Finding out Topics in Educational Materials Using their Components

Márcio de Carvalho Saraiva
Institute of Computing
University of Campinas (UNICAMP)
Campinas,Brazil, CEP 13083-852
Email: marcio.saraiva@ic.unicamp.br

Claudia Bauzer Medeiros
Institute of Computing
University of Campinas (UNICAMP)
Campinas,Brazil, CEP 13083-852
Email: cmbm@ic.unicamp.br

***Abstract*—The Web is witnessing an exponential growth of distributed and heterogeneous educational material. This hampers distinguishing among contents of these materials, as well as their retrieval. While information retrieval and classification mechanisms concentrate on corpus analysis, annotation approaches either target specific formats or require that a document follows interoperable standards. Rather than target only textual characteristics, our strategy is mainly based on components of educational material. The header, body, footer and numbering of slides and progress bar are examples of components of slides and videos. Though our work is general purpose, it is being tested against slides and videos from Coursera, a web platform that provides universal access to online education material and courses from universities and organizations around the world.**

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 [1], [2], [3]. 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 [4]. 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 [5], [6]. 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 these for e-learning [7].

Topic Modeling are 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 [8]). 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 present 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.

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

2. Concepts and Related Work

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 [9] research in EDM is formed by intersection of the areas: Data mining and machine learning, Computer-based education and Learning analytics.

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.

2.3. Document Analysis

Currently, document analysis is concentrated in three main strategies to deal with large volumes of complex and heterogeneous documents [10]: 1) to convert the original document 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.

2.4. 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 [11], [12] .

Probabilistic topic models allow work such as [13], [14], 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.

2.5. 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 [15], 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. [5], [16]). However, some authors (e.g [17]) 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 [17], [18], [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 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. [20] 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.

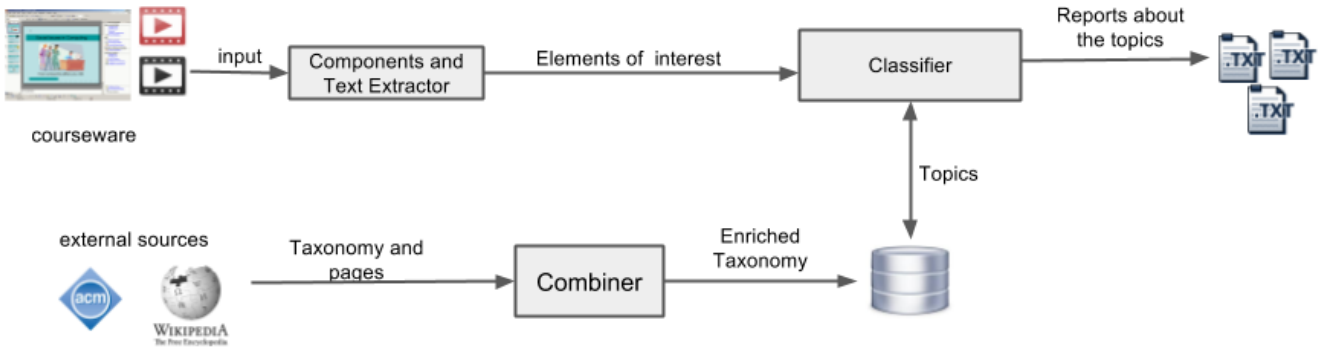

Figure 1. Overview of our research steps.

3. Research steps

Figure 1 shows the steps proposed for our methodology to find out topics in educational materials topics using their components. The first step "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 step, 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 step, the "Classifier". Here, we access a database which stores an enriched taxonomy created in the "Combiner" step. 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 [21], 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 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 step, 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.

4. 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 and 3 we can observe the components and texts, respectively highlighted through ellipses and rectangles that will be used for classification.

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 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

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

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

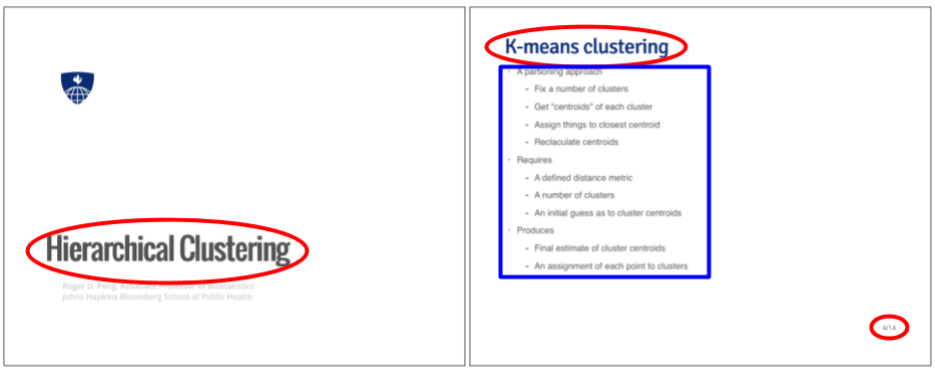

Figure 2. Components and text extracted from slides.

