Constructing knowledge map for MOOC using data mining methods

Junmeng Hou,Ruifang Liu,Yu Zhang

School of Information and Communication Engineering

Beijing University of Posts & Telecommunications,Beijing, China

e-mail: hjunmeng@163.com

Keywords: MOOC; data mining; k-means clustering; knowledge map

Abstract. Although Massive Open Online Courses(MOOCs) have become a way of online study used by millions of peopleacross the world, so many websites and courses often confuse people that they can't choose the courses they need quickly and accurately. In this paper, we present an approach that constructknowledge map for courses using data mining methods. The data mining algorithm combines all the courses of several MOOC websites, and takes the text descriptions of the courses given in the website into account. People can choose corresponding knowledge map for a series of courses according to their needs.

Introduction

Amassive open online course (MOOC) is an online course aimed at unlimited participation and open access through the web. MOOCs are widely seen as a major part of a larger disruptive innovation taking place in higher education. The MOOCs broaden the scope of the online learning in addition to traditional courses materials such as filmed lectures and readings, many MOOCs provide interactive user forums to support community interactions between students and professors. The MOOC provides a more convenient way to learn the course that we need.

In the past few years, with their dramatically increasingpopularity,MOOCs have become a way of online learningused by millions of people across the world.As a result of efforts conducted by academia and industry, more and more MOOC providers have emerged and too many online courses have been provided[1].This situation has resulted in people can't find the courses they needed quickly and accurately. The process of choosing wastes a lot of time. And what they need maybe a series of courses. To solve this problem, we try to combine the algorithms of data mining with course information to construct theknowledge map.

Data mining is the process of discovering interestingknowledgefrom the large database. Data mining has been used in many fields include telecommunication industry, education[2], and etc. Data mining involves tasks such as association rule learning, classification, clustering,regression and summarization.The combine of data mining and MOOC[3] at present is mainly based on people's behavior[4, 5] to enhance web-based learning environments for the educator to better evaluate the learning process. For example,we can predict MOOC dropout overweeks[6],or we can recommend courses for people based on historical data[7]. But the problem we want to solve is when people start online learning, how to choose a series of courses they need from massive websites and courses. And the data mining algorithm takes the text information of the courses given in the website into account.

The rest of this paper is organized as follows.We begin with the related work about MOOC in section II.We describeour model to make the knowledge map in section III. Next, we describe the data set and

preprocessing stage, and the experiment results are presented in section IV.Finally,weconclude our workin section V.

Related work

The task of combination of data mining and MOOCs has been the topic of several recent studies. In this context, there are two main research directions. One is to predict MOOC dropout over weeks using machine learning methods[8,9]. Another is to recommend courses for students.

Discussion forum of MOOC websites serves as a rich source of information that offers insights into many aspects of students' behavior[10,11]. This is the theme of studies such as Wen and Yang[12], which focuses on using computational linguistic models to measure learners'motivation and cognitive engagement from the text of forum posts. Marius and Felix[13] seek to present an approach that works on click-stream data to predict MOOC dropout over weeks. Among other features,the machine learning algorithm takes theweekly history of students' data into accountand thus is able to notice changesinstudents'behavior over time.

In KDD cup 2015[14], they predicted dropout on Xue-tangX, one of the largest MOOC platforms in China. The participants of the competition needed to predict whether a user will drop a course within the next 10 days based on his or her prior activities. If a user leaves no records for course C in the log during the next 10 days, we define it as dropout from course C. The prediction is based on user's prior activities and the data is only from one MOOC website.

The recommendation of courses always bases on the historical data. In this context,Sunita and Aher[15] recommend courses to a new student who has recently enrolled for some courses through the choice of other students for particular set of courses collected from Moodle. Naveen[16] presented the existing efforts in the direction of adaptive learning. In this approach, they organized learning material into a four layered architecture. The representation is the combination of hierarchical and network based approach.

Here in this paper,we take the course information given in the website into account. The problem that MOOCs need to solve is to serve student's autonomy rather than simply giving a planned learning program. In this study,the task is to construct the knowledge map for the courses so that students can choose a series of system courses that they need quickly and accurately.

