Towards the recommendation of resources in Coursera

Carla Limongelli[1], Matteo Lombardi[2], and Alessandro Marani[2]

[1] Engineering Department, Roma Tre University,
Via della Vasca Navale, 79 - 00146 Roma, Italy
limongel@ing.uniroma3.it
[2] School of Information and Communication Technology, Griffith University,
170 Kessels Road, Nathan, QLD, 4111 Australia
{matteo.lombardi,alessandro.marani}@griffithuni.edu.au

Abstract. Technology Enhanced Learning (TEL) largely focuses on the retrieval and reuse of educational resources from Web platforms like Coursera. Unfortunately, Coursera does not provide educational meta-data of its content. To overcome this limitation, this study proposes a data mining approach for discovering Teaching Contexts (TC) where resources have been delivered in. Such TCs can facilitate the retrieval of resources for the teaching preferences and requirements of teachers.

Keywords: MOOCs, Educational Data Mining, Coursera

1 Introduction

Many contributions in TEL propose Information Retrieval (IR) systems of educational web resources [4, 5] and most of these works are focused on students; only some recent contributions address the role of instructors [2–4]. The web hosts many e-learning platforms that help instructors in delivering their resources or courses. Particularly attractive is the idea of Massive Open Online Courses (MOOCs) where courses are delivered and publicly available worldwide. Coursera is an on-line platform with plenty of reliable MOOCs delivered by prestigious universities: a very attractive source of educational data. However, it does not offer educational meta-data of its content, so this paper suggests a clustering technique for deducing some representative TCs useful for IR systems in TEL. In this contribution, a TC consists of i) the teaching preferences of an instructor, ii) course information and iii) lesson information.

2 Clustering Coursera data

This study suggests to apply Hierarchical Clustering (HC) on DAJEE [1], which is a TEL dataset built from Coursera data. The TCs are deduced from the analysis of the three main educational entities or hierarchies in the dataset: instructors, courses and lessons where the resources have been delivered in. The features obtained after the pre-processing of data are the following: *average duration of resources*, *average number of resources per concept* and *average duration*

of concepts for instructors, *duration* and *semantic density*[3] for both courses and lessons. To identify distinctive educational models at each level, it is suggested to cluster data in an hierarchical manner.

1. **Instructors**: for clustering instructors instances, both K-Means (with K from 2 to 242) and Expectation-Maximization (EM) have been run. The best configuration indicated by the Calinski-Harabasz (CH) index [6] is K-Means with K=2.
2. **Courses**: each cluster of instructors is further divided considering courses models taught by instructors in a same cluster. It is harder to suggest a value of K for K-Means that is appropriate for any change introduced by the first level of clustering. Moreover, the CH index is sensible to the data [6], so CH cannot be used for finding K once-for-all. Therefore, we suggest EM for this level, so that the most appropriate mixture models are defined for any partition of data derived by the upper level.
3. **Lessons**: each course cluster is split using the models of lessons; this is the same situation of the upper level, so EM is suggested.

Finally, the resources used for lessons in a same lesson cluster are grouped together. A total of 27 clusters indicate the most representative TCs in Coursera.

3 Conclusions

With the proposed HC method, 27 TCs have been discovered from data in Coursera. These contexts can be used for retrieving resources from MOOCs appropriate for the specific teaching situation of an instructor. In the near future, this assumption has to be proved with a large experimentation of an IR system based on our method.

References

1. Vladimir Estivill-Castro, Carla Limongelli, Matteo Lombardi, and Alessandro Marani. Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. In *Proceedings of the 39th International ACM SIGIR Conference*. ACM, 2016.
2. Carla Limongelli, Matteo Lombardi, Alessandro Marani, and Filippo Sciarrone. A teacher model to speed up the process of building courses. In *Human-Computer Interaction. Applications and Services*, pages 434–443. Springer, 2013.
3. Carla Limongelli, Matteo Lombardi, Alessandro Marani, and Filippo Sciarrone. A teaching-style based social network for didactic building and sharing. In *Artificial Intelligence in Education*, pages 774–777. Springer, 2013.
4. Carla Limongelli, Matteo Lombardi, Alessandro Marani, Filippo Sciarrone, and Marco Temperini. A recommendation module to help teachers build courses through the moodle learning management system. *New Review of Hypermedia and Multimedia*, pages 1–25, 2015.
5. Matteo Lombardi and Alessandro Marani. A comparative framework to evaluate recommender systems in technology enhanced learning: a case study. In *Advances in Artificial Intelligence and Its Applications*, pages 155–170. Springer, 2015.
6. Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12):1650–1654, 2002.

[3] In this work, semantic density is the ratio of number of concepts on the duration of the educational entity (i.e. courses or lessons) following IEEE 1484.12.1-2002.