Received 2 December 2016
Revised 10 January 2017
17 January 2017
Accepted 19 January 2017

Application of keyword extraction on MOOC resources

Zhuoxuan Jiang
Peking University, Beijing, China

Chunyan Miao
Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore, Singapore and School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore, and

Xiaoming Li
Peking University, Beijing, China

Abstract

Purpose – Recent years have witnessed the rapid development of massive open online courses (MOOCs). With more and more courses being produced by instructors and being participated by learners all over the world, unprecedented massive educational resources are aggregated. The educational resources include videos, subtitles, lecture notes, quizzes, etc., on the teaching side, and forum contents, Wiki, log of learning behavior, log of homework, etc., on the learning side. However, the data are both unstructured and diverse. To facilitate knowledge management and mining on MOOCs, extracting keywords from the resources is important. This paper aims to adapt the state-of-the-art techniques to MOOC settings and evaluate the effectiveness on real data. In terms of practice, this paper also tries to answer the questions for the first time that to what extend can the MOOC resources support keyword extraction models, and how many human efforts are required to make the models work well.

Design/methodology/approach – Based on which side generates the data, i.e instructors or learners, the data are classified to teaching resources and learning resources, respectively. The approach used on teaching resources is based on machine learning models with labels, while the approach used on learning resources is based on graph model without labels.

Findings – From the teaching resources, the methods used by the authors can accurately extract keywords with only 10 per cent labeled data. The authors find a characteristic of the data that the resources of various forms, e.g. subtitles and PPTs, should be separately considered because they have the different model ability. From the learning resources, the keywords extracted from MOOC forums are not as domain-specific as those extracted from teaching resources, but they can reflect the topics which are lively discussed in forums. Then instructors can get feedback from the indication. The authors implement two applications with the extracted keywords: generating concept map and generating learning path. The visual demos show they have the potential to improve learning efficiency when they are integrated into a real MOOC platform.

Research limitations/implications – Conducting keyword extraction on MOOC resources is quite difficult because teaching resources are hard to be obtained due to copyrights. Also, getting labeled data is tough because usually expertise of the corresponding domain is required.

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, China NSFC with Grant No. 61532001 and No. 61472013, and China MOE-RCOE with Grant No. 2016ZD201.

International Journal of Crowd Science
Vol. 1 No. 1, 2017
pp. 48-70
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-12-2016-0003

Practical implications – The experiment results support that MOOC resources are good enough for building models of keyword extraction, and an acceptable balance between human efforts and model accuracy can be achieved.

Originality/value – This paper presents a pioneer study on keyword extraction on MOOC resources and obtains some new findings.

Keywords Concept map, Graph model, Keyword extraction, Learning path, Massive Open Online Courses (MOOCs)

Paper type Research paper

1. Introduction

In recent years, massive open online courses (MOOCs) have benefited tens of millions of students all over the world. A very important characteristic of MOOC is it provides a one-stop online learning environment which consists of lecture videos, assignments, email notifications, discussion forum, and quizzes and examinations. Along with the popularity of MOOCs, a large amount of online educational resources of various subject areas, ranging from humanity to science, are unprecedentedly produced. Not only instructors can provide videos, subtitles, lecture notes, questions, etc., but also learners can generate forum contents, Wiki posts, log of homework submissions, etc. In fact, each MOOCs platform is a large-scale "knowledge base" where the educational resources can be regarded as the outcome of crowd intelligence (from both instructors and learners). However, those resources are unstructured and diverse. For example, subtitles are well-organized and formal, as they are usually produced by instructors, whereas the contents of posts are written by different learners; thus, they are colloquial and informal. As Figure 1 shows, MOOC resources can be considered as teaching resources and learning resources. By proposing proper models to discover knowledge from the crowd intelligence, it is promising to implement knowledge management, knowledge mining and even smart education for MOOCs.

This paper explores the task of keyword extraction and its applications on MOOC resources. The reason for conducting this task is that in most work of knowledge engineering, e.g. construction of knowledge graph and knowledge management, entity extraction is the first step. As for our task, we call the "entity" as keyword. The meaning of keyword in the educational setting is regarded general and intuitively can include concept, terminology, named entity and so on. Keyword extraction from MOOC resources may face several difficulties:

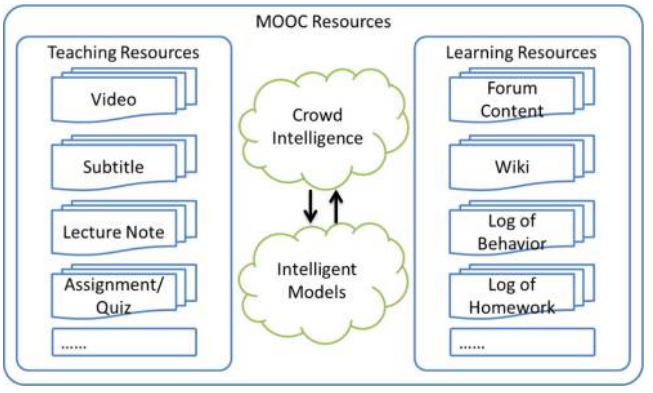

Figure 1.
Overview of MOOC resources

- MOOCs are of different subject areas, any domain-specific method should not help much, and as such, the method we use should be instructor- and course-agnostic;
- obtaining labeled training data set is extremely expensive, as usually domain expertise is required; and
- the volume of data is usually large and the textual styles are various.

Despite those difficulties, once keywords are well extracted, many subsequent applications are feasible, e.g. construction of course-specific and domain-specific concept map, management of cross-domain concepts, knowledge discovery from crowd and even personalized learning by mining learners' behaviors.

Based on the partition of who generates the MOOC resources, i.e. instructors and learners, the research design of this paper is composed by three parts:

(1) keyword extraction on resources generated by instructors;

(2) keyword extraction on resources generated by learners; and

(3) applications with keywords in MOOC settings.

As to the first part, it is difficult to collect entire instructor-generated resources of many courses. Also, labeling the data requires expertise in the corresponding subject area. Even so, we invite the instructors and teaching assistants (TAs) to help label the teaching resources of one course, as we expect to use human knowledge to learn a classifier by supervised machine learning methods. Moreover, we design a semi-supervised learning framework to test whether using less labeled data is practical. We regard this task as a problem of natural language processing, i.e. word sequence labeling. Sutton and McCallum (2011) believe that the probabilistic graphical models, especially conditional random fields (CRFs), can obtain the state-of-the-art performance in many sequence labeling tasks like part of speech (POS), named entity recognition (NER) and word segment, so we leverage this kind of model to extract keywords on MOOC teaching resources.

As to the second part of keyword extraction on resources generated by learners, i.e. discussion forum contents, it is relatively convenient to collect the contents of many courses. However, the number of posts may be quite large, e.g. over ten thousands, so it is difficult to use human knowledge through labeled data. On the other hand, as a kind of social media, forums have relational information between learners and contents. By referring to many methods of keyword extraction for social media, we model the MOOC forum to a heterogeneous network for each course. Then through graph-based random walk algorithm, keywords are extracted by ranking the importance of each word. We regard the top words in the ranking list as keywords.

After keywords are extracted, lots of novel educational applications can be developed within the MOOC settings. In the third part of this paper, we introduce two preliminary applications: generation of concept map and generation of learning paths. Romero and Ventura (2010) proposes that in the educational field, concept map is useful to organize, design and manage the course resources for instructors. We propose a new concept map which is called semantic concept map (SCM). The main difference with traditional concept maps is that the edge, i.e. relationship between keywords, is defined as semantic similarity. This kind of concept map can be easily extended to various courses. Then based on the SCM, we propose a method to automatically generate learning paths which have the potential for personalized learning.

In what follows, we review the related work in Section 2. Section 3 introduces data sets used in this paper. Section 4 introduces the method of keyword extraction on the side of teaching resources. The corresponding method on the learning side, i.e. forum contents, is introduced in Section 5. Then in Section 6, we report the experiment results obtained from both sides of resources respectively. In Section 7, we state the two demo applications with extracted keywords. Finally, we conclude this paper in Section 8.