Methods

Course representation

The description of a course can be thought as a document. First we process the document,including word segmentation and removing the stopping words. Then each document can be represented as a vector of words like equation(1).

$$d = \{(t_{1,}w_{1d}), (t_{2,}w_{2d}), \dots (t_{s,}w_{sd})\}. \quad (1)$$

Wheret_k is a word of the document and w_{kd} is the weight oft_k. The weight of a word in a document is often calculated as equation(2).

$$w_{kd} = \mathrm{TF}(t, d) \times \mathrm{IDF}(t). \quad (2)$$

Term Frequency(TF) is intended to reflect how important a word is to a document in a collection or corpus. TF(*t, d*) is the ratio of the frequency of the word*t* to the maximal frequency of the word in the document.The Inverse Document Frequency(IDF) is a measure of how much information the word provides,that is,whether the term is common or rare across all documents.IDF(*t*) is the logarithmically scaled fraction of the documents that contain the word, obtainedbydividing the total number of

documents by the number of documents containing the term *t*,and then taking the logarithm of that quotient. TF-IDF is the product ofthetwo statics.

In our study,we try to extract keywords from the course description, and get the weight of keywords with TF-IDF. The resultof the pre-processingis used to calculate the value of TF-IDF. The calculation of the document collection will generate a two-dimensional matrix.

Course grouping

Each row of the matrix represents one course,and we randomly select *k* samples. Then we use k-means clustering model to divide all the courses into different groups.K-means clustering is a method of vector quantization, and it is popular for cluster analysis in data mining because of the method's simplicity and efficiency. The main idea of k-means clustering is to partition n courses into *k* clusters in which each course belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given a set of courses$\{d_{1,}d_2, d_{3,}\cdots, d_n\}$, whereeach course is a real vector, we use k-means clustering to partition the n courses into k (≤n) sets $S = \{S_{1,}S_2, S_{3,}\cdots, S_k\}$ so as to minimize the within-cluster sum of similarity (WCSS).The objective function is equation(3).

$$\arg\min_{\mathrm{s}} \sum_{i=1}^{\mathrm{k}} \sum_{\mathrm{d}\in \mathrm{S_i}} \mathrm{sim}(d, u_i). \quad (3)$$

where u_i is the mean of points in S_i.

In order to calculate the similarity between a document and a cluster, we adopt the metric as equation (4).

$$\mathrm{sim}(d, u_i) = \frac{\sum_{t=1}^{M} w_{td} \times w_{tu_i}}{\sqrt{\sum_{t=1}^{\mathrm{M}} w_{td}^2 \times \sum_{t=1}^{M} w_{tu_i}^2}}. \quad (4)$$

In cluster analysis,the k-means algorithm can be used to partition the input data set into *k* clusters. The most common algorithm uses an iterative refinement technique.Given an initial set of *k* means, the algorithm proceeds by alternating between two steps:

Assignment step: Assign each course to the cluster whose mean yields the WCSS. Since the sum of squares is the squaredEuclideandistance, this isintuitively the "nearest" mean.

Updatestep:Calculate the new means to be the centroids of the courses in the new clusters.

Knowledge map

The results we get after k-means clustering are a variety of different groups of courses. The courses in the same clustering group are similar in the content, and the courses belong to the different clustering groups are dissimilar. Users can easily figure out what they are interesting.

One knowledge map is one clustering group we get from the k-means clustering. And the courses in the map are the courses belong to the same group. From the map users can find all courses of one kind, and the details about the courses including the name of the course, the category of the course,the description of the course,the source of the course, the university (or provider) and etc. On the basis of this information, users can make comparison of the similar courses so that they can choose the courses they really need. What is more important, they can compare the assessment information of the similar courses so that they can find which course is better.

Data and Experiment

Data set

In this paper,we crawled and studied all the courses information given in the web from the three biggest MOOC platforms in China, including XuetangX, icourse163 and open163.

Thedataset consists of 1528 courses.The information of each courseconsists of name,description,university (or provider).

Experiment result

The description of a course can be thought as a document. Each description is about hundreds of words. Through calculating the TF-IDF value of the word segmentation results, we can get a two-dimensional TF-IDF matrix.

