Wu, P., Yu, S., & Wang, D. (2018). Using a Learner-Topic Model for Mining Learner Interests in Open Learning Environments. *Educational Technology & Society, 21* (2), 192–204.

Using a Learner-Topic Model for Mining Learner Interests in Open Learning Environments

Pengfei Wu[1,2,3], Shengquan Yu [1,2*] and Dan Wang[1,2]

[1]School of Educational Technology, Faculty of Education, Beijing Normal University, China // [2]Advanced Innovation Center for Future Education, Beijing Normal University, China // [3]School of Education, Shijiazhuang University, China // wupengfei_2000@163.com // yusq@bnu.edu.cn // woxinwing@163.com

[*]Corresponding author

ABSTRACT

The present study uses a text data mining approach to automatically discover learner interests in open learning environments. We propose a method to construct learner interests automatically from the combination of learner generated content and their dynamic interactions with other learning resources. We develop a learner-topic model to discover not only the learner's knowledge interests (interest in generating content), but also the learner's collection interests (interest in collecting content generated by others). Then we combine the extracted knowledge interests and collection interests to yield a set of interest words for each learner. Experiments using a dataset from the Learning Cell Knowledge Community demonstrate that this method is able to discover learners' interests effectively. In addition, we find that knowledge interests and collection interests are related and consistent in their subject matter. We further show that learner interest words discovered by the learner-topic model method include learner self-defined interest tags, but reflect a broader range of interests.

Keywords

Open learning environments, Learner interest model, Educational data mining, Learning cell knowledge community, Interaction behaviors, Resource organization

Introduction

Web 2.0 not only brings new ideas and forms of communication for human beings, but also provides opportunities for social interactions focused on knowledge generation, collaborative learning and sharing, and the exchange of knowledge, experience and resources (Kuswara & Richards, 2011). Users' Web 2.0 interactive content and behavior can offer insight into their learning interests. Detecting user interests based on user-related content and behavior is an important task for enhancing value from online services.

With increasing educational applications of Web 2.0, open learning environments have become an important platform and space for multiple learners to collaboratively create, share and acquire knowledge (Wu & Yu, 2015). Open learning environments involve many different learners, each with different interests that change dynamically. To annotate and manage learner interests, open learning environments usually provide learners with the means to self-define their interests by tagging topics (an example is shown in Figure 1). However, it is difficult for learners to describe their interests in detail. Furthermore, learners will not necessarily update their interest tags as their interests change. In addition, many learners do not actively tag their interests. Thus, it is worth exploring how to discover learner interests automatically in open learning environments.

Learner interests play an important role in Web-based learning environments and are positively related to learning outcomes (Wigfield & Cambria, 2010). Learner interests are reflected in learner generated content and dynamic interactions between the learner and the web-based resources. Essentially, learners express their knowledge interests and knowledge requirements through their online behavior. In open learning environments, the massive amount of data, on both learner generated content and their interactions with online resources, provides opportunities to detect learner interests automatically. At the same time, use of this data enables open learning environments to improve their educational services by adaptively discovering learner needs and automatically recommending relevant resources.

Educational data mining can support the construction of smart learning environments. Mining learner generated content and interaction behaviors to construct a learner interest model is important for offering adaptive learning services in open learning environments. In this paper, we use a text data mining approach to explore the problem of discovering learner interests automatically in open learning environments. Text mining techniques can be employed to mine learner interests from content that they generate and from the text of their online interactions. Topic modelling has attracted recent research attention and been applied in the fields of educational mining and text content analysis. In this study, we present a method to automatically construct a learner interest model in open learning environments using a learner-topic model (LTM). For each learner, interests are composed of two

ISSN 1436-4522 (online) and 1176-3647 (print). This article of the Journal of Educational Technology & Society is available under Creative Commons CC-BY-ND-NC 3.0 license (https://creativecommons.org/licenses/by-nc-nd/3.0/). For further queries, please contact Journal Editors at ets.editors@gmail.com

types: knowledge interests and collection interests. Knowledge interest refers to the learner's interest in creating content, while collection interest refers to the learner's interest in collecting content generated by others. We develop an LTM to discover not only learner knowledge interests, but also collection interests.