2. Related work

Keyword is the word which people regard as important in a text. In different situations, keyword can be named entity, proper noun, terminology or concept. In this paper, the meaning of keyword is general. So if not otherwise specified, their differences are neglected.

In the past decades, Finkel *et al.* (2005), Nadeau and Sekine (2007) and Ratinov and Roth (2009) have studied the tasks of keyword extraction by machine learning methods, e.g. NER, terminology extraction and key phrases extraction. NER methods focus on named nouns, such as person name, location, time and address, and they are for constructing knowledge base, as seen from the papers by Dong *et al.* (2014) and Nickel *et al.* (2015). Terminology extraction methods are developed to extract domain-specific words. Recently, Nojiri and Manning (2015) and Qin *et al.* (2013) propose the methods based on machine learning for keyword extraction. However, methods for one kind of keywords extraction may not be used to another kind. For example, Nojiri and Manning (2015) exhibit that directly applying existing methods of NER to terminology extraction will not perform well. It is different to our task that we have labeled data.

Apart from supervised machine learning methods with human knowledge, another perspective for solving keyword extraction is the unsupervised approach. For example, Justesona and Katza (1995) propose the rule-based method, Frantzi *et al.* (2000) and Bin and Shichao (2011) propose the statistical methods. In this paper, we leverage a graph-based method proposed by Sonawane and Kulkarni (2014), which can model the social relationship between words to a network and then rank all the words in accordance with their importance.

To our knowledge, a large number of studies of data analytics on MOOC data have been proposed in recent years. For example, Anderson *et al.* (2014) try to classify MOOC learners after analyzing their behavior patterns. It also studies how to use a badge system to produce incentives based on learners' activity and contribution in the forum. Huang *et al.* (2014) analyze the behaviors of superposters in 44 MOOCs forums and finds that MOOCs forums are mostly healthy. Wen *et al.* (2014) study the sentiment analysis in MOOCs discussion forums and find that no positive correlation exists between the sentiment of posts and the course dropout. Wang *et al.* (2015) study the learning gain reflected through forum discussions. Jiang *et al.* (2015) conduct an analysis from the perspective of influence by modeling the MOOC forum to a heterogeneous network. Kizilcec *et al.* (2013) conduct a research on the behavior of learner disengagement. Moreover, some statistical reports and case study papers analyze behavior of MOOC learners, such as Ho *et al.* (2013) and Breslow *et al.* (2013). However, few studies of keyword extraction have been conducted on MOOC data.

Romero and Ventura (2010) and Novak and Cañas (2006) define that a concept map is a connected graph that shows relationships between concepts and expresses the hierarchical structure of knowledge. To our knowledge, plenty of work of automatically constructed concept map has been studied with data mining techniques. For example, Tsenga *et al.* (2007), Lee *et al.* (2009) and Qasim *et al.* (2013) leverage association-rule mining; Chen *et al.* (2008), Lau *et al.* (2009) and Huang *et al.* (2006, 2015) base on text mining; and Marian and

Maria (2009) and Chu *et al.* (2007) design specific algorithms. However, the majority of those methods are domain-specific, e.g. for specific courses or specific learning settings. We expect to explore new methods by reducing their dependency on domains, so a new kind of semantic relationship is leveraged in this paper.

3. Overview of a data sets

Also based on the partition of sources of generated resources, i.e. instructors and learners, we introduce the available MOOC data we have respectively.

3.1 Resources on teaching side

We collect the resources of an interdisciplinary course conducted in the fall of 2013 on Coursera. The course involves computer science, social science and economics. Textual content includes video subtitles, PPTs, questions and forum contents (i.e. threads, posts and comments). Table I shows the statistics of resources. We invited the instructor and two TAs to help label the data. As seen in Table I, the number of keywords in questions and PPTs are much smaller than that in subtitles. Based on our observation during labeling the data, the instructor and TAs would still spend much time on understanding each sentence, even though they should be more familiar with the contents than any others. We guess it is because the resources are composed by different people. During the activity of labeling data, everyone would spend about 8 h on labeling 3,000 sentences (in average 10 s per sentence).

A preprocessing step of word segment for Chinese may be necessary. We adopt the Stanford Word Segmenter[1] proposed by Chang *et al.* (2008). All data are randomly shuffled before they are processed late.

3.2 Resources on learning side

We collect data from 12 courses offered by Peking University from Coursera. They were offered in Fall Semester of 2013 and Spring Semester of 2014. There are totally over 4,000 threads and over 24,000 posts. For convenience later in the paper, Table II lists the pairs of course codes and course titles. Table III shows the statistics of the data sets per course. The "posts" denotes both posts and comments.

4. Keyword extraction on teaching side

The resources generated by instructors in MOOCs mainly include lecture notes, subtitles, PPTs and questions. In order to extract keywords from the teaching resources, we regard this task as a sequence labeling problem. It is similar to other sequence labeling tasks, e.g. NER and part-of-speech annotation. So probabilistic graphical models are the natural solution to this kind of tasks. Sutton and McCallum (2011) exhibits CRFs can achieve the state-of-the art performance. And we define instructor- and course-agnostic features in order to reduce the domain dependency. Moreover, we propose a semi-supervised learning framework to reduce human efforts of labeling data.

Table I.
Statistics of resources generated by the instructor: people and network

Source	# Sentence	# Word	# Keyword
Subtitles	3,036	69,437	402
PPTs	2,823	22,334	249
Questions	268	7,138	95

Course code	Course title
peopleandnetworks-001	People and networks
arthistory-001	Art history
dsalgo-001	Data structures and algorithms A
pkuic-001	Introduction to computing
aoo-001	Advanced object-oriented technology
bdsalgo-001	Data structures and algorithms B
criminallaw-001	Criminal law
pkupop-001	Practice on programming
chemistry-001	General chemistry (Session 1)
chemistry-002	General chemistry (Session 2)
pkubioinfo-001	Bioinformatics: Introduction and methods (Session 1)
pkubioinfo-002	Bioinformatics: Introduction and methods (Session 2)

Table II. Pairs of course code and course title

Course	# registrants	# threads	# posts	Maximum # posts per thread	Average # posts per thread	# votes	# overall certificates (ratio)	$P(C\|F)$	$P(F\|C)$
peopleandnetworks-001	10,807	219	1,206	38	5.5	304	149 (0.013)	0.271	0.584
arthistory-001	16,395	273	2,181	124	8.0	1,541	237 (0.014)	0.272	0.620
dsalgo-001	13,197	283	1,221	36	4.3	266	57 (0.004)	0.180	0.930
pkuic-001	14,462	1,029	5,942	141	5.8	595	285 (0.020)	0.290	0.782
aoo-001	9,563	97	515	63	5.3	204	53 (0.006)	0.160	0.528
bdsalgo-001	852	319	1,299	48	4.1	132	248 (0.320)	0.853	0.774
criminallaw-001	8,190	118	763	58	6.5	648	–	–	–
orgchem-001	4,374	28	85	12	3.0	11	12 (0.004)	0.091	0.250
pkupop-001	18,410	1,085	6,443	92	5.9	977	205 (0.012)	0.255	0.780
chemistry-001	9,124	110	591	40	5.4	65	116 (0.013)	0.400	0.448
chemistry-002	6,782	167	715	92	4.3	678	125 (0.020)	0.336	0.336
pkubioinfo-001	18,367	361	2,139	201	5.9	1,474	581 (0.032)	0.362	0.370
pkubioinfo-002	16,714	170	942	51	5.5	235	510 (0.032)	0.571	0.212
Overall	147,237	4,259	24,042	–	–	–	–	0.309	0.508

Notes: $P(C|F)$ represents the ratio of certificated forum learners to overall forum learners; $P(F|C)$ represents the ratio of certificated forum learners to overall certificated learners of the course

Table III. Statistics of resources generated by learners

4.1 Conditional random field's model

The problem of keyword extraction can be formally described as solving the conditional probability $P(Y|X)$. The random variable X refers to features of each sentence which follows a word sequence $x = \{x_1, x_2, \ldots, x_T\}$, and the random variable Y is a label sequence of the sentence $y = \{y_1, y_2, \ldots, y_T\}$. The label of a word is defined as three classes: *NO*, *ST* and *IN*. They respectively mean *not a keyword*, *the beginning word of a keyword* and *the middle word of a keyword*. So the label variable is $Y \in \{NO, ST, IN\}$.