For example, the corpus is consisted of part of description of three courses. The corpus is

"*The course carefully selects classics of Chinese past dynasties as teaching content.*

The course mainly introduces the single variable differential and integral.

Calculus is the floorboard of differential and integral."

We can get the TF-IDF matrixwhich is 3*16.

The first row of the matrix is

{0.27626457, 0.36325471, 0.36325471,0.36325471, 0.36325471, 0.36325471, 0.36325471, 0.36325471, 0, ⋯, 0 }.

The second row of the matrix is

{0.34949812, 0, ⋯, 0, 0.45954803, 0.45954803, 0.45954803, 0.34949812, 0.34949812, 0, 0, 0}.

And the third row of this matrix is

{0, ⋯, 0, 0.37302199, 0.37302199, 0.49047908, 0.49047908, 0.49047908}.

From the matrix we can find that the second description is more similar with the third description.When the corpus consists of all courses description, we can get a matrix which is1528*11680. Each row of the matrix represents a document,and each column of the matrix represents the weighted value of the word for courses. Through the analysis of the matrix,we can find that most of the value is 0 for each course. Then we use the two-dimensional matrix to cluster. And the clustering result is shown in Fig. 1.

In Fig. 1, each number of the figure represents a clustering group, and the course which has the same number belongs to the same group. We can easily find that the courses close in distance corresponding to similar or cross disciplines. Through the clustering method,we divide all the 1528 courses from three MOOC platforms into 15 groups, and the courses of each group are in the similar theme.

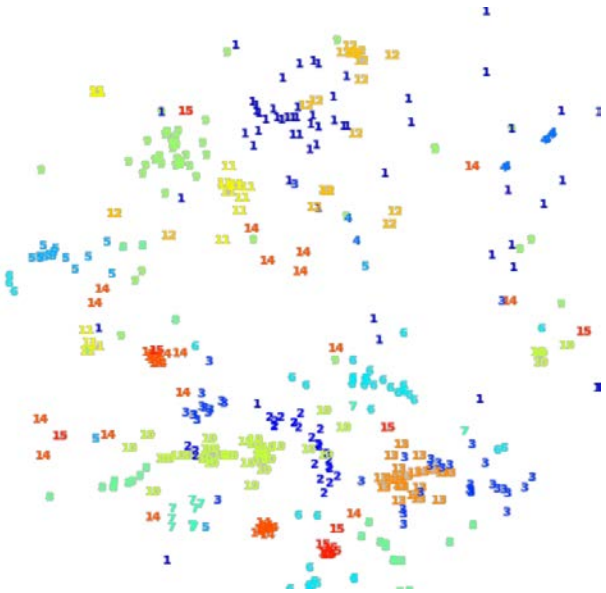

Figure 1. Clustering result

TABLE I. KNOWLEDGE MAP OF ENTREPRENEURSHIP COURSES

Course name	Category in the website	Source
Entrepreneurship education	economic	open.163
How to startup	managment	open.163
Collegestudents'entrepreneurship base	economic	open.163
Entrepreneurshiptechnology manage-ment	economic	open.163
Learn entrepreneurship together	economic	open.163
Entrepreneurship management	economic	open.163
Start a new business	economic& managment	XuetangX
Entrepreneurship trip	economic& managment	XuetangX

TableI is one knowledge map analysis.Courses listed in the table are all courses under one group from the clustering results. From the table we found that all courses listed are related to entrepreneurship, although they are different category in the same website or they are from different MOOC website.

So we can conclude that the knowledge map for courses in MOOC websites we get through data mining methods is effective. There are some course groups that were not originally in the website,such as entrepreneurship. Or some courses in our knowledge map belong to different categories in the same website, forexample,some courses related to entrepreneurship belong to economic class,but some similar courses belong to management class. The most common situation is the courses are from different website,such as how to "startup" is from open.163, while "start a new business" belongsto XuetangX.

We hold the opinion that the knowledge map is useful for students,they can find a series of courses they need quickly including the detailsof these courses. They can learn all the courses of the knowledge map, but also they can compare to choose the course which is better for them.