Figure 1. An example of user self-defined interest tagging

This paper contributes to a better understanding of learner interests in open learning environments. Our study examines different types of learning content data (a) to discover learner knowledge interests and collection interests, (b) to compare the two sets of mined interest data, and (c) to explore the characteristics of and differences between the mined interest data and user self-defined interest tags. Our study aims to answer the following research questions:

Question 1. What differences exist between a learner's knowledge interests and collection interests?
Question 2. What differences exist between the mined interest words and self-defined interest tags?
Question 3. Is the learner-topic model effective for mining learner interests?

Literature review

Open learning environments

With the spread of open resources, open learning environments, such as open knowledge communities and massive open online courses, have become increasingly popular (Yang, Qiu, Yu, & Tahir, 2014). Open learning environments provide learners with opportunities for individual knowledge construction, resource annotation, social collaboration, participation and communication (Wu & Yu, 2015). In open learning environments, educational big data are generated from learners' various interactions and learner generated content. Mining this educational big data in open learning environments is important for offering learners better learning services.

User interest

In Web 2.0 environments, user models have attracted increasing attention. User modeling represents several aspects of users such as their knowledge of a subject, their interests, their goals, their backgrounds and other individual traits (Brusilovsky, 2007). Representation of user interests in user models is becoming increasingly popular. Access to user interests makes it easier to satisfy users' personal needs in recommendation systems,

question answering systems (Ni, Lu, Quan, Liu, & Hua, 2012) and information retrieval and filtering systems (Liu, Chen, Xiong, Ding, & Chen, 2012).

In open learning environments, learners are an important category of users. A learner interest model is a key component of adaptive hypermedia and adaptive educational systems that track learner behaviors and make inferences according to learner interests. Learning behavior actions could offer insight into learners' topic interest profiles in open learning environments (Peng, Liu, Liu, Gan, & Sun, 2016; Zhao, Cheng, Hong, & Chi, 2015). Knowledge interest and collection interest are reflected in learner generated contents and dynamic learning behavior actions. The smart and personalized educational systems research community has conducted substantial research into the construction of models able to represent user interests (e.g., Zhang, Zhu, Zhao, Gu & Ting, 2008; Gong, 2012; Li, Sagl, Mburu, & Fan, 2016; You, Bhatia, & Luo, 2016). Learner modeling is the process of collecting and computing learning relevant data in educational systems. In the educational environment, analysis of "big data" offers opportunities for constructing such learner interest models.

Text mining in eLearning

Mining of educational data can play a supportive role in eLearning. Hwang (2005) proposed a data mining approach to assist teachers in providing information tailored to guide individual students in their learning efforts. In recent years, text mining has become popular in educational data mining applications. Text mining aims to find and extract useful, latent or interesting patterns and models from unstructured text documents. Text mining can be used to identify, extract, integrate, and exploit knowledge for eLearning efficiently and effectively (He, 2013).In recent years, a number of studies have used text mining techniques to analyze learning-related data. For example, text mining techniques were used to automatically analyze data from online questions, interactions and chat messages and predict final student grades (Abdous, He, & Yen, 2012; He, 2013). Other studies analyzed student attitudes towards learning through mining lecture data, and explored correlations between learning attitudes and learning achievement through analyzing the texts of student answers to a questionnaire (Minami & Ohura, 2013; Minami & Ohura, 2015). Goda and Mine (2011) estimated learning situations based on text mining of student comments.

Topic modeling is a text mining method for estimating topics in documents and clustering documents based on latent topics. Sekiya, Matsuda and Yamaguchi (2010) used latent topic modeling to analyze course syllabi. Sorour, Goda and Mine (2015) used two types of machine learning techniques to learn the relationships between comment data analyzed by Latent Semantic Analysis (LSA) and final student grades. Sorour, Goda and Mine (2017) applied Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (pLSA) to predict student grades in each lesson through mining student comment data.