We consider the conditional probability of labeling sequence Y, i.e. $p(Y|X)$, rather than their joint probability $p(Y, X)$, so linear chain CRFs framework proposed by Lafferty *et al.* (2001) is the natural choice. The conditional distribution over label sequence y, given an observation word sequence x, can be defined as:

$$p(y|\vec{x}) = \frac{1}{Z(\vec{x})} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, \vec{x}_t)\right) \tag{1}$$

where $Z(\vec{x}) = \sum_{y} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, \vec{x}_t)\right)$ and $\mathcal{F} = \{f_k(y_{t-1}, y_t, \vec{x}_t)\}_{k=1}^{K}$ are the set of feature functions defined on given x; $\Theta = \{\lambda_k\} \in \Re^K$ are parameter vector. N is the length of sentence and K is the number of features.

Given a training data set, the model $\Theta = \{\lambda_k\}_{k=1}^{K}$ could be learned by maximum likelihood estimation. To avoid overfitting, we add a regularized term to the function. Then the log-likelihood function of $p(y|x, \lambda)$ based on the Euclidean norm of $\lambda \sim (0, \sigma^2)$ is represented as:

$$L(\Theta) = \sum_{x,y} \log p(y|x, \Theta) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \tag{2}$$

So the gradient function is:

$$\frac{\partial L}{\partial \lambda_k} = \sum_{x,y}\sum_{t=1}^{T} f_k(y_{t-1}, y_t, x_t) - \sum_{x,y}\sum_{t=1}^{T}\sum_{y,y'} f_k(y, y', x_t) p(y, y'|x) - \frac{\lambda_k}{\sigma^2} \tag{3}$$

The detail of learning the CRFs model can be referred to Sutton and McCallum (2011). Then, given a new word sequence x^* and a learned model $\Theta = \{\lambda_k\}_{k=1}^{K}$, the optimal label sequence y^* could be calculated by:

$$y^* = arg \max_{y \in \mathcal{Y}} p(y|x^*, \Theta) \tag{4}$$

where $\mathcal{Y}$ is the set of all possible label sequences for the given sentence x^*. We use L-BFGS algorithm to learn the model and Viterbi algorithm to infer the optimal label sequence y^*.

4.2 Feature engineering

A crucial part of CRFs framework is the definition of feature functions. Based on our observation, we define five kinds of features which are adapted to our educational data. All the features are course-agnostic and make our framework flexible for scalability.

4.2.1 Text style features

- whether the target word is English;
- whether the two neighbor words are English;
- whether the word is the first word in a sentence;
- whether the word is the last word in a sentence; and
- whether the target word is in a quotation.

Text style features capture the stylistic characteristics. Some keywords usually appear at the beginning or the last of a sentence in instructor's language, e.g. "*Netwrok* means[. . .]" or "[. . .]This is the definition of *Network*". Because our data are from a Chinese MOOC, we

regard whether the word is English as a feature. Obviously, when it comes to English MOOCs, capitalization is the key feature of English keywords. So this kind of features are flexible to different situations.

4.2.2 Structure features

- POS tag of the target word;
- POS tag of the previous word; and
- POS tag of the next word.

We treat the POS as a feature because fixed combination of POS, e.g. adjective + noun or noun + noun, may indicate keyword phrases. We use the Stanford Log-linear POS Tagger [2] proposed by Toutanova *et al.* (2003) to assigns POS to each word. Note that as to the corresponding feature functions, we adopt binary value, 0 or 1, to every POS. For example, there is a function to capture whether the target word is a noun and so on.

4.2.3 Context features

- term frequency and inverted document frequency (TF-IDF) value of the target word and two neighbor words;
- normalized uni-gram BM25 score of the target word;
- normalized bi-gram BM25 score of the target word; and
- normalized bi-gram BM25 score of the two neighbor words.

Context features capture the importance of words and word-level information within the whole document. The training set is partitioned to documents based on video clips. Statistical metric of normalized bi-grams BM25 scores proposed by Robertson *et al.* (2004) is used to quantify word relevance by default parameters.

4.2.4 Semantic features

- semantic similarity of the target word with the previous two words; and
- semantic similarity of the target word with the next two words.

Some frequent-co-occurrence words may be keywords. Also, close words in the semantic space may be keywords. So by learning the word semantics, features of adjacent words can be captured. The similarity of two adjacent words in semantic space is calculated with the corresponding word vectors trained by Word2Vec[3] proposed by Mikolov *et al.* (2013). All textual contents are used to learn the word embeddings. The corpus size is 145,232 words and the vector dimension is set as 100 by default.

4.2.5 Dictionary features

- whether the target word and two neighbor words are in the dictionary; and
- whether the two neighbor words are in the dictionary.

As in most tasks about natural language processing, a dictionary is useful. We therefore design a run-time dictionary which is just a set of keywords in training data set.

4.3 Semi-supervised learning framework

Because the effort for labeling training data is extremely expensive, we propose the semi-supervised framework. We leverage the ideas of self training proposed by Liu *et al.* (2009) and k nearest neighbors (KNN). The intuition is that if an unlabeled sample is similar to a labeled sample in semantic space, the unlabeled sample is very probable to be successfully

inferred by the model which is learned from all the current labeled data. Then, the unlabeled sample is turned to a labeled one and can be added into the labeled dataset with model-inferred labels. A new model can be learned. The new thing proposed here is that we use the word embeddings learned by Word2Vec to calculate the similarity between two sentences. Sentence vector is denoted as:

$$VecSentence_i = \frac{1}{T}\sum_{t=1}^{T} VecWord_t \tag{5}$$

where *VecWord* is the word vector. Algorithm 1 is the details of the semi-supervised version of training process.

The time complexity of Algorithm 1 is $O(NM^2) + \frac{M}{c}O(\text{TrainCRF})$ where N and M are the sizes of labeled set and unlabeled set, respectively, and c is the number of unlabeled data which are selected to be inferred in each loop. The additional computing cost is rewarding, as human effort can be largely reduced, especially when N and M is not large.

Algorithm 1 Semi-supervised learning based on KNN self-training

INPUT: labeled data set $X_L = \{(x,y)\}$, unlabeled data set $X_U = \{x\}$, number of candidates c

OUTPUT: model Θ

repeat
  $\Theta = \text{TrainCRF}(X_L)$
  $X_{c-nearest} = \emptyset$
  for $i = 1:c$
    $x = \text{argmin}_{x \in X_U} \text{Cosine\_distance}(x, X_L)$
    $X_U = X_U - \{x\}$
    $X_{c-nearest} = X_{c-nearest} \cup \{x\}$
  $Y_{c-nearest} = \text{InferCRF}(X_{c-nearest}, \Theta)$
  $X_L = X_L \cup \{(X_{c-nearest}, Y_{c-nearest})\}$
until $X_U = \emptyset$
$\Theta = \text{TrainCRF}(X_L)$
return Θ

5. Keyword extraction on learning side

Due to the difficulty and complexity of labeling massive data of MOOC forums, we leverage unsupervised approaches to extract keywords from contents generated by learners. As the discussion forums are a kind of social media, we build a graph to model the relationship of post-reply. Then, a random walk algorithm is proposed to rank the importance of words. Finally, we regard the top words as keywords.

The intuition to build a graph model is that the more words are replied to, the more important the word is, and the more important word A is when word A replies to word B, the more important word B would be. This is similar to the algorithms for ranking Web pages, e.g. PageRank proposed by Brin and Page (1998).

5.1 Data model of massive open online course forum

To better model the importance of keywords, we design a heterogeneous network. Two kinds of entities are involved, learners and words. In the following, we introduce definition of the data model. Then, we explain the intuition for designing such a network.

