seguinte afirmação:

3. **As disciplinas e as semelhanças apresentadas nos documentos podem ser utilizadas para planejar novas atividades entre as disciplinas.** *
   1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente
   Marcar apenas uma oval.

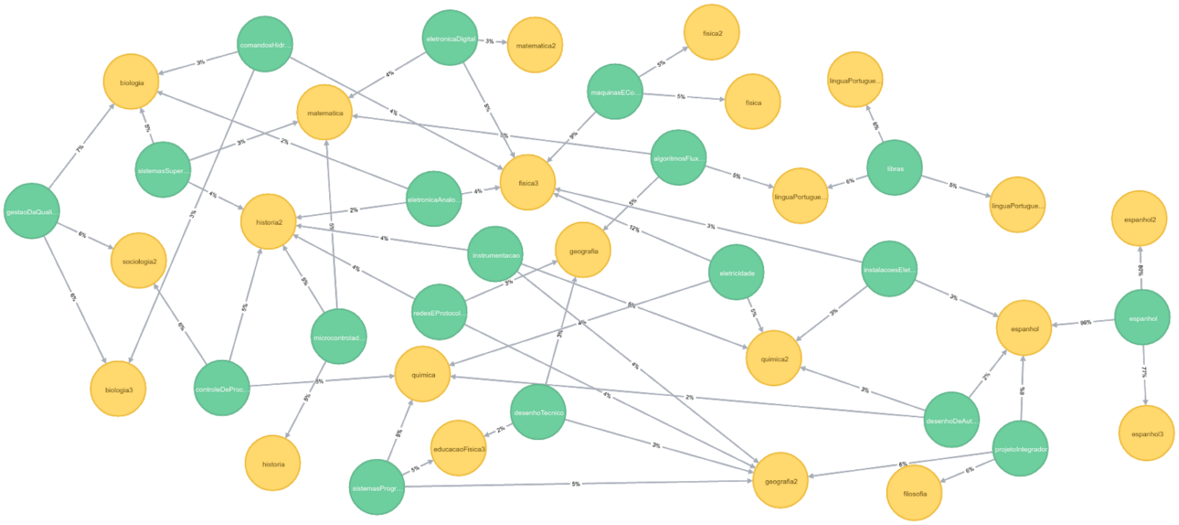
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

4. **Comentários**

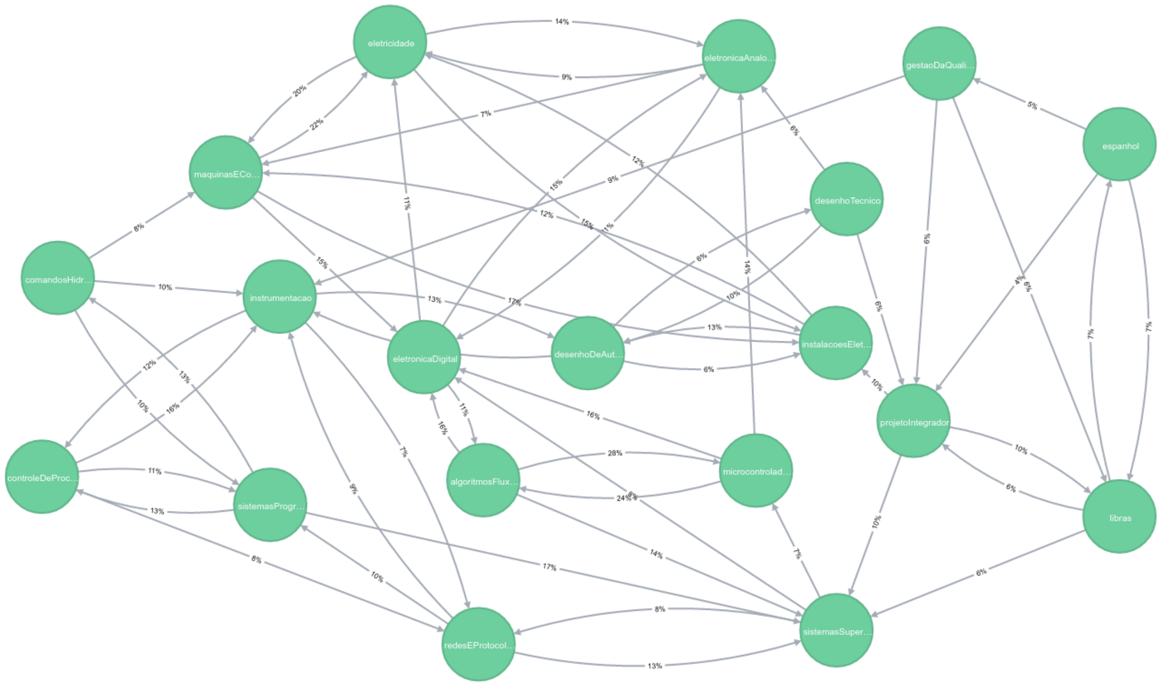
Sobre a utilização de grafos

Dado os seguintes grafos. Dê sua opinião sobre a seguinte afirmação

Semelhanças entre as disciplinas de Automação e do Núcleo Comum - Para melhor visualização: https://goo.gl/lCzsY8