There have also been several studies using topic modeling focused on modeling users and mining user interests in Web 2.0 environments such as Microblogs and Twitter. Pennacchiotti and Popescu (2011) applied topic modeling techniques to classify users by considering user profiles, behaviors, content and social network features. Ishii, Mizoguchi, Kimita and Shimomura (2015) proposed a topic model for clustering learners based on educational counseling content. Michelson and Macskassy (2010) discovered topics of interest for Twitter users based on their posts. Xu, Ru, Xiang and Yang (2011) proposed a method for discovering an author's interest on Twitter with a twitter-user model. Collectively, these efforts demonstrate that content features are highly valuable, in general, and that topic modeling techniques are reliable and effective for social media user classification.

Currently, in the eLearning field, there are fewer studies using topic modeling to mine learner interests. For example, Zhang, Zhu, Zhao, Gu and Ting (2008) used behavioral analysis for interest mining in e-learning. Tobarra, Robles-Gómez, Ros, Hernández and Caminero (2014) analyzed student behaviors and relevant topics in virtual learning communities. Peng, Liu, Liu, Gan and Sun (2016) explored learners' topic interests by mining course reviews using an LDA-like model , showing that learner interactions with these texts were helpful in building learner topic interest profiles; moreover, the combination of interactive behavioral features with textual content was useful for mining learner topic interests (Peng, Liu, Liu, Gan, & Sun, 2016). To sum up, learners not only create content that interests them. They are also more likely to collect content created by other learners, which interests them. Consequently, in this study, we use the topic model mining approach to discover learner interests automatically by integrating learner generated content and data about their interactions with resources available in open learning environments. We propose a method to mine learner interests automatically in open learning environments using a learner-topic model.

Methodology

We introduce a LTM for mining learner interests based on two types of learning-related data. An overview of our methodology is shown in Figure 2. In this section, we first describe the collection of the two types of learning-related data, which includes the learner creation data and the collection data. Next, we briefly review the LDA model. Then based on LDA, we introduce our learner-topic model and demonstrate how to mine learner interests with our LTM.

