Enhancing Categorization of Learning Resources in the DAtaset of Joint Educational Entities

Carla Limongelli[1], Matteo Lombardi[2], Alessandro Marani[2], Davide Taibi[3]

[1]Engineering Department, Roma Tre University, Italy
limongel@ing.uniroma3.it

[2]School of Information and Communication Technology, Griffith University, Australia
{matteo.lombardi,alessandro.marani}@griffithuni.edu.au

[3]Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy
davide.taibi@itd.cnr.it

Abstract. The DAtaset of Joint Educational Entities (DAJEE) is a repository which hosts more than 20,000 educational resources crawled from the MOOC platform Coursera. The resources are divided per category according to the MOOC categorization on Coursera, which is, however, very shallow. This contribution focuses on a more meaningful categorization of the resources in DAJEE, tailored to their content. To achieve such goal, our approach enriches the resources in DAJEE with semantic entities by applying state-of-the-art semantic techniques. The result is a significant improvement of the categorization of the resources in DAJEE than the previous version.

Keywords: Semantic Entities, OER, Linked Data for Education

1 Introduction

The DAtaset of Joint Educational Entities (DAJEE) provides a huge variety of learning resources coming from the popular Massive Open Online Course (MOOC) platform Coursera[1]. The novelty of DAJEE is the contextualization of the delivery of learning resources in lessons and courses [1]. This information has a potential for describing the teaching approaches of the author of the course, like, for example, concept sequencing and semantic density [2].

In Coursera, MOOCs are grouped in 10 top-level categories, each one with a different number of sub-categories, while the resources do not have any categorization. In DAJEE, the category of a resource is inherited from the related MOOC (reproducing the same categorization stated by authors). However, Coursera offers only a shallow categorization with at most two levels of categories, with some top-level categories without any sub-category. Hence, hundreds of resources are grouped into just one category with no additional diversification among them. Also, educational resources

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

may belong to categories that differ from the ones of their course. This limited categorization of the resources is currently replicated in DAJEE.

We propose a method to enhance DAJEE with a more-in-depth categorization of the resources based on their content, instead of their course. The proposed method firstly exploits Semantic Web techniques and data offered by DBpedia for interlinking the content of a resource with semantic entities in DBpedia [4]. The hierarchical structure of the categories of DBpedia is used for building the category graph of the resource. Then, we propose an application of Dijkstra's algorithm and the Spreading Activation technique [3] to reduce the noise that may be introduced when interlinking the resources with DBpedia entities, and, so, to refine the overall category graphs. The result of this method is a significant improvement of the categorization in DAJEE, more detailed and tailored directly to the content of the resources. A fine-grained categorization improves the browsing of the educational resources and supports further applications such as category-based retrieval and recommendation system.

2 Categories in DAJEE

The shallow category structure of Coursera is linked to the courses, not to the resources. Instead, DBpedia offers a much more detailed category structure that can directly categorize the resources. As an example, the category *Math and Logic* in Coursera has no subcategories, while in DBpedia the equivalent category (named *Mathematics*) has several subcategories with different levels of depth.

In DBpedia, semantic entities have many categories. For achieving our goal, resources in DAJEE should be associated with categories that are tailored to the semantic entities extracted from their transcripts. To keep trace of the categorization of a semantic entity, a sub-graph of the DBpedia category structure describes an entity. Such sub-graph starts with the source category and ends with the categories stated in the DBpedia page of the entity. Since each transcript has a number of entities, each one with a sub-graph, the main problem is how to properly merge these sub-graphs for a correct and meaningful categorization of a resource or a text.

A simple merge of the category-graphs of the entities associated with the resource is not an efficient solution. DBpedia entities can present a very wide set of categories and some of them may be poorly or not at all related to the resource. Let us consider the resource *Generic birthday attack* from course *Cryptography I* as a reference. The DBpedia entity *dbr:Cryptographic_hash_function*[2] has been found in its transcript, and Figure 1 shows its category graph. The graph includes *Science*, *Mathematics*, and *History*, while other categories are descendant of *Business* and *Belief* which seem to be unrelated to the resource. For a more effective categorization of the resource, we suggest a novel method for filtering DBpedia categories based on the Spreading Activation (SA) [3] and Dijkstra's algorithm. For each entity extracted from the resource text, the spreading phase starts giving one activation unit only to the categories explicitly stated in the DBpedia page of the entity. Then, edges are weighted in three

[2] http://dbpedia.org/resource/Cryptographic_hash_function

steps: 1st step) Dijkstra's algorithm finds the shortest paths among the categories stated in the DBpedia page of the entity, if any; 2nd step) the edges receive a weight which is 1 if they are part of those paths; 3rd step) the weight is set to 1 for all the edges that connect the categories on the DBpedia page with the root of the graph.