Semelhanças entre as disciplinas de Automação - Para melhor visualização: https://goo.gl/Le0WfG



5. **O uso de grafos, onde as disciplinas são nós (círculos) e as semelhanças entre elas estão nas arestas (ligações), facilita a visualização das informações, análise e o planejamento?**
1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

6. **Qual a melhor forma de visualização das informações?**
Marcar apenas uma oval.

○ Documentos textuais
○ Grafos

7. **Justifique a resposta acima?**

Pare de preencher este formulário.

Para os professores de Informática

De acordo com os resultados apresentados nos links abaixo:

1- https://goo.gl/Rv3zmR
2- https://goo.gl/dkXTeo

Dê sua opinião sobre a seguinte afirmação:

8. **As disciplinas e as semelhanças apresentadas nos documentos podem ser utilizadas para planejar novas atividades entre as disciplinas.** *
1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente
Marcar apenas uma oval.

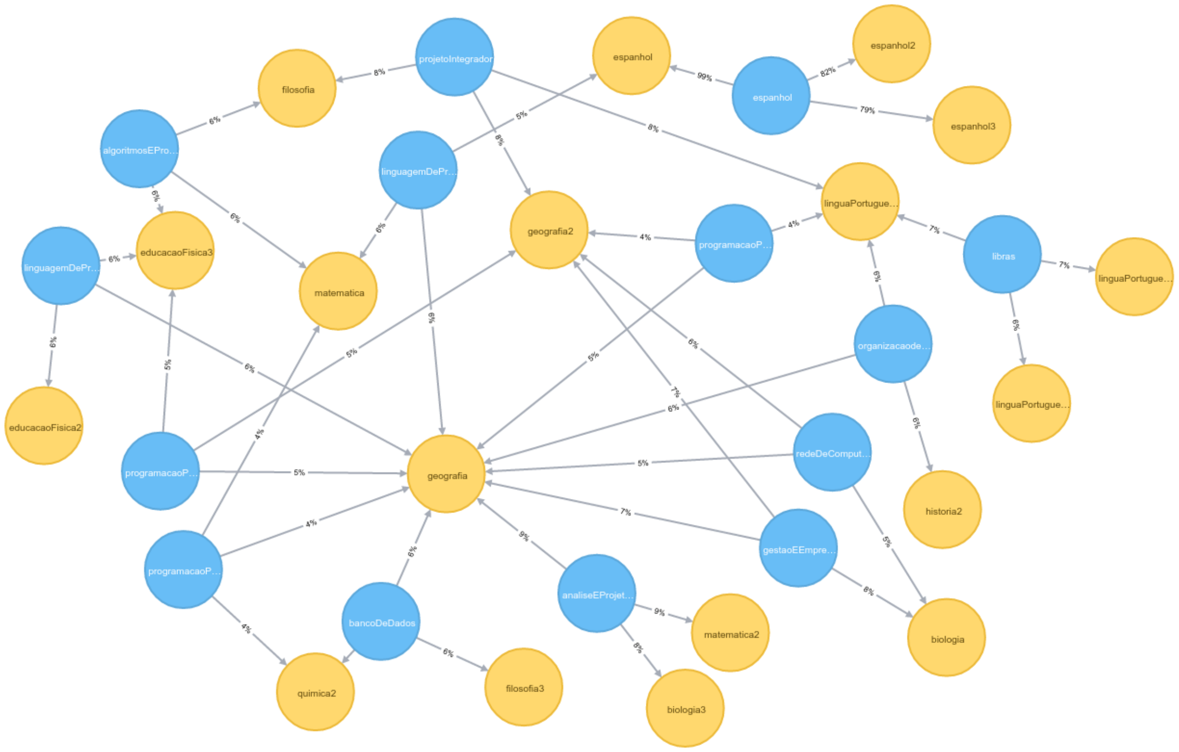
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