Conclusion and Future Works

Conclusion

In this paper, we present a study on how to construct the knowledge map for courses in MOOCs. We take the course description into account and using data mining methods. We extract keywords using TF-IDF,input the results to the k-means model.Through the k-means clustering, every group we get is the knowledge map we need and the courses belong to the group are in the same theme. The clustering results analysis validates that our knowledge map is effective. The knowledge map contains a list of courses from different website of the same theme so that people can find the courses they need quickly and accurately.

Future works

This study can also be used to recommend courses to the student who has already joined one course. They may need other related courses. We have constructed the basic implementation model but, the requirements are not fully captured in formal way. In our method to construct knowledge, we use the description of the course only.In this sense,we can make use of some other text information to train the model more comprehensively, such as evaluation of students to the course. The objective of introducing student's evaluation is to contrast the similar courses from different website. With the rapid development of MOOCs website this approach of constructing knowledge map for courses can be immensely be useful.

References

[1] J. Mackness, S. Mak, and R. Williams, “The ideals and reality of participating in a mooc,” Proc. Networked Learing Conference, 2010, pp. 266-275.

[2] Sunita B. Aher, and L.M.R.J. Lobo,“Data mining in educational system using WEKA,” Proc. Proceedings on International Conference on Emerging Technology Trends(ICETT), IJCA (International Journal of Computer Application), Mar. 2011, pp. 20-25.

[3] Hamalainen, J. Suhonen, E. Sutinen, and H. Toivonen,“Data mining inpersonalizing distance education courses,” Proc. World Conference on OpenLearning and Distance Education, 2004, pp. 1–11.

[4] JWang Y, Kraut R, and Levine J. M., “To stay orleave? the relationship of emotional and informationalsupport to commitment in online health support groups,” Proc. Proceedings of Computer Supported CooperativeWork, 2012, pp. 833-842.

[5] Tanmay Sinha1, Patrick Jermann2, Nan Li3, and Pierre Dillenbourg3, “Your click decides your fate: Inferring Information Processing andAttrition Behavior from MOOC Video Clickstream Interactions,” Proc. EMNLP 2014 workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Oct. 2014, pp. 3-14.

[6] Girish, Balakrishnan, “Predicting Student Retention inMassive Open Online Courses using Hidden Markov Models,” Proc. EECS Department, University of California, 2013.

[7] Sunita B. Aher,L.M.R.J. Lobo, “A comparative study of association rulealgorithms for course recommender system,” E-learning, InternationalJournal of Computer Applications, vol. 39, Jan. 2013, pp. 48-52.

[8] Carolyn Rose ,and George Siemens, “Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses.,” Proc. EMNLP 2014 workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Oct. 2014, pp. 39-41.

[9] Carolyn Rose ,and George Siemens, “Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses.,” Proc. EMNLP 2014 workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Oct. 2014, pp. 39-41.

[10] Beukeboom, C. J., Tanis, M., and Vermeulen,I. E, “TheLanguage of Extraversion Extraverted People Talk MoreAbstractly, Introverts Are More Concrete,”Journal of Languageand Social Psychology, vol. 32, Feb. 2013, pp. 191-201.

[11] Roscoe, R. D., and Chi, M. T. H, “Tutor learning: Therole of explaining and responding to questions.,”InstructionalScience, vol. 36, 2008, pp. 321-350.

[12] Miaomiao Wen,Diyi Yang, and Carolyn PensteinRos´e., “Linguistic reflections of studentengagement in massive open online courses,” Proc. ICWSM, Aug. 2014.

[13] Marius Kloft, Felix Stiehler, and Zhilin Zheng,“Predicting MOOC Dropout overWeeks Using Machine Learning Methods,” Proc. EMNLP 2014 workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Oct. 2014,pp. 60-65.

[14] https://kddcup2015.com.

[15] Sunita B. Aher, L.M.R.J. Lobo, “Combination of machine learning algorithms for recommendation ofcourses in E-Learning System based on historical data,” Knowledge-Based systems, vol. 51, Oct. 2013,pp. 1-14.

[16] SunitaB.Aher, L.M.R.J. Lobo, “A Frameworkfor Recommendation of courses inE-learning System,”International Journal of Computer Applications,vol.35, Apr. 2011, pp. 21-28.