Definition 1. Heterogeneous network with learners and words. Given all the learners' records of a MOOC forum, heterogeneous network $G = (V, E, W) = (V_L \cup V_D, E_L \cup E_D \cup E_{LD}, W_L \cup W_D \cup W_{LD})$ where $V_L = \{v_1^L, v_2^L, \ldots, v_{n_L}^L\}$ and $V_D = \{v_1^D, v_2^D, \ldots, v_{n_D}^D\}$ are sets of learners and words, respectively. E_L is the set of directed edges which mean the co-occurrence of two learners in the same thread. The leaner who posts later points to the other. E_D is the set of directed and bidirectional edges which mean the co-occurrence of two words in the same thread. Directed edges mean the two words belong to different posts. And the one which appears later points to the other. The bidirectional edges mean the two words belong to the same post. E_{LD} is the set of bidirectional edges which mean a learner's contents contain the word and in reverse a word appears in the learner's contents. W_L, W_D and W_{LD} are the sets of weight values which mean the times of co-occurrence of two entities on corresponding edges. Self co-occurrence is meaningless and is consistently ignored.

Figure 2 is a demo of the heterogeneous network with learners and words of a MOOC forum. By the way, we denote $G_L = (V_L, E_L, W_L)$ as a weighted directed graph of learners, $G_D = (V_D, E_D, W_D)$ as a weighted directed and bidirectional graph of words and $G_{LD} = (V_{LD}, E_{LD}, W_{LD})$ as the weighted bipartite graph of authorship between students and keywords. Denote $n_L = |V_L|$ and $n_D = |V_D|$ are the numbers of entities in V_L and V_D, respectively.

Such a heterogeneous network can embody the latent post-reply relationship between learners and words. In G_L and G_D, the more edges point to an entity, the more important it is. Moreover, if more important entities point to a specific entity, the entity would be more important. Similarly seeing from G_{LD}, the more edges point to a word, the more popular it is while also more important, if an important leaner points to it. All the weight values can capture the importance degree of relationship. The transmission of importance between learners (in G_L) can be transited to G_D. It is a process of mutual reinforcement between the two subnetworks.

5.2 Jump random walk algorithm

We design an algorithm for co-ranking learners and words, named Jump-Random-Walk (JRW) which simulates two random surfers jumping and walking between different types of entities. Figure 3 shows the framework of JRW algorithm. G_L is the subnetwork of learners and G_D is the subnetwork of words. G_{LD} is the subnetwork of authorship. β is the probability of walking along an edge within the homogeneous subnetwork. λ is the probability for jumping to the other subnetwork. $\lambda = 0$ means the two random surfers are independent to jump and walk within respective homogeneous subnetworks. We assume the probabilities of jump and walk are consistent.

Denote $\mathbf{l} \in \mathbb{R}^{n_L}$ and $\mathbf{d} \in \mathbb{R}^{n_D}$ are the ranking result vectors, also probability distributions, whose entries are corresponding to entities of V_L and V_D, subject to $||\mathbf{l}||_1 \leq 1$

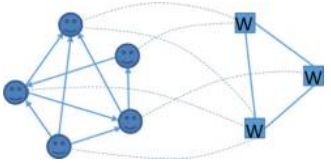

Figure 2.
Demo of the heterogeneous network with learners (left circles) and words (right rectangles) of a MOOC forum

and $||\mathbf{d}||_1 \leq 1$ due to existence of no-out-degree entities. Denote four transition matrixes of G_L, G_D, G_{LD} and G_{DL} as $L \in \mathbb{R}^{n_L \times n_L}$, $D \in \mathbb{R}^{n_D \times n_D}$, $LD \in \mathbb{R}^{n_{LD} \times n_{LD}}$ and $DL \in \mathbb{R}^{n_D \times n_L}$, respectively. Adding the probability of random jumping for avoiding trapped in small set of entities or no-out-degree entities, the iteration functions are:

$$\mathbf{l} = (1-\lambda)(\beta L\tilde{\mathbf{l}} + (1-\beta)\mathbf{e}_{n_L}/n_L) + \lambda LD\tilde{\mathbf{d}} \tag{6}$$

$$\mathbf{d} = (1-\lambda)(\beta D\tilde{\mathbf{d}} + (1-\beta)\mathbf{e}_{n_D}/n_D) + \lambda DL\tilde{\mathbf{l}} \tag{7}$$

where the former terms right the equal signs are iteration functions within a homogeneous subnetwork and the latter are across the two homogeneous subnetworks. λ is the probability of jumping to the subnetwork. β is the probability of walking along an edge within a homogeneous subnetwork. $\mathbf{e}_{n_L} \in \mathbb{R}^{n_L}$ and $\mathbf{e}_{n_D} \in \mathbb{R}^{n_D}$ are the vectors whose all entries are 1. The four transition matrixes are:

$$L_{i,j} = \frac{w^L_{i,j}}{\sum_i w^L_{i,j}} \quad where \sum_i w^L_{i,j} \neq 0, \tag{8}$$

$$D_{i,j} = \frac{w^D_{i,j}}{\sum_i w^D_{i,j}} \quad where \sum_i w^D_{i,j} \neq 0, \tag{9}$$

$$LD_{i,j} = \frac{w^{LD}_{i,j}}{\sum_i w^{LD}_{i,j}}, \tag{10}$$

$$DL_{i,j} = \frac{w^{DL}_{i,j}}{\sum_i w^{DL}_{i,j}} \quad where \sum_i w^{DL}_{i,j} \neq 0. \tag{11}$$

$w^L_{i,j}$ is the weight of the edge from V^L_i to V^L_j, $w^D_{i,j}$ is the weight of the edge between V^D_i and V^D_j, $w^{LD}_{i,j}$ is the weight of the edge between V^L_i and V^D_j and $w^{DL}_{i,j}$ is the weight of the edge between V^D_i and V^L_j. Actually, $w^{LD}_{i,j} = w^{DL}_{j,i}$. When $\sum_i w^L_{i,j} = 0$, it means the student V^L_j always posts the last in a thread. If $\sum_i w^D_{i,j} = 0$, it means the keyword V^D_j always has no peer in a thread. Actually, this situation almost never happens in our filtered words. $\sum_i w^{LD}_{i,j} = 0$ is also impossible, which means every word would have at least one author

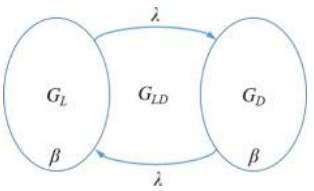

Figure 3.
The overview of JRW algorithm for co-ranking learners and words in the heterogeneous network

(learner). On the contrary, it does not make sure that every student would post at least one keyword because maybe someone's post has nothing valuable and contains no keyword. All the transition matrixes are non-negative. Algorithm 1 is the detail of JRW algorithm in the heterogeneous network.

Algorithm 2 JRW on G

INPUT $L, D, LD, DL, \beta, \lambda, \epsilon$

$\mathbf{l} \leftarrow \mathbf{e}/n_L$

$\mathbf{d} \leftarrow \mathbf{e}/n_D$

repeat

$\tilde{\mathbf{l}} \leftarrow \mathbf{l}$

$\tilde{\mathbf{d}} \leftarrow \mathbf{d}$

$\mathbf{l} = (1-\lambda)(\beta L\tilde{\mathbf{l}} + (1-\beta)\mathbf{e}_{n_L}/n_L) + \lambda LD\tilde{\mathbf{d}}$

$\mathbf{d} = (1-\lambda)(\beta D\tilde{\mathbf{d}} + (1-\beta)\mathbf{e}_{n_D}/n_D) + \lambda DL\tilde{\mathbf{l}}$

until $|\mathbf{d} - \tilde{\mathbf{d}}| \leq \epsilon$

return $\mathbf{l}, \mathbf{d}$

Finally, we can actually get two ranking lists of learners and words, but we only consider the ranking list of words within this paper.

6. Experiment

Again, based on the partition of two kinds of resources, as well as an extra experiment, this section consists of three parts.