9. **Comentários**

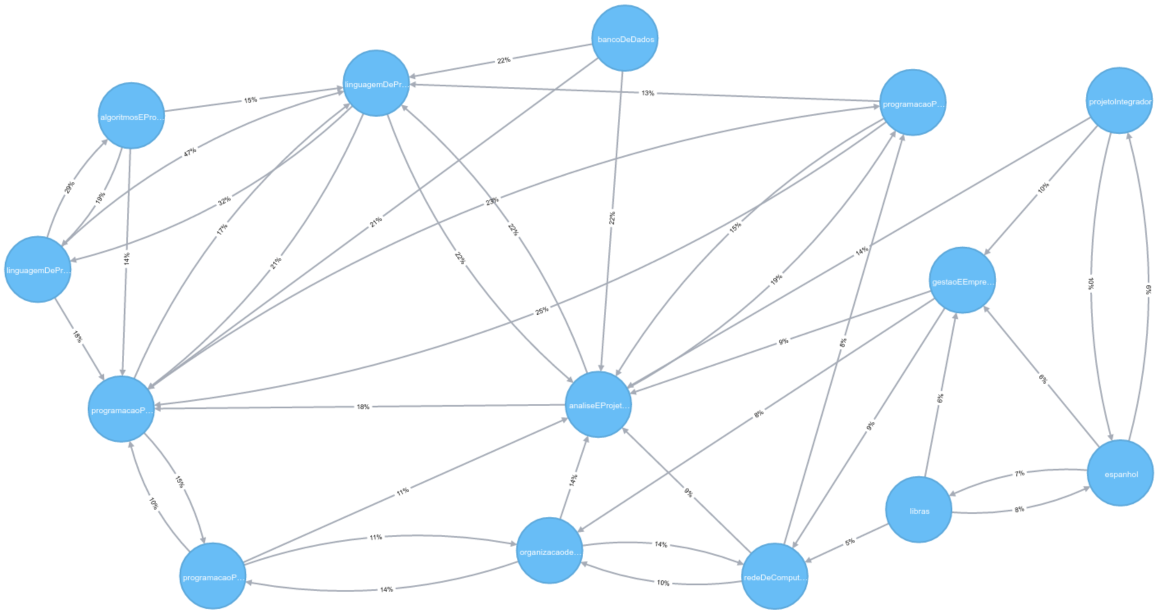
Sobre a utilização de grafos

Dado os seguintes grafos. Dê sua opinião sobre a seguinte afirmação

Semelhanças entre as disciplinas de Informática e do Núcleo Comum. - Para melhor visualização: https://goo.gl/g9rG7j



Semelhanças entre as disciplinas de Informática. - Para melhor visualização: - Para melhor visualização: https://goo.gl/hJCpcp



10. **O uso de grafos, onde as disciplinas são nós (círculos) e as semelhanças entre elas estão nas arestas (ligações), facilita a visualização das informações, análise e o planejamento?**
1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

11. **Qual a melhor forma de visualização das informações?**
Marcar apenas uma oval.

○ Documentos textuais
○ Grafos

12. **Justifique a resposta acima?**

Pare de preencher este formulário.

Para os professores de Mecânica

De acordo com os resultados apresentados nos links abaixo:

1- https://goo.gl/d29Eqe
2- https://goo.gl/VQDeQ0

Dê sua opinião sobre a seguinte afirmação:

13. **As disciplinas e as semelhanças apresentadas nos documentos podem ser utilizadas para planejar novas atividades entre as disciplinas.** *
1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente
Marcar apenas uma oval.

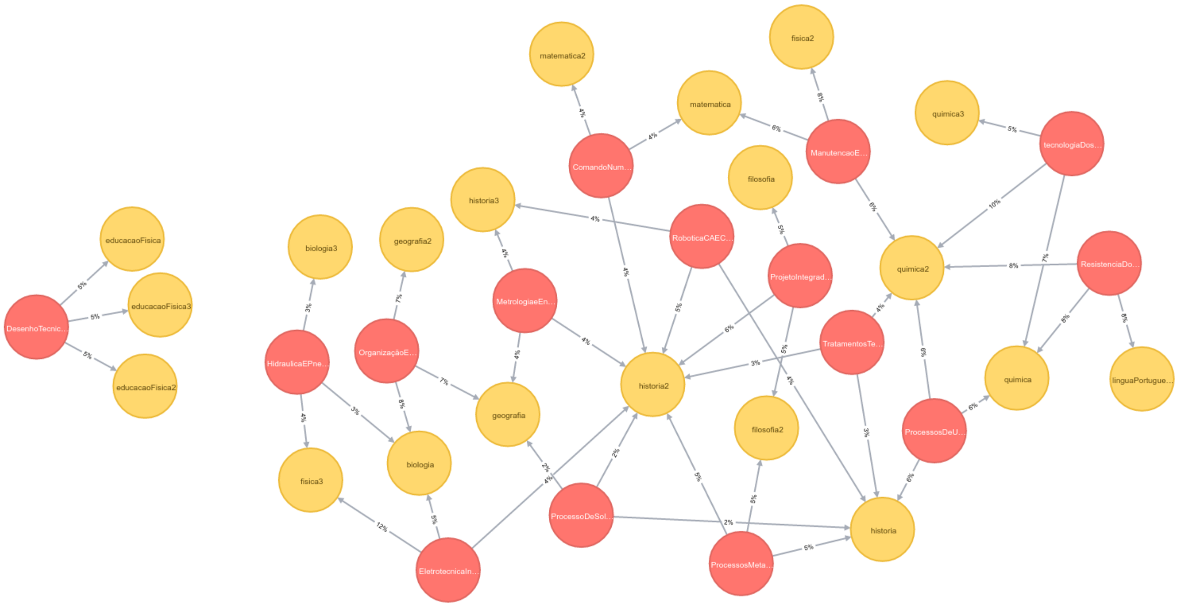
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