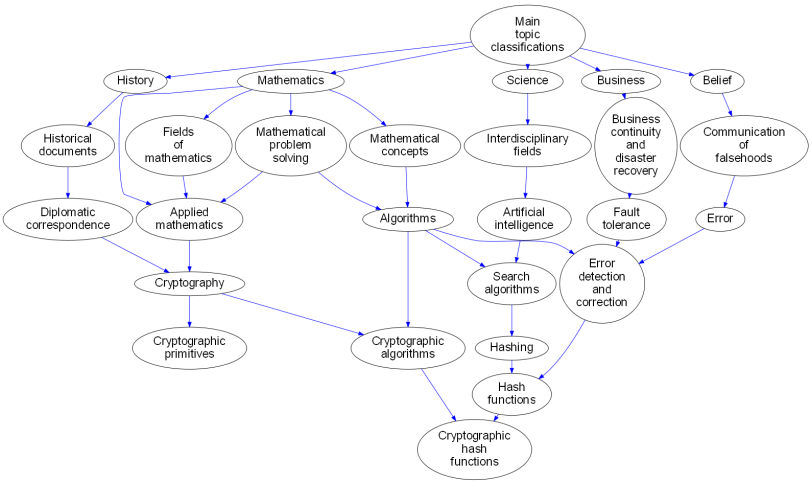

Figure 1: Category graph for entity *dbr:Cryptographic_hash_function*.

For deducing the most important top-level categories for an entity, the activation is spread throughout the category graph opposite to the edges direction (from "child" to "parent"). The activation for a category *j* is regulated as follows:

$$IngoingActivation(j) = \sum_{i \in ingoingEdges(j)} w_{ij}$$

$$OutgoingActivation(j) = \begin{cases} 1, & if\ IngoingActivation(j) > 0 \\ 0, & otherwise \end{cases}$$

where w_{ij} is the weight of an edge from category *i* to *j*. The algorithm stops when the *OutgoingActivation* is 0. Figure 2 (left) reports the activations of the top-level categories, finding that *Generic birthday attack* is mostly about *Mathematics* (57% of the activations). The same process can also filter the sub-categories of the most frequent top-level categories, removing edges with weight 0 and sub-categories with no activation from the graph of the resource[3]. Interestingly, the resource *Alpha Beta Pruning* is identified as *Mathematics* as well, but it is different from *Generic birthday attack.* Our method shows that the category graph for *Alpha Beta Pruning*[4] is focused on graph theory and algorithms, while *Generic birthday attack* presents many connections to mathematical analysis and algebra. So, we can further distinguish resources

[3] The final category graph for the resource *Generic birthday attack* is available at: http://virtuosa.pa2.itd.cnr.it/iswc17/generic_birthday_attack.png

[4] The resulting graph is available at: http://virtuosa.pa2.itd.cnr.it/iswc17/alpha_beta_pruning_graph.png

belonging to a same top-level category, like *Generic birthday attack* and *Alpha Beta Pruning*. Applying our methodology, a more meaningful categorization tailored on the resource content is now included in DAJEE.

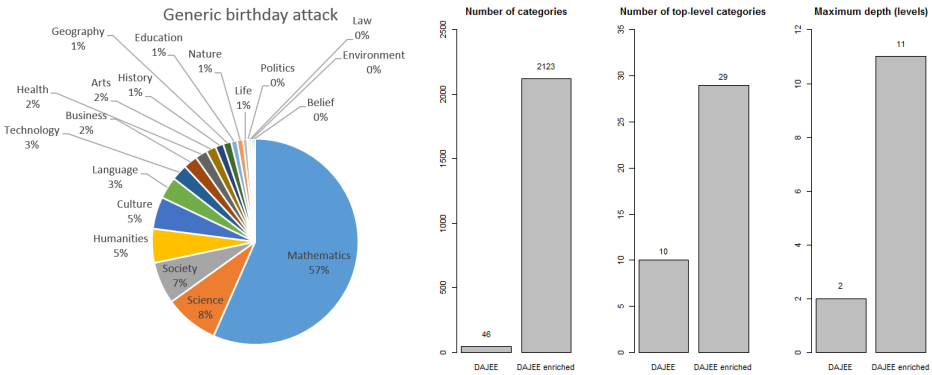

Figure 2: Activations of different top-level categories for the entities in resource *Generic birthday attack* (left) and Overall analysis of the effects of the proposed categorization on DAJEE (right).

Figure 2 (right) reports the improvement of the categorization of the resources in DAJEE. The total number of categories is now more than 2,000, originally only 46. The top-level categories are now 29 instead of just 10. Categories previously grouped together (e.g., Arts and Humanities are now two separate top-level categories) are now separated, and other top-level topics of resources in DAJEE are discovered. Also, the maximum number of levels for the category tree increases from 2 to 11.

The application of our method successfully provides a much more detailed categorization of the resources in DAJEE, finally enabling a more accurate analysis of the content in DAJEE and textual resources in general.

3 References

1. Estivill-Castro, V., Limongelli, C., Lombardi, M., and Marani, A. Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (2016), ACM, pp. 681-684.
2. Limongelli, C., Lombardi, M., Marani, A., and Taibi, D. Enrichment of the dataset of joint educational entities with the web of data. In 17th IEEE International Conference on Advanced Learning Technologies (ICALT'17) (2017), IEEE.
3. Crestani, F. Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11, 6 (1997), 453-482.
4. Dietze S., Yu H. Q., Giordano D., Kaldoudi E., Dovrolis N., Taibi D. 2012. Linked Education: interlinking educational Resources and the Web of Data. ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications.