Generating Semantic Concept Map for MOOCs

Zhuoxuan Jiang[1], Peng Li[1], Yan Zhang[2], Xiaoming Li[1]
School of Electronics Engineering and Computer Science
Peking University, Beijing, China
[1]{jzhx,lipengcomeon,lxm}@pku.edu.cn, [2]zhy@cis.pku.edu.cn

ABSTRACT

The task of re-organizing the teaching materials to generate concept maps for MOOCs is significant to improve the experience of learning process, e.g. adaptive learning. This paper introduces a novel and tailored Semantic Concept Map (SCM), and we design a two-phase approach based on machine learning methods to generate it.

1. INTRODUCTION

With the increasing development of Massive Open Online Courses (MOOCs) in recent years, it is believed that how to efficiently re-organize the course materials to serve for better learning is worthy of discussion [6].

In the traditional computer-assisted education, concept map is useful but usually involves domain experts. Considering the large amount of MOOCs, an information system that behaves like an expert and provides the skeleton of a concept map can be more effective.

Unlike partially organized e-textbooks, we can not directly identify concepts from various MOOC materials merely through stylistic features, so machine learning based method is leveraged. Moreover, in order to reduce the cost of labelling, semi-supervised framework is adopted in this paper. Rather than generating various relationships between concepts, we define a novel Semantic Concept Map (SCM) which considers semantic similarity as the only relationship without regard to complex and hierarchy ones. Due to its concision and universality, this map can be applied widely to more courses. Figure 1 shows the two-phase approach including 1) concept extraction and 2) relationship establishment.

2. RELATED WORK

Plenty of work about automatically constructing concept maps has been studied with data mining techniques, such as association-rule mining, text mining and specific algorithms [7]. However, these methods are designed for either specific data sources or special learning settings. Due to the diversity of MOOCs settings, they can hardly be leveraged here.

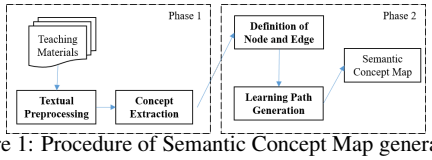


Figure 1: Procedure of Semantic Concept Map generation.

The task of terminology extraction in computer science field is similar to our machine learning based concept extraction [1], but those methods mainly concern about proper nouns or named entity recognition (NER) for generating knowledge graph [5]. Actually this kind of task is corpus-dependent.

3. GENERATING SEMANTIC CONCEPT MAP

Semantic Concept Map. SCM is composed of entities and edges. Formally, denote $SCM = \{C, R\}$ where $C = \{c_1, c_2, ..., c_n\}$ is a set of concepts. Each concept c_i is denoted by a terminology (including phrase), and unique in C. $R = \{r_{11}, r_{12}, ..., r_{ij}, ..., r_{nn}\}$ is a set of relationships between concepts. Each weight value r_{ij} means the degree of semantic similarity between c_i and c_j. The key steps shown in Figure 1 are following.

1.Textual Preprocessing. This step includes tokenization, filtering stop words and removing code and html tags, as well as word segment for Chinese if necessary. We also conduct conflation. All data are randomly shuffled before being learnt and tested, which is partially equivalent to cross-fold validation.

2.Concept Extraction. We leverage CRF+semi-supervised framework to solve this task as a problem of sequence annotation [2]. The labels needed to be predicted of each word are defined as three categories: B, I and O, which respectively mean the beginning word of a concept, the internal word of a concept and not a concept. Feature definition is a key part of machine learning method. Then we design the course- and instructor-agnostic features to meet the diverse materials including stylistic, structural, contextual, semantic and dictionary features. In order to reduce the heavy cost of human labeling, the idea of self-training is leveraged when training data [3].

3.Definition of Node and Edge. The weights of nodes could have different definitions. For example, the more frequent a concept is present in the lecture notes, the more fundamental it is. So the metric of term frequency (tf) can be defined as the node weights, named for *fundamentality*. The diverse teaching materials put together are partitioned to documents corresponding to each video. Moreover, low-frequency concepts may be the key ones of each corresponding unit. So we can define the second metric, Term Frequency and

Table 1: Performance of different concept extraction methods.

	Precision	*Recall*	*F1*
TF@500	0.402	0.500	0.446
TF@1000	0.600	0.746	0.665
BT	0.099	0.627	0.171
SC-CRF	**0.890**	**0.842**	**0.865**
SSC-CRF	0.875	0.783	0.826

Inverted Document Frequency (*tfidf*) which is ideal for quantifying the importance of a concept. As to the weights of edges, the Cosine distance of two word vectors of concepts are defined as the semantic similarity, because the word vectors learnt by word2vec have a natural trait that semantically similar vectors are close in the Cosine space and vise verse [4].