14. **Comentários**

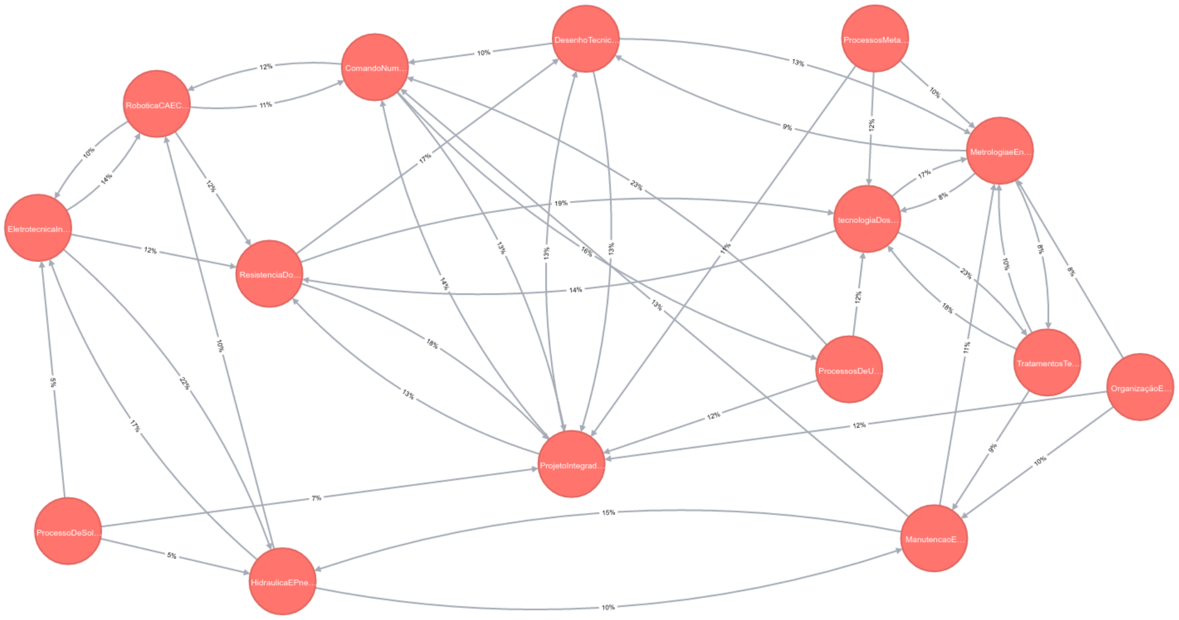
Sobre a utilização de grafos

Dado os seguintes grafos. Dê sua opinião sobre a seguinte afirmação

Semelhanças entre as disciplinas de Mecânica e do Núcleo Comum. - Para melhor visualização: https://goo.gl/AL303P



Semelhanças entre as disciplinas de Mecânica. - Para melhor visualização: https://goo.gl/9bRGpN



15. **O uso de grafos, onde as disciplinas são nós (círculos) e as semelhanças entre elas estão nas arestas (ligações), facilita a visualização das informações, análise e o planejamento?**

1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente

Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

16. **Qual a melhor forma de visualização das informações?**

Marcar apenas uma oval.

○ Documentos textuais

○ Grafos

17. **Justifique a resposta acima?**

Pare de preencher este formulário.

Para os professores do Núcleo

De acordo com os resultados apresentados nos links abaixo:

1- https://goo.gl/kKxKoH
2- https://goo.gl/dkXTeo
3- https://goo.gl/F61iZt

Dê sua opinião sobre a seguinte afirmação:

18. **As disciplinas e as semelhanças apresentadas nos documentos podem ser utilizadas para planejar novas atividades entre as disciplinas.** *

1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente

Marcar apenas uma oval.