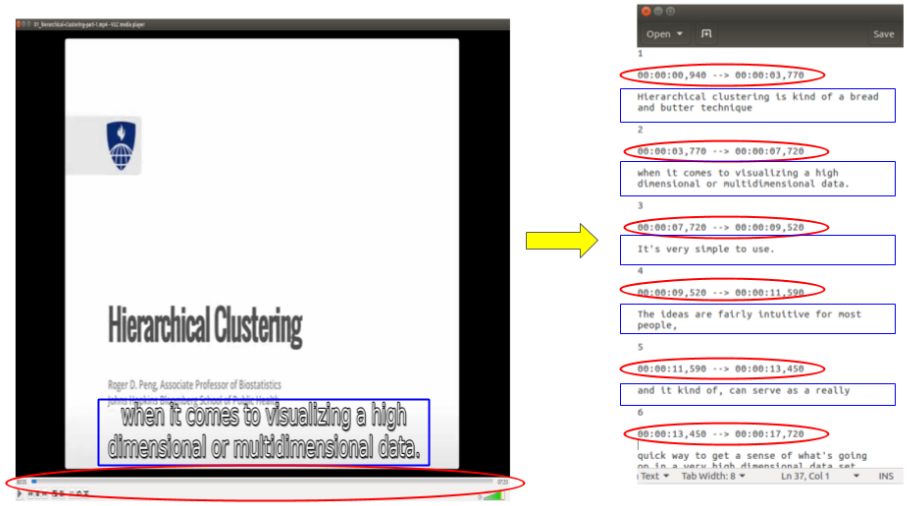

Figure 3. Components and text extracted from video subtitles.

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.

5. 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 five topics must 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). Figure 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.

6. 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. [22], 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

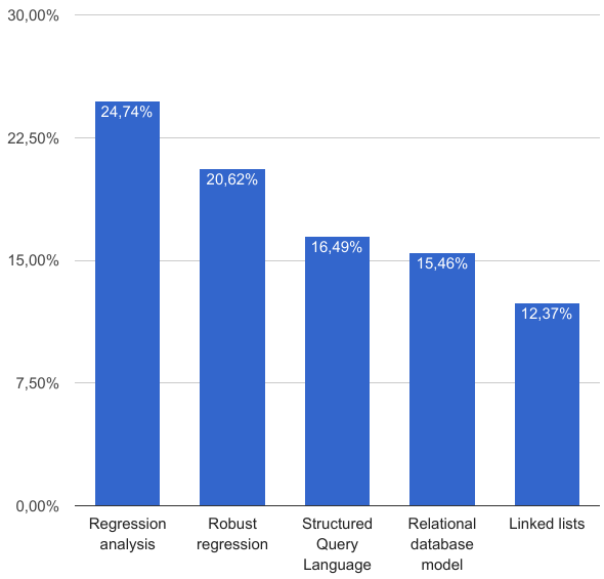

Figure 4. Top 5 topics covered in the Specialization course in Data Science at Coursera.

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.

7. 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 work, 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.

Acknowledgments

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

References

[1] R. 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*. IEEE, 2015, pp. 886–890.

[2] R. Machova, J. Komarkova, and M. Lnenicka, "Processing of big educational data in the clou d using apache hadoop," *Technical Co-Sponsored by IEEE UK/RI Computer Chapter*, p. 46, 2016.

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

[4] S. Changuel, N. Labroche, and B. Bouchon-Meunier, "Resources sequencing using automatic prerequisite–outcome annotation," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 1, pp. pages 6:1–6:30, Mar. 2015.

[5] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[6] Y. Zhuang, "Bag-of-discriminative-words (bodw) representation via topic modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 977–990, 2017.

[7] K. Sathiyamurthy, T. V. Geetha, and M. Senthilvelan, "An approach towards dynamic assembling of learning objects," in *ICACCI*. NY, USA: ACM, 2012, pp. 1193–1198.

[8] R. G. Rossi, S. O. Rezende, and A. A. Lopes, "Term network approach for transductive classification," vol. 9042. Springer International Publishing, 2015, pp. 497–515.

[9] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[10] M. S. Mota and C. B. Medeiros, "Introducing shadows: Flexible document representation and annotation on the web." *ICDE Workshops*, pp. 13–18, 2013.

[11] P. Ahmadi, M. Tabandeh, and I. Gholampour, "Persian text classification based on topic models," in *Electrical Engineering (ICEE), 2016 24th Iranian Conference on*. IEEE, 2016, pp. 86–91.

[12] Y. Zhou, R. Xu, and L. 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*, 2016, pp. 356–361.

[13] E. Asgari, M. Ghassemi, and M. A. Finlayson, "Confirming the themes and interpretive unity of ghazal poetry using topic models," in *Neural Information Processing Systems (NIPS) Workshop for Topic Models*, 2013.

[14] E. Asgari and J.-C. Chappelier, "Linguistic resources & topic models for the analysis of persian poems," in *Proceedings of the Second Workshop on Computational Linguistics for Literature (CLfL 2013)*, pp. 23–31.

[15] M. Allahyari and K. Kochut, "Automatic Topic Labeling using Ontology-based Topic Models," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 259–264.

[16] S. Samarawickrama, S. Karunasekera, and A. Harwood, "Finding high-level topics and tweet labeling using topic models," in *Parallel and Distributed Systems (ICPADS), 2015 IEEE 21st International Conference on*. IEEE, 2015, pp. 242–249.

[17] J. O. Diogo Nolasco, "Detecting knowledge innovation through automatic topic labeling on scholar data," vol. 00, no. undefined. Los Alamitos, CA, USA: IEEE Computer Society, 2016, pp. 358–367.

[18] R. V. Lindsey, W. P. Headden, III, and M. J. Stipicevic, "A phrase-discovering topic model using hierarchical pitman-yor processes," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 214–222.

[19] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. 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*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 437–445.

[20] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1536–1545.

[21] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *IJCAI*. CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.

[22] A. Khan, Y. Wu, and X. Yan, "Emerging graph queries in linked data," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012, pp. 1218–1221.