6.1 On teaching side

In this subsection, we use teaching resources, i.e. subtitles, PPTs and questions, to evaluate the supervised learning model. We introduce several baselines to extract keywords for comparison:

- *Term frequency (TF)*: Words are ranked by their term frequency. If a word is a keyword, the instructor may say it repeatedly in lecture.
- *Bootstraping (BT)*: Instructors may have personal language styles to give talks. So we design the rule-based algorithm by giving several patterns containing keywords. This method is actually course- and instructor-dependent.
- *Stanford Chinese NER (S-NER)*: This is an exiting tool developed for NER, whose model is already trained, and we just use it to infer keywords in our educational data sets[4] proposed by Nadeau and Sekine (2007).
- *Terminology extraction (TermExtractor)*: This is an exiting tool for terminology extraction[5]. The well-trained model is also only used to infer keywords in our data sets.
- *Supervised keyword-CRF (SK-CRF)*: This is a method of supervised learning based CRFs with all features as defined before.
- *Semi-supervised keyword-CRF (SSK-CRF)*: This is the semi-supervised version for keyword extraction. The parameter of c, number of candidates, is empirically set as 20.

We adopt three metrics, precision, recall and F1-value, to measure the results.

6.1.1 Results and analysis. Table IV shows the comparison of performance between baselines. We use 30 per cent data of subtitles as training data for SK-CRF and SSK-CRF,

and the rest are for evaluation. Especially for SSK-CRF, half of the training data are unlabeled. The statistic-based methods (TF@500 and TF@1000) are unreliable because many stopwords may degrade the performance. The rule-base method (BT) is highly dependent on human experience, and the low precision means plenty of subsequent work for filtering the outputs is required. On the other hand, Stanford Chinese NER and TermExtractor do not perform well maybe because of two reasons, namely, named entity and terminology are actually different from the keywords in our data, and the models are not learned from our data set. The semi-supervised CRF is comparable to the supervised version.

Figure 4 manifests that the semi-supervised learning would be comparable to the supervised version, especially when less than 20 per cent data are used for training. Half of training data is identically regarded as unlabeled by SSK-CRF. Note that the amount of labeled data when using 10 per cent training data by SK-CRF is equivalent to that of using 20 per cent training data by SSK-CRF, but SSK-CRF performs better than SK-CRF. This result means the semi-supervised framework can obtain satisfactory performance by only labeling a handful of data.

Now, we evaluate the different model abilities among various MOOC textual content. As shown in Table V, the items in row are training data set, while those in column are testing data set. This table can explain some common situations of educational settings. Subtitles can cover almost all the keywords. They are ideal to be regarded as the training data. PPTs is also decent to be as training data seeing from the precisions, but the recalls are low. Maybe due to usually in PDF format, PPTs may cause incomplete sentences when being converted to text. Questions could lead to lower recalls than PPTs because not all keywords are present in questions as shown in Table I. In summary, different kinds of MOOC textual content have different model ability, so they should be separately considered.

6.1.2 Feature contribution. We analyze how the different kinds of features contribute to the model. The result is shown in Table VI. Dictionary feature has a predominant influence on the final results, and structure feature is the second important. Other features are also contributive, but the difference is small. Even so, every kind of features contribute to the model positively.

6.2 On learning side

After building a heterogenous network for each course, Table VII shows the parameters of the network per course.

The important of keywords ranked at top is hard to evaluate. Table VIII lists the top ten high-frequency words and top ten keywords ranked by JRW, respectively. We can see the

Baselines	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
S-NER	0.131	0.080	0.099
TermExtractor	0.202	0.107	0.140
SK-CRF	*0.914*	*0.897*	*0.905*
SSK-CRF	0.889	0.825	0.856

Table IV.
Performance of baselines

Notes: SK-CRF and SSK-CRF use 30% data of subtitles for training; half of the training data as unlabeled for SSK-CRF; the italic data mean they are the best results among all the baselines

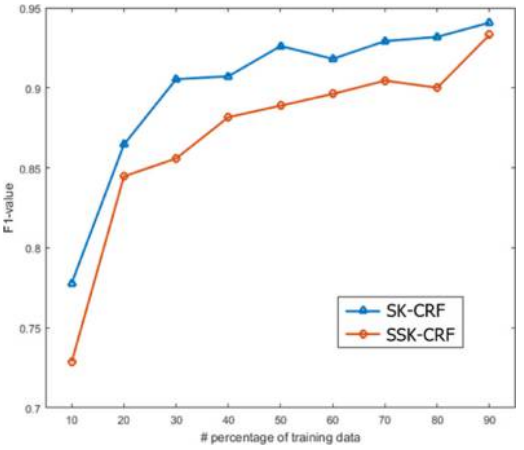

Figure 4.
Performance between supervised and semi-supervised models

	Subtitles			PPTs			Questions		
Data set	P	R	F1	P	R	F1	P	R	F1
Subtitles	–	–	–	0.816	0.838	0.827	0.860	0.800	0.829
PPTs	0.868	0.764	0.813	–	–	–	0.857	0.685	0.761
Questions	0.846	0.349	0.494	0.722	0.360	0.480	–	–	–

Note: The rows are training data and the columns are testing data

Table V.
Mutual learning between various content

Methods	Precision	Recall	F1
All	0.780	0.775	0.777
Without text style feature	0.768	0.776	0.772
Without structure feature	0.722	0.683	0.702
Without context feature	0.757	0.753	0.755
Without semantic feature	0.772	0.757	0.764
Without dictionary feature	0.689	0.235	0.350

Note: 10% of data are used for training by SSK-CRF

Table VI.
Efficacy of features

two kinds are highly overlapped, but the order is slightly different. The bold keywords are related to course content, and the italic ones are mainly about the course quiz, assignment, video and other course stuff.

Table IX shows the statistics of the top three "important posts", meaning that the posts contain the top 20 keywords. The more frequency of keywords they contain, the higher they rank. From Table IX, we can first find that the content lengths are mostly long, which is obvious by our definition of "important posts". From the dimension of vote, we cannot find some insight of the numbers. Author rank means the ranking of the post author in the ranking list of important learners. We find they are truly the "important learners" of each

course. Also, the important posts are mostly at the top of a thread, seeing from position in thread. It means the initial authors in a thread are inclined to express important information. By the way, the lengths of a thread, i.e. # Post in Thread, are significantly correlated to the important posts. Some empirical conclusions can be summarized as below:

6.3 Extra experiment

Considering the available data in our hand, although we do not have labels of forum data, we can learn a classifier from labeled teaching resources and conduct a task of identifying the need of concept comprehension on forum contents. This task can be regarded as a binary classification of forum threads, that is to identify whether a thread is about concept comprehension. So if the question contains keywords of the course, it is much likely to ask for the explanation of some concepts. The result is post-evaluated which means: to each thread, if the score is marked as "1", two situations are included as the following:

- if no concept is identified and this thread is not about need of concept comprehension; and
- if at least one concept is identified and the definition of identified concepts can answer the question.

Other situations are marked as "0".

We use 30 per cent of subtitles to learn a classifier by the semi-supervised method. Only threads title and the initial post are involved in this experiment, instead of all the posts. Table X exhibits the result. The accuracy is not bad. The relatively high recall is meaningful because this can accurately remind instructors which threads to intervene. Moreover, this method not only can identify whether a thread is about concept comprehension but also can identify which concept needs to be explained.

7. Application with keywords for massive open online course

After keywords are extracted from teaching resources of one course, we exhibit two intelligent applications with keywords in the MOOC settings: generation of concept map and generation of learning path. We conduct the applications on the course of people and network.