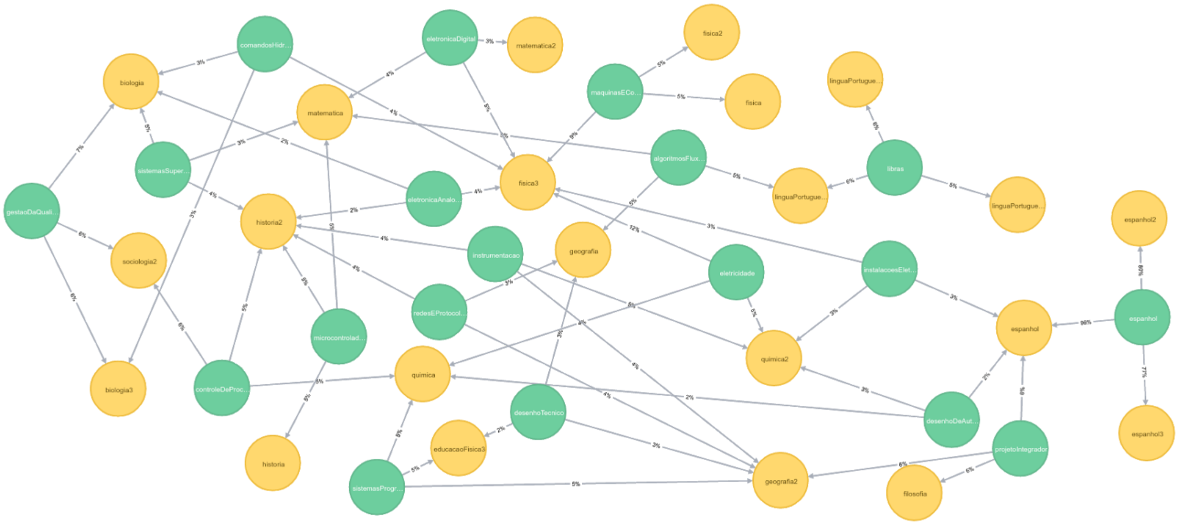
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