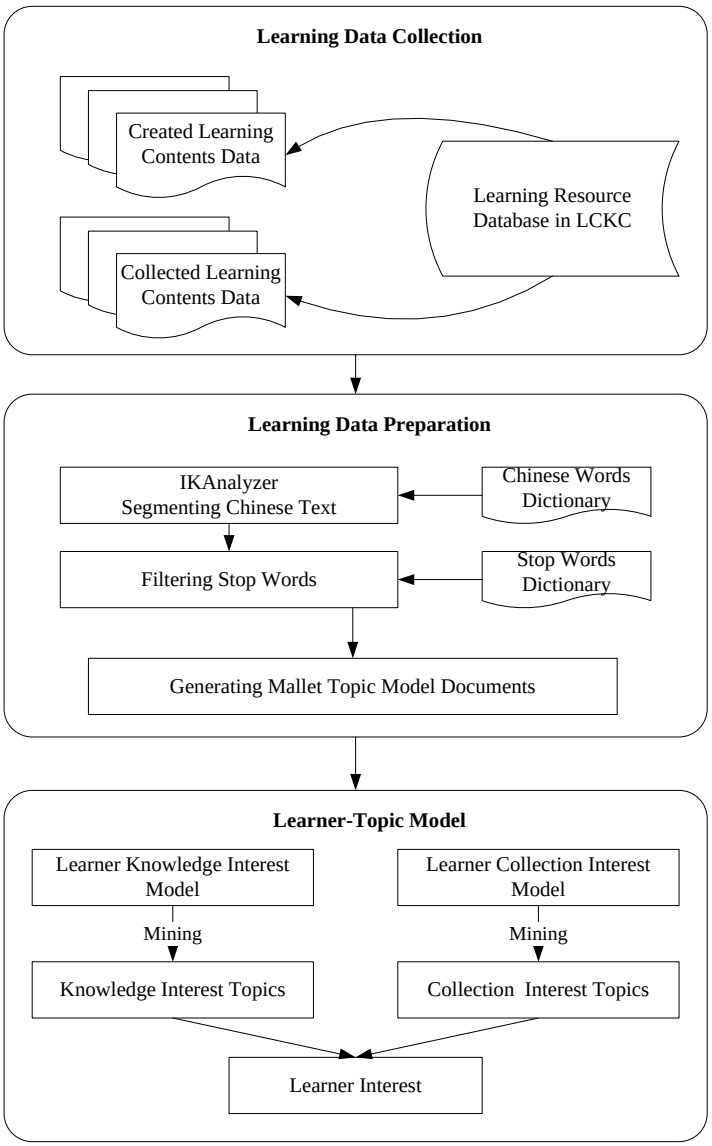

Figure 2. A framework of the methodology

Learning data collection

Learning Cell is a resource organization model for ubiquitous learning in a seamless learning space (Yu, Yang, Cheng, & Wang, 2015). The Learning Cell Knowledge Community (LCKC) (see http://lcell.bnu.edu.cn), inaugurated in May 2011, is an open knowledge community constructed based on Learning Cell (Yang, Qiu, Yu, & Tahir, 2014). As of April 20, 2017, LCKC had 24508 registered users and 81218 learning cells.

Our study uses learning-related data from the learning resource database of LCKC. For privacy protection, we removed learners' real names from the dataset. Before using topic modeling for interest mining, we first extract learning content data and obtain the sets of created learning content data and collected learning content data for each learner. An example of learner creation data and collection data for a single LCKC learner is shown in Figure 3.

Figure 3. Examples of learner creation data and collection data

Data preparation

We use the natural language segmentation system of IKAnalyzer (see https://www.oschina.net/p/ikanalyzer) to segment the Chinese text obtained from LCKC and extract Chinese words. In our study, we use the stop word dictionary to filter out stop words.

LDA model

The primary goal of this study is to develop a model to mine the learning-related data for learner interests. The LDA topic model (Blei, Ng, & Jordan, 2003) is a general Bayesian probabilistic framework for modeling documents linked by a layer of latent topics. It is a method for clustering the documents based on latent topics. The LDA topic model assumes that words in each document were generated from a mixture of latent topics, where each latent topic is represented as a multinomial probability distribution over words. A document is modeled as a set of draws from a mixture distribution over a set of latent topics and a topic is modeled as a probability distribution over words. There are many methodological extensions to LDA. LDA topic models have been applied and have demonstrated reliability in many text mining tasks. The following learner-topic model builds on LDA and its extensions.

Learner-Topic model

In open learning environments, learners not only create learning content, but also collect learning content created by other learners in which they are interested. These activities suggest that the generative process for learning content should meet the following rules for a topic model:

- Learning content created and collected by a learner are related to the learner's interests. The learning content originates from the learner's topic distribution. In the generative process of learning content, we should choose a latent topic from the learner's topic distribution for each word in the learning content.
- Learner interest is composed of two parts: knowledge interest and collection interest. Knowledge interest refers to the learner's interest in creating learning content. Collection interest refers to the learner's interest in collecting learning content.
- Learning content created by a learner should be generated from the learner's knowledge topic distribution, while learning content collected by a learner should be generated from the collection topic distribution.

Based on the above rules, we propose an LTM for topic mining in open learning environments. The LTM models the generation process of the learner's learning content of interest, and then we deduce learner interest from the model. For each learner, learning content of interest is composed of (1) created learning content and (2) collected learning content. A learner creates learning content based on his/her knowledge interests, and therefore the topic distribution of the created learning content should be determined by the learner's knowledge interests. Thus, we

can model the generation process for created learning content based on the learner's knowledge interests. Collected learning content is created by other learners. The current learner collects that learning content, which reflects the learner's collection interests. Therefore, we can derive the learner's collection interests by aggregating topic distributions of the collected learning content. In our proposed learner-topic model, we aggregate all the learning content created by a learner as a single document $D_{created}$ and all the learning content collected by a learner as a single document $D_{collected}$. Thus, each document essentially corresponds to a learner.