4.Learning Path Generation. The learning path depends on the definition of node and edge in the last step. For example in terms of importance, starting from some concept, each time we choose top k most semantically similar concepts and regard the most important one within the top k as the next node of the path. When choosing the subset of top k candidates, we also consider their locational order of first appearance in the lecture notes.

4. EXPERIMENTS

We collect the teaching materials of an interdisciplinary course conducted on Coursera, including lecture notes (video transcripts), PPTs, questions. The instructors and two TAs help label the data.

We select several baselines to extract concepts from MOOCs materials for comparison. The preprocessing is identical for baselines.

- **Term Frequency (TF)**: This is a statistic baseline.
- **Bootstraping (BT)**: A rule-based iterative algorithm given several patterns which contain true concepts.
- **Supervised Concept-CRF (SC-CRF)**: A supervised CRF with all features but semi-supervised algorithm.

Table 1 shows the performance between baselines and our approach (SSC-CRF). The results also show the necessity of machine learning based methods. Figure 2 manifests that semi-supervised learning is competitive with supervised learning. But considering only half labor consumed, semi-supervised learning is feasible and necessary.

Based on the definitions of node and edge mentioned before, the two kinds of SCMs generated look like Figure 3. Starting from the most fundamental concept, *Node*, the first five successors on the path are: *Edge* → *Element* → *Set* → *Alternative* → *Vote*, which are from basic concepts to advanced ones. Starting from the most important concept, *PageRank*, the first five successors on the path are: *PageRankAlgorithm* → *SmallWorld* → *Balance* → *NashBalance* → *StructuralBalance*. We can see they are not only important along with the course syllabus, but also semantically similar.

5. CONCLUSION

In this paper we mainly propose an approach to re-organize existing teaching materials to generate a novelly-defined SCM for facilitating the learning process in MOOCs. This work is a promising start for content-based adaptive learning since hierarchical and multiple relationships of a complete concept map can be incrementally replenished, and meanwhile this map can be extended to more courses and domains. Experiments show a good efficacy of the semi-supervised

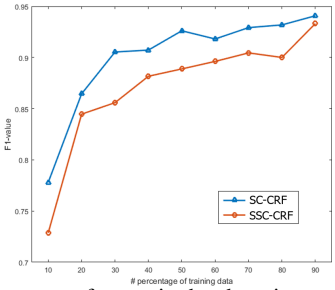
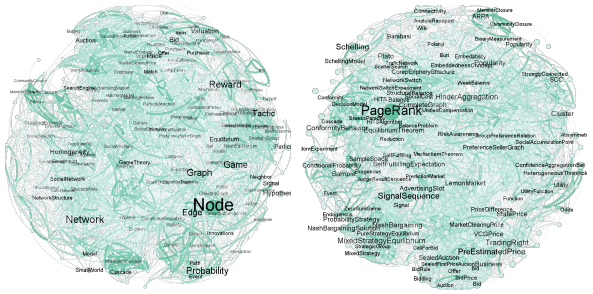


Figure 2: Performance of supervised and semi-supervised learning.



(a) For fundamentality (b) For importance

Figure 3: Two kinds of Semantic Concept Map.

machine learning algorithm and the CRF framework. And the learning paths defined based on SCMs can be humanly modified further to satisfy the requirements of different learners. In future work SCM could be utilized for generating course Wiki via crowdsourcing, hinting concept in forum discussions, etc. Large-scale student knowledge tracing in MOOCs is also doable by associating concepts with questions. Moreover, methods of transfer learning and deep learning may be more effective to extract the abstract concepts from multiple courses and diverse materials.

6. ACKNOWLEDGMENTS

This research is supported by NSFC with Grant No.61532001 and No.61472013, and MOE-RCOE with Grant No.2016ZD201.

7. REFERENCES

[1] K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey of the state of the art. In *ACL*, pages 1262–1273, 2014.

[2] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, pages 282–289, 2001.

[3] A. Liu, G. Jun, and J. Ghosh. A self-training approach to cost sensitive uncertainty sampling. *Machine Learning*, 76(2-3):257–270, 2009.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. `http://arxiv.org/abs/1301.3781`.

[5] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs, 2015. `http://arxiv.org/abs/1503.00759v3`.

[6] Z. A. Pardos, Y. Bergner, D. T. Seaton, and D. E. Pritchard. Adapting bayesian knowledge tracing to a massive open online course in edx. In *EDM'13*, pages 137–144, 2013.

[7] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. on Systems, Man, and Cybernetics*, 40(6):601–618, 2010.