19. **Comentários**

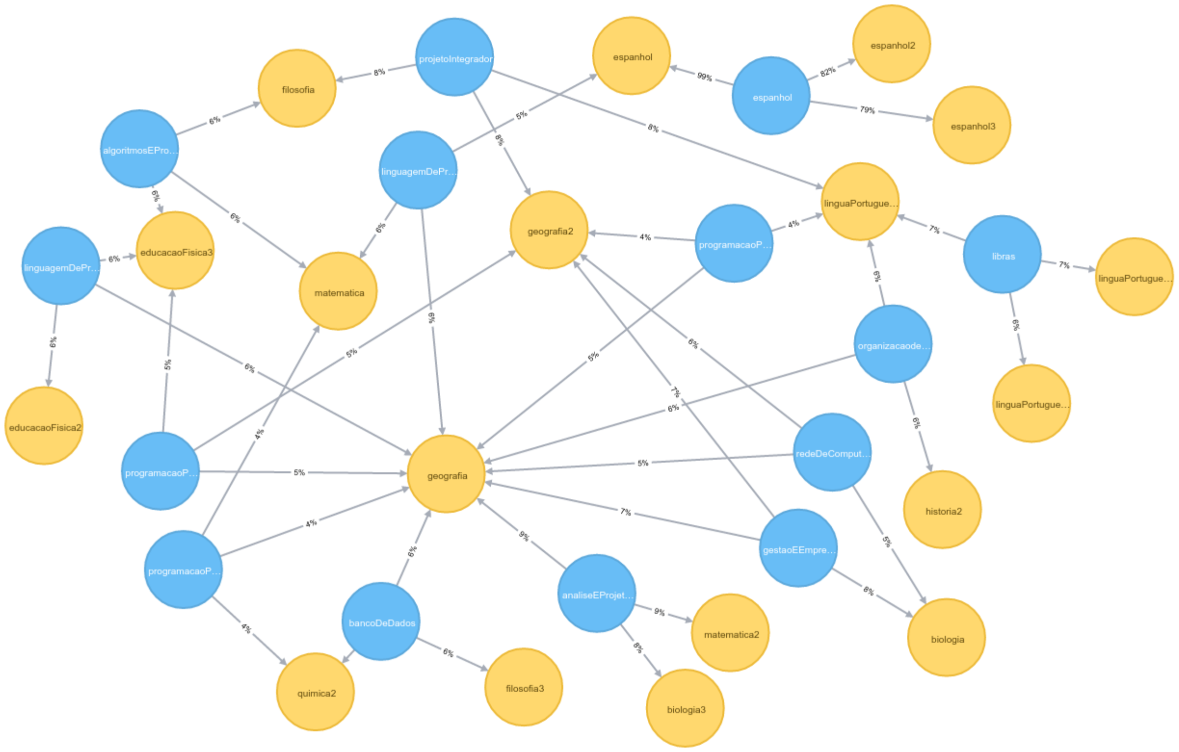
Sobre a utilização de grafos

Dado os seguintes grafos. Dê sua opinião sobre a seguinte afirmação

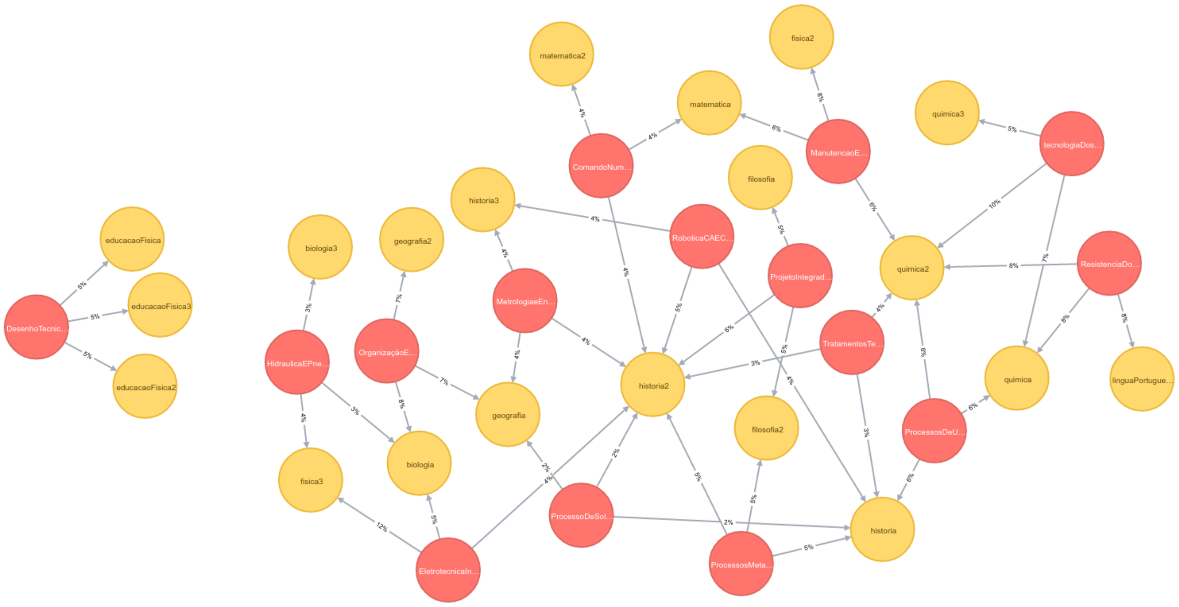
Semelhanças entre as disciplinas de Automação e do Núcleo Comum - Para melhor visualização: https://goo.gl/lCzsY8



Semelhanças entre as disciplinas de Informática e do Núcleo Comum. - Para melhor visualização: https://goo.gl/g9rG7j



Semelhanças entre as disciplinas de Mecânica e do Núcleo Comum. - Para melhor visualização: https://goo.gl/AL303P



20. **O uso de grafos, onde as disciplinas são nós (círculos) e as semelhanças entre elas estão nas arestas (ligações), facilita a visualização das informações, análise e o planejamento?**

1- Discordo totalmente; 2- Discordo parcialmente; 3- Indiferente ; 4- Concordo parcialmente; 5- Concordo totalmente

Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

21. **Qual a melhor forma de visualização das informações?**

Marcar apenas uma oval.

○ Documentos textuais

○ Grafos

22. **Justifique a resposta acima?**

____

____

____

____

____

Powered by
Google Forms

Appendix C

Approval of the UNICAMP Ethics Committee

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: CIMAL: Courseware Integration under Multiple relations to Assist Learning
Pesquisador: MARCIO DE CARVALHO SARAIVA
Área Temática:
Versão: 2
CAAE: 64404816.5.0000.5404
Instituição Proponente: Instituto de Computação
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.997.453

Apresentação do Projeto:

Professores e alunos precisam ter acesso a diversos materiais educacionais para entender um novo assunto ou atualizar seus conhecimentos. No entanto, o aumento da quantidade de material educativo disponível na internet faz com que essa tarefa seja bastante laboriosa e exija um grande dispêndio de tempo. Este projeto visa criar e desenvolver um conjunto de ferramentas computacionais para ajudar professores e os alunos a navegar através de coleções de material didático. Trabalhos relacionados, relativos à integração e visualização de conteúdo educativo, mostram que ainda existem muitos desafios nesse campo. Além disso, há ainda a necessidade para a detecção de diferenças entre o materiais didáticos produzidos por professores distintos ou ainda por um único professor, mas em diferentes pontos no tempo. Em sites como o Banco Internacional de Objetos Educacionais, ACM Learning Center and the ACM Techpack, Coursera, ARIADNE Foundation, MERLOT e SlideShare, pesquisadores criaram repositórios para material didático. No entanto, observou-se que até mesmo consultas simples para encontrar materiais nesses repositórios podem resultar em um grande número de itens, o que torna difícil entendê-los e selecionar os relevantes. O objetivo deste trabalho é permitir a integração de materiais educativos usando relações entre o conteúdo para auxiliar no processo de aprendizagem e facilitar a manipulação de materiais que estão relacionados entre si. Este projeto irá concentrar-se em materiais no formato de slides e vídeos. A solução proposta no nosso trabalho pode ser esteestendida para outros tipos de material. Nesta pesquisa, será projetada e construida uma infra-estrutura de software que irá implementar um conjunto de ferramentas; chamada de CIMAL (em inglês: Courseware Integration under Multiple relations to Assist Learning. O CIMAL será avaliado por professores e alunos do Instituto de Computação da Unicamp por meio de questionários e experimento. As três principais contribuições esperadas da presente proposta são assim: (1) novos algoritmos para analisar material didático utilizando grafos; (2) novos métodos para interligar materiais didáticos de formatos diferentes (vídeos e slides) de diversas fontes, destacando-se as relações entre os conteúdos; (3) Construção da infra-estrutura CIMAL, através da qual professores e estudantes de diversas áreas serão capazes de navegar através de diversos materiais didáticos, usando as relações que emergem progressivamente entre os assuntos, para orientar seus estudos.

Objetivo da Pesquisa:

Objetivo Primario:

O objetivo principal deste trabalho e ajudar os usuarios a encontrar conteudo didatico relevante em repositorios de materiais educativos. Para atingir esse objetivo, o projeto ira projetar e desenvolver algoritmos para extrair relacoes ocultas entre o conteudo desses materiais. O projeto concentra-se em material disponibilizado por educadores, isto e, slides e videos de aulas. Estas relacoes vao ajudar no processo de aprendizagem e facilitar a manipulacao de materiais que sao (indiretamente) relacionados uns aos outros.

Objetivo Secundario:

* Definicao de tipos de relacionamento que serao analisados entre os materiais educativos disponibilizados.
* Obtencao de diferentes tipos de relacionamentos entre os conteudos dos materiais educativos utilizados.
* Definicao de algoritmos para identificacao de conjuntos de topicos abordados em um mesmo material didatico, uma vez que em uma unica aula um professor pode abordar mais do que um topico.
* Adaptacao de medidas de similaridade para distinguir materiais educacionais.
* Permitir a analise dos materiais educativos utilizados nesta pesquisa, enfatizando a rede de relacionamentos criados entre os conteudos desses materiais.

Avaliação dos Riscos e Benefícios:

De acordo com o pesquisador, o projeto apresenta os seguintes riscos e benefícios:

Riscos:

Esse estudo não traz riscos previsíveis aos voluntários, pois durante os testes só utilizarão um dispositivo com os qual já estão familiarizados, isto é,computadores, em atividades de estudo.
Benefícios:
É previsto que os voluntários realizem atividades de estudo mais dinâmicas que as convencionais, além de facilitar a busca, em grandes repositórios de material didático, do item mais adequado para aprendizagem de algum conceito novo, tornando seu aproveitamento em disciplinas potencialmente mais interessante.

Comentários e Considerações sobre a Pesquisa:
Este protocolo se refere ao Projeto de Pesquisa de Doutorado intitulado " CIMAL: Courseware Integration under Multiple relations to Assist Learning" que será desenvolvido pelo pesquisador responsável Marcio De Carvalho Saraiva sob supervisão da Profa Dra Claudia Maria Bauzer Medeiros. A pesquisa foi enquadrada na Área de Ciências Exatas e da Terra e embasará a Tese de Doutorado do pesquisador. A Instituição Proponente é o Instituto de Computação da UNICAMP. Segundo as Informações Básicas do Projeto, a pesquisa tem orçamento estimado em R$ 1800,00 (Um mil e oitocentos reais) e o cronograma apresentado contempla início do estudo para outubro de 2016, com término em março de 2018, com a coleta de dados iniciando em agosto de 2017. Serão abordados ao todo 30 pessoas, sendo 5 professores e 25 alunos. O objetivo deste trabalho é permitir a integração de materiais educativos usando relações entre o conteúdo para auxiliar no processo de aprendizagem e facilitar a manipulação de materiais que estão relacionados entre si. Nesta pesquisa, será projetado e construído uma infra-estrutura de software que irá implementar um conjunto de ferramentas; chamamos essa infra-estrutura de CIMAL (em inglês: Courseware Integration under Multiple relations to Assist Learning). Para verificar se a meta foi atingida, o CIMAL será avaliado com os materiais didáticos por meio de questionários e testes de uso de software respondidos por professores e alunos do Instituto de Computação da Unicamp.