Table VII.
Summary of the constructed heterogeneous network per course

Course	n_L	$\|E_L\|$	$\|E_L\|/n_L^2$	n_D	$\|E_D\|$	$\|E_D\|/n_D^2$	$\|E_{LD}\|$	$\|E_{LD}\|/(n_L+n_D)^2$
peopleandnetworks-001	321	3,287	0.032	1,193	104,821	0.074	4,814	0.002
arthistory-001	540	17,022	0.058	3,376	1,019,289	0.089	14,195	0.001
dsalgo-001	295	1,876	0.022	1,152	124,118	0.094	5,009	0.002
pkuic-001	768	19,801	0.034	2,302	302,989	0.057	14,599	0.002
aoo-001	175	1,963	0.064	783	73,208	0.119	2,597	0.003
bdsalgo-001	225	2,369	0.047	781	23,540	0.039	3,133	0.003
criminallaw-001	219	2,971	0.062	1,224	123,737	0.083	4,577	0.002
pkupop-001	628	12,883	0.033	1,748	88,035	0.029	13,807	0.002
chemistry-001	130	886	0.052	1,055	111,026	0.100	2,685	0.002
chemistry-002	125	2,341	0.150	964	61,425	0.066	2,574	0.002
pkubioinfo-001	594	22,275	0.063	686	46,768	0.099	1,946	0.001
pkubioinfo-002	189	1746	0.049	380	16662	0.115	784	0.002

Course	Top 10 high-frequency words	Top 10 JRW ranked words
peopleand networks-001	**relationship**, **people**, **node**, *homework*, *question*, *teacher*, *course*, *problem*, *video*, *answer*	**people**, **relationship**, **node**, *homework*, *teacher*, *question*, *problem*, *video*, *course*, **network**
arthistory-001	**art**, **art history**, people, *question*, *teacher*, *course*, *class*, *classmate*, *homework*, **artistic work**	**art**, **art history**, *teacher*,people, *course*, *question*, *class*, *classmate*, **artistic work**
dsalgo-001	*teacher*, *course*, *question*, *class*, **data**, *problem*, *homework*, **code**, *video*, **algorithm**	*teacher*, *course*, *class*, *question*, **data**, people, **code**, *homework*, **structure**, *problem*
pkuic-001	*question*, **program**, *teacher*, *homework*, **code**, *course*, **array**, *problem*, *mistake*, result	*question*, **program**, *teacher*, *course*, **computer**, *homework*, **code**,people, **array**, *video*
aoo-001	**object**, **model**, **method**, **software**, *question*, **code**, **system**, *teacher*, **graph**, *video*	**object**, **model**, **method**, **software**, *question*, **code**, **system**, *video*, **graph**, *teacher*
bdsalgo-001	*question*, **data**, *problem*, **code**, **pointer**,occasion, *teacher*, **function**, **table**, **algorithm**	*question*, **data**, *problem*, **table**, *teacher*, occasion, **code**, **algorithm**, **function**, **array**
criminallaw-001	**penal law**, **law**, **behavior**, people, **judicature**, **guilt**, *question*,classmate, **country**, *teacher*	**law**, **penal law**, people, **behavior**, **judicature**, **guilt**, classmate, *question*, **society**, **Part B**
pkupop-001	**function**, *question*, *problem*, *homework*, **program**, **code**, *teacher*, **object**,result, **array**	**function**, *question*, **code**, **program**, **object**, *homework*, *problem*, *teacher*, **array**, *time*
chemistry-001	**chemistry**, *teacher*, *course*, *class*, *question*, *problem*, *video*, **door**,people, *answer*	**chemistry**, *teacher*, *course*, *class*, *question*, *problem*, *video*, **door**,people, *answer*
chemistry-002	**chemistry**, **electron**, *teacher*, *course*, **chemicalbond**, *question*, **atom**, **orbit**, *class*, **radius**	**chemistry**, *teacher*, **electron**, *course*, university, *class*, student, *question*, **atom**, **radius**
pkubioinfo-001	**sequence**, *teacher*, **biology**, *course*, *question*, *class*, **informatics**, *video*, **information**, **data**	**biology**, *teacher*, **sequence**, **informatics**, *question*, *course*, *class*, *data*, *video*, **information**
pkubioinfo-002	*course*, *question*, *teacher*, **biology**, **sequence**, *video*, **door**, classmate, *certificate*, **genome**	*course*, *question*, **sequence**, *teacher*, *classmate*, **biology**, **door**, content, **species**, *video*

Note: The bold words are course content-related and the italic words are course resource-related

Table VIII. Top 10 high-frequency words and top 10 keywords per course

7.1 Concept map

For generating a concept map, we need to define the meaning of nodes, edges and their weights. The nodes are concepts, and the edges are defined as the semantic similarity which is general for every course. We call the new concept map as SCM. Based on our observation, there are two kinds of node weight definitions, i.e. term frequency (TF) and TF-IDF.

It can be observed that the more frequent a concept appears, the more fundamental it is. For example, the top ten concepts, *Node, Network, Reward, Probability, Graph, Game, Edge, Tactic, Hypothesis and Price*, are all the fundamental knowledge points of the course. So the metric of TF can capture the feature of *fundamentality*. The formal definition is:

$$NodeWeight_i^{(F)} = \sum_k f_{ki} \tag{12}$$

Course	Rank	Content length	Vote	Author rank	Position in thread	# post in thread
peopleandnetworks-001	1	1,515	3	5	3	5
	2	716	1	10	7	8
	3	664	1	2	1	5
arthistory-001	1	5,146	2	14	4	11
	2	4,670	3	11	1	5
	3	2,380	6	1	1	4
dsalgo-001	1	757	1	14	1	9
	2	757	0	6	9	11
	3	682	0	8	1	3
pkuic-001	1	1,790	0	39	1	4
	2	1,418	0	23	4	5
	3	619	0	4	2	3
aoo-001	1	1,991	3	1	1	2
	2	1,468	0	1	1	2
	3	1,394	5	1	1	6
bdsalgo-001	1	3,501	2	1	2	3
	2	855	0	1	2	5
	3	1,257	2	1	1	10
criminallaw-001	1	1,683	0	3	1	2
	2	1,967	2	2	27	28
	3	1,480	1	1	12	14
pkupop-001	1	368	0	1	2	4
	2	457	0	22	1	2
	3	579	0	8	1	4
chemistry-001	1	1,664	1	2	1	9
	2	359	0	2	3	9
	3	1,539	0	5	3	4
chemistry-002	1	424	16	2	1	5
	2	472	2	1	3	5
	3	306	1	4	5	6
pkubioinfo-001	1	769	-2	3	1	7
	2	1,322	3	3	1	7
	3	1,956	2	2	1	6
pkubioinfo-002	1	309	0	1	2	3
	2	422	0	8	1	4
	3	515	0	1	3	5

Table IX.
Summary of top three important posts which contain the top 20 keywords per course

Table X.
Result of identifying threads about concept comprehension by SSK-CRF

Accuracy	Precision	Recall	F1
0.822	0.523	0.784	0.627

where f_{ki} is the times of the ith concept existing in the kth document; a document corresponds to a video clip in MOOCs.

However, on the other hand, low-frequency concepts often are the important knowledge points. So TF-IDF is ideal to measure the *importance* of a concept. For example, the top ten important concepts are *PageRank, SignalSequence,*

PreEstimatedPrice, MixedStrategyEquilibrium, TradingRight, HinderAggregation, Cluster, ConformityBehavior, NashBargaining and Popularity. The formal definition of TF-IDF weights is:

$$NodeWeight_i^{(I)} = \frac{1}{n_i}\sum_k \log(f_{ki}+1)\cdot\log(N/n_i) \tag{13}$$

where N is the number of video clips and n_i is the times of video clips in which the ith concept appears.

Considering the word embeddings learned by Word2Vec have the characteristic that semantically similar words are close in the embedding space, so we use the similarity as the weights of edges between two concepts. For example, the most semantically similar concepts around *Network* are: *NetworkAnalysis, SocialNetwork, ResidualNetwork, ComplexNetwork, NetworkSwitch, ComplexNetworkAnalysis, SocialNetwork, TraficNetwork, SocialNetworkAnalysis and NetworkSwitchExperiment*.

Figure 5 shows the demos of SCM of the course of People and Network for fundamentality and importance, respectively. We find the map can visually reveal the degree of semantic relationships between concepts. This is beneficial for learners to build a "concept map" in their brain and remember concepts easily. We use the tool of Gephi to draw the maps.

7.2 Learning path

Based on the SCM, learners can also learn the course in line with their own pace. Here, we propose an algorithm (Algorithm 3) to generate a primary learning path according to the definition of SCM. Then, the learning path can be revised by both the instructors and learners as required.

The basic idea of the algorithm is simple. Every time a current concept is taken, then a candidate set of k the most semantically similar neighbors of the concept are selected. Among the candidate set, TF or TF-IDF of concepts is calculated. Then, the top concept is selected as a node in the path and as the next current concept. The algorithm can start from any concept. Note that the concepts which are selected in the candidate set should appear later than the current concept along with the course because learners may be confused to learn concepts through the path which does not conform to the instructor's design.

By taking the concept *Node* as the starting point and setting $k = 10$, the first ten concepts in the learning path with metric of TF are: *Node* → *Edge* → *Element* → *Set* → *Alternative* → *Vote* → *MajorityVoting* → *MajorityVotingRule* → *IndividualRanking* → *GroupRanking*.

By taking the concept with the highest TF-IDF as the starting point, the first ten concepts are: *PageRank* → *PageRankAlgorithm* → *SmallWorld* → *Balance* → *NashBalance* → *StructuralBalance* → *EquilibriumTheorem* → *MixedStrategyEquilibrium* → *NashBargaining* → *NashBargainingSolution*. We can see these concepts are all important along the course:

Algorithm 3 Generation of learning path

INPUT: $SCM = \{C, R\}$, starting concept c_i, number of candidates k

OUTPUT: learning path $p_i = \{n_1, n_2, \ldots, n_{|C|}\}$

1: $j = 1$

2: $n_j = c_i$

3: $p_i = \{n_j\}$

4: $C' = C - \{n_j\}$

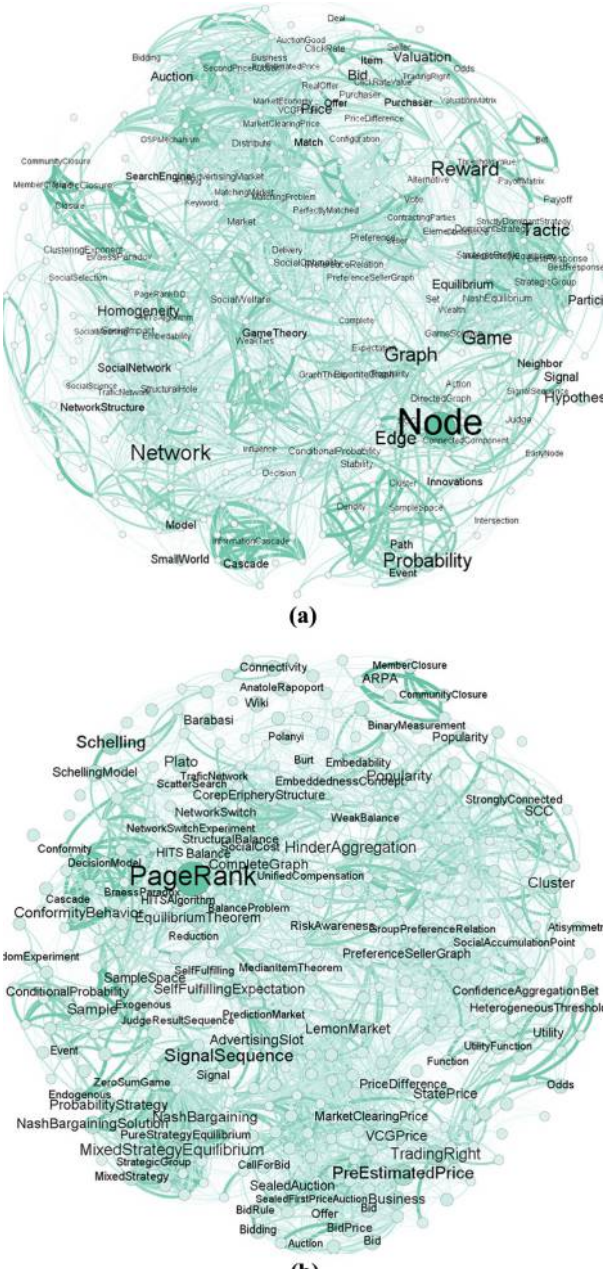

Figure 5.
Two kinds of SCM based on different concepts metrics

Notes: (a) For fundamentality; (b) For importance

```
repeat
T = the k most semantically similar and later appeared concepts to n_j in C'
j = j + 1
n_j = the concept selected by some metric (TF or TF-IDF) in T
p_i = p_i ∪ {n_j}
C' = C' − {n_j}
until C' = ∅
return p_i
```

Admittedly, the two demo learning paths are very primitive. They cannot support personalized learning and adaptive learning yet. However, by analyzing the learners' behavior and log of homework, the learning paths can be more intelligent. We leave this for the future work.

8. Conclusion

Along with the development of MOOCs, massive online educational resources are unprecedentedly produced from crowd. Instructors can provide videos, subtitles, lecture notes, questions, etc., while learners can generate forum content, Wiki, log of homework, etc. How to process these data from unstructured to structured is a challenging problem. In this paper, we explore the task of keyword extraction on MOOC resources.

Keyword extraction can benefit a lot of subsequential applications. First, it is a kind of annotation for MOOC resources. The annotation can be used for studying machine learning methods for MOOC-related natural language processing tasks, such as information extraction, information retrieval and question answering. Second, keyword extraction can pick up domain-specific or cross-domain knowledge points from complex text. This result can be further processed to build knowledge graph or concept map. With the graph (or the map), instructors can better organize the course, and learners can plan their own learning paths more easily. Then by collecting the feedback from learners, the whole teaching and learning process can be a virtuous cycle. Thus finally, crowd intelligence can lead to intelligent education.

Back to the task of this paper, we are faced with two challenges: MOOCs are cross-domain, labeling training data is extremely expensive. So we propose a flexible framework based on semi-supervised machine learning with domain-agnostic features. Experiments demonstrate the efficacy of our framework. Using a very little labeled data can achieve decent performance. We find that various kinds of MOOC content, e.g. subtitles and PPTs, have different modeling ability for keyword extraction. So they should be separately treated in future work. Our framework also can be applied to the task of concept identification on MOOC forum content. Moreover, unsupervised method based on graph model is proposed by modeling MOOC forum to a heterogeneous network. Although the top keywords in MOOC forums are not as the same as those keywords extracted from teaching resources, they can indicate the concerned topics which are discussed in forums. At least instructors can get feedback from the information.

In the future, methods of transfer learning and deep learning may be better for extracting cross-domain keywords. External resources of knowledge, e.g. Wikipedia, may be helpful. The relationship between keywords is deserved to be paid more attention for building a domain-specific or even cross-domain concept map.

Notes

1. Stanford Chinese word segment: http://nlp.stanford.edu/software/segmenter.shtml
2. Stanford Log-linear POS Tagger: http://nlp.stanford.edu/software/tagger.shtml
3. Word2Vec: https://code.google.com/p/word2vec/
4. Stanford Chinese NER: http://nlp.stanford.edu/software/CRF-NER.shtml
5. Terminology extraction by translated labs: http://labs.translated.net/terminology-extraction/

References

Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. (2014), "Engaging with massive online courses", *WWW'14 Proceedings of the 23rd International Conference on World Wide Web*, pp. 687-698.

Bin, Y. and Shichao, C. (2011), "Term extraction method based on mutual information with threshold interval", *Applied Informatics and Communication*, Vol. 227 No. 4, pp. 186-194.

Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D. and Seaton, D.T. (2013), "Studying learning in the worldwide classroom: Research into edX's first MOOC", *Research & Practice in Assessment*, Vol. 8 No. 1, pp. 13-25.

Brin, S. and Page, L. (1998). "The anatomy of a large-scale hypertextual web search engine", *Proceedings of the 7th International Conference on World Wide Web, WWW '1998*, Elsevier Science Publishers, pp. 107-117.

Chang, P.-C., Galley, M. and Manning, C. (2008). "Optimizing Chinese word segmentation for machine translation performance", *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 224-232.