In open learning environments, each learner can create learning content and collect other learners' learning content. Therefore, we can model the generation process of all the learning content for each learner by considering learner interests in terms of knowledge interests and collection interests. This is the reason why we divide learner interest into two parts in our model. We can get the topics of knowledge interest and the topics of collection interest for a particular learner by learner-topic modeling. Because learner interests consist of two parts, we can generate the learner interest words by aggregating knowledge interests and collection interests.

Learner knowledge interest model

For a learner, the generating probability of the word *w* is given as follows:

$$P(w|l,\theta,\emptyset) = \sum_{z=1}^{K} P(w|z,\emptyset_z)P(z|l,\theta_l) \qquad (1)$$

where the document is created by learner *l*, *z* is a topic, *K* is all topics, θ_l is the multinomial distribution of the learner *l* over topics, $\emptyset_z$ is the multinomial distribution of the topic *z* over words.

The generation process algorithm for the Learner Knowledge Interest Model is as follows:

1. **for** each topic $z \in [1, K]$ **do**
2. choose a distribution over word $\emptyset_z$ from a Dirichlet distribution with parameter β, $\emptyset_z \sim Dir(\beta)$
3. **end for**
4. **for** each learner $l \in [1, L]$ **do**
5. choose a distribution over topics θ_l from a Dirichlet distribution with parameter α, $\theta_l \sim Dir(\alpha)$
6. **for** the *n*th token in the document set $D_{created}$ created by *l* **do**
7. choose a latent topic z_{ln} from the multinomial distribution θ_l,
8. $z_{ln} \sim Multi(\theta_l)$
9. generate the word token w_{ln} from the multinomial distribution $\emptyset_{z_{ln}}$,
10. $w_{ln} \sim Multi(\emptyset_{z_{ln}})$
11. **end for**
12. **end for**

Learner collection interest model

For a learner, the generating probability of the word *w'* is given as follows:

$$P(w'|l,\theta',\emptyset') = \sum_{z'=1}^{K} P(w'|z',\emptyset'_{z'})P(z'|l,\theta'_l) \qquad (2)$$

where the document is collected by learner *l*, *z'* is a topic, *K* is all topics, θ' is the multinomial distribution of the learner *l* over topics, $\emptyset'_{z'}$ is the multinomial distribution of the topic *z'* over words.

The generation process algorithm of the Learner Collection Interest Model is as follows:

1. **for** each topic $z' \in [1, K]$ **do**
2. choose a distribution over word $\emptyset'_{z'}$ from a Dirichlet distribution with parameter β',

$\emptyset'_{z'} \sim Dir(\beta')$
3. **end for**
4. **for** each learner $l \in [1, L]$ **do**
5. choose a distribution over topics θ'_l from a Dirichlet distribution with parameter α', $\theta'_l \sim Dir(\alpha')$
6. **for** the *n*th token in the document set $D_{collected}$ collected by l **do**
7. choose a latent topic z'_{ln} from the multinomial distribution θ'_l, $z'_{ln} \sim Multi(\theta'_l)$
8. generate the word token w'_{ln} from the multinomial distribution $\emptyset'_{z'_{ln}}$,
9. $w'_{ln} \sim Multi(\emptyset'_{z'_{ln}})$
10. **end for**
11. **end for**

Discovering learner interests

We can find the word proportions over each topic and extract the representative words for each latent topic. Then we can find the latent topic proportions over the created learning content and extract the knowledge interests for each learner. We can also get the latent topic proportions over the collection learning content and extract the collection interests for each learner. The learner-topic model thus mines learner interests from two aspects: knowledge interests and collection interests. We use the java open source software of Machine Learning for Language Toolkit named Mallet (see http://mallet.cs.umass.edu) to implement the learner-topic model.

Results

Dataset

For