Considerações sobre os Termos de apresentação obrigatória:
Foram apresentados: 1) projeto de pesquisa (ProjetoDePesquisa.pdf); 2) folha de rosto, devidamente preenchida, datada e assinada pelo diretor da unidade na qual o pesquisador tem vínculo (FolhaDeRostoMarcio.pdf); 3) termo de consentimento livre e esclarecido (TCLE), de acordo com as normas da Res. CNS-MS 466/12 (TCLEMarcio.pdf); 4) COmprovante de vínculo do pesquisador na instituição (ComprovanteDeMatricula.pdf) . 5) Carta de resposta as pendências (cartaResposta.pdf)

Conclusões ou Pendências e Lista de Inadequações:

Todas as pendências foram respondidas adequadamente. O projeto encontra-se apto para o desenvolvimento em seres humanos

Considerações Finais a critério do CEP:

- O participante da pesquisa deve receber uma via do Termo de Consentimento Livre e Esclarecido, na íntegra, por ele assinado (quando aplicável).

- O participante da pesquisa tem a liberdade de recusar-se a participar ou de retirar seu consentimento em qualquer fase da pesquisa, sem penalização alguma e sem prejuízo ao seu cuidado (quando aplicável).

- O pesquisador deve desenvolver a pesquisa conforme delineada no protocolo aprovado. Se o pesquisador considerar a descontinuação do estudo, esta deve ser justificada e somente ser realizada após análise das razões da descontinuidade pelo CEP que o aprovou. O pesquisador deve aguardar o parecer do CEP quanto à descontinuação, exceto quando perceber risco ou dano não previsto ao participante ou quando constatar a superioridade de uma estratégia diagnóstica ou terapêutica oferecida a um dos grupos da pesquisa, isto é, somente em caso de necessidade de ação imediata com intuito de proteger os participantes.

- O CEP deve ser informado de todos os efeitos adversos ou fatos relevantes que alterem o curso normal do estudo. É papel do pesquisador assegurar medidas imediatas adequadas frente a evento adverso grave ocorrido (mesmo que tenha sido em outro centro) e enviar notificação ao CEP e à Agência Nacional de Vigilância Sanitária – ANVISA – junto com seu posicionamento.

- Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEP de forma clara e sucinta, identificando a parte do protocolo a ser modificada e suas justificativas e aguardando a aprovação do CEP para continuidade da pesquisa. Em caso de projetos do Grupo I ou II apresentados anteriormente à ANVISA, o pesquisador ou patrocinador deve enviá-las também à mesma, junto com o parecer aprovatório do CEP, para serem juntadas ao protocolo inicial.

- Relatórios parciais e final devem ser apresentados ao CEP, inicialmente seis meses após a data deste parecer de aprovação e ao término do estudo.

-Lembramos que segundo a Resolução 466/2012 , item XI.2 letra e, “cabe ao pesquisador apresentar dados solicitados pelo CEP ou pela CONEP a qualquer momento”.

-O pesquisador deve manter os dados da pesquisa em arquivo, físico ou digital, sob sua guarda e responsabilidade, por um período de 5 anos após o término da pesquisa.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_814054.pdf	21/03/2017 13:47:05		Aceito
Projeto Detalhado / Brochura Investigador	ProjetoDePesquisa.pdf	21/03/2017 13:46:23	MARCIO DE CARVALHO SARAIVA	Aceito
Outros	cartaResposta.pdf	21/03/2017 13:45:24	MARCIO DE CARVALHO SARAIVA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLEMarcio.pdf	21/03/2017 13:40:33	MARCIO DE CARVALHO SARAIVA	Aceito
Outros	ComprovanteDeMatricula.pdf	31/01/2017 15:48:48	MARCIO DE CARVALHO SARAIVA	Aceito
Folha de Rosto	FolhaDeRostoMarcio.pdf	18/11/2016 16:07:10	MARCIO DE CARVALHO SARAIVA	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

CAMPINAS, 03 de Abril de 2017

Assinado por:
Renata Maria dos Santos Celeghini
(Coordenador)