Chen, N.-S., Kinshuk, Wei, C.-W. and Chen, H.-J. (2008), "Mining e-learning domain concept map from academic articles", *Computers & Education*, Vol. 50 No. 3, pp. 1009-1021.

Chu, H.-C., Hwang, G.-J., Wu, P.-H. and Chen, J.-M. (2007). "A computer-assisted collaborative approach for e-training course design", *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies*, IEEE, pp. 36-40.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T. and Zhang, S.S.W. (2014). "Knowledge vault: a web-scale approach to probabilistic knowledge fusion", *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 601-610.

Finkel, J.R., Grenager, T. and Manning, C. (2005), "Incorporating non-local information into information extraction systems by gibbs sampling", *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 363-370.

Frantzi, K., Ananiadou, S. and Mima, H. (2000), "Automatic recognition of multi-word terms: the C-value/NC-value method", *International Journal on Digital Libraries*, Vol. 3 No. 2, pp. 115-130.

Ho, A.D., Reich, J., Nesterko, S.O., Seaton, D.T., Mullaney, T., Waldo, J. and Chuang, I. (2013), "HarvardX and MITx: the first year of open online courses", HarvardX and MITx Working Paper No. 1, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263

Huang, C.-J., Tsai, P.-H., Hsu, C.-L. and Pan, R.-C. (2006), "Exploring cognitive difference in instructional outcomes using text mining technology", *Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 2116-2120.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J. and Sanders, M. (2014). "Superposter behavior in MOOC forums", *Proceedings of the first ACM Conference on Learning @ Scale Conference, Atlanta, GA, 4-5 March*, ACM, New York, NY, pp. 117-126.

Huang, X., Yang, K. and Lawrence, V. (2015), "Classification-based approach to concept map generation in adaptive learning", *Proceedings of the IEEE 15th International Conference on Advanced Learning Technologies*, IEEE, pp. 19-23.

Jiang, Z., Zhang, Y., Liu, C. and Li, X. (2015), "Influence analysis by heterogeneous network in MOOC forums: what can we discover?", paper presented at the International Conference on Educational Data Mining, Madrid, pp. 242-249.

Justesona, J.S. and Katza, S.M. (1995), "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, Vol. 1 No. 1, pp. 9-27.

Kizilcec, R.F., Piech, C. and Schneider, E. (2013). "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses", *Proceeding of LAK 2013*, ACM Press, pp. 170-179.

Lafferty, J.D., McCallum, A. and Pereira, F.C.N. (2001), "Conditional random fields: probabilistic models for segmenting and labeling sequence data", *Proceedings of the 18th International Conference on Machine Learning 2001 ICML'01*, pp. 282-289.

Lau, R.Y., Song, D., Li, Y., Cheung, T.C. and Hao, J.-X. (2009), "Toward a fuzzy domain ontology extraction method for adaptive e-learning", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 21 No. 6, pp. 800-813.

Lee, C.-H., Lee, G.-G. and Leu, Y. (2009), "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 1675-1684.

Liu, A., Jun, G. and Ghosh, J. (2009), "A self-training approach to cost sensitive uncertainty sampling", *Machine Learning*, Vol. 76 Nos 2/3, pp. 257-270.

Marian, S. and Maria, B. (2009), "Automatic concept relationships discovery for an adaptive e-course", *Proceedings of the 2nd International Conference on Educational Data Mining*, IEDMS, pp. 171-178.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality", *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Curran Associates, pp. 3111-3119.

Nadeau, D. and Sekine, S. (2007), "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, Vol. 30 No. 1, pp. 3-26.

Nickel, M. Murphy, K. Tresp, V. and Gabrilovich, E. (2015). "A review of relational machine learning for knowledge graphs", available at: http://arxiv.org/abs/1503.00759v3

Nojiri, S. and Manning, C.D. (2015). "Software document terminology recognition", *AAAI Spring Symposium*, pp. 49-54.

Novak, J.D. and Cañas, A.J. (2006), "The theory underlying concept maps and how to construct and use them", *Technical Report IHMC CmapTools 2006-01 Rev 2008-01*, available at: http://cmap.ihmc.us/docs/theory-of-concept-maps.php

Qasim, I., Jeong, J.-W., Heu, J.-U. and Lee, D.-H. (2013), "Concept map construction from text documents using affinity propagation", *Journal of Information Science*, Vol. 39 No. 6, pp. 719-736.

Qin, Y., Zheng, D., Zhao, T. and Zhang, M. (2013), "Chinese terminology extraction using em-based transfer learning method", *14th International Conference, CICLing 2013*, pp. 139-152.

Ratinov, L. and Roth, D. (2009), "Design challenges and misconceptions in named entity recognition", *Proceedings of the 13th Conference on Computational Natural Language Learning*, ACL, Boulder, pp. 147-155.

Robertson, S., Zaragoza, H. and Taylor, M. (2004), "Simple bm25 extension to multiple weighted fields", *CIKM'04 Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 42-49.

Romero, C. and Ventura, S. (2010), "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40 No. 6, pp. 601-618.

Sonawane, S.S. and Kulkarni, P.A. (2014), "Graph based representation and analysis of text document: a survey of techniques", *International Journal of Computer Applications*, Vol. 96 No. 19, pp. 1-8.

Sutton, C. and McCallum, A. (2011), "An introduction to conditional random fields", *Machine Learning*, Vol. 4 No. 4, pp. 267-373.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network", *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 252-259.

Tsenga, S.-S., Sue, P.-C., Su, J.-M., Weng, J.-F. and Tsai, W.-N. (2007), "A new approach for constructing the concept map", *Computers & Education*, Vol. 49 No. 3, pp. 691-707.

Wang, X., Yang, D., Wen, M., Koedinger, K. and Rosé, C.P. (2015). "Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains", *The 8th International Conference on Educational Data Mining EDM'15, Madrid*, pp. 226-233.

Wen, M., Yang, D. and Rosé, C. (2014), "Sentiment analysis in MOOC discussion forums: what does it tell us?", *EDM'14*, pp. 130-137.

About the authors

Zhuoxuan Jiang is currently pursuing a PhD degree at the School of Electronics Engineering and Computer Science, Peking University, China. His current research interests include machine learning, MOOC data mining and online education. Zhuoxuan Jiang is the corresponding author and can be contacted at: jzhx@pku.edu.cn

Chunyan Miao is a Professor with the School of Computer Science and Engineering (SCSE) at Nanyang Technological University (NTU), Singapore. She is the Director of the NTU-UBC Joint Research Centre of Excellence in Active Living for the Elderly (LILY). Prior to joining NTU, she was an Instructor and Post-Doctoral Fellow with the School of Computing, Simon Fraser University, Canada. Her research focuses on studying the cognitive and social characteristics of intelligent agents in multi-agent and distributed AI/CI systems, such as trust, emotions, incentives, motivated learning, ecological and organizational behavior. She has worked on new disruptive Artificial intelligence (AI) approaches and theories that synergize human intelligence, artificial intelligence and behavior data analytics (AI powered by humans). Her current research interests include human-agent interaction, cognitive agents, human computation and serious games.

Xiaoming Li is a Professor in Computer Science and Technology and the Director of Institute of Network Computing and Information Systems (NCIS) at Peking University, China. His current research interest is in Web data mining and online education. He led the effort of developing a Chinese search engine (Tianwang) since 1999 and is the founder of the Chinese Web archive (Web InfoMall). Related papers have been published in *WWW*, *KDD*, *Computer Networks*, *Journal of Software and Systems*, *Journal of Web Engineering*, etc. Under his direction, the Institute is focused on the areas of search engine and Web mining, peer-to-peer computing, distributed systems, mobile computing, high productivity computing and database systems. He serves on the editorial boards of several journals, including *Concurrency and Computation* (John Wiley), *Journal of Web Engineering* (Rinton), etc. He is a Senior Member of IEEE and a Member of Eta Kappa Nu. He also serves as a Vice President for Chinese Computer Federation and is chairing the Advisory Subcommittee for Undergraduate Computing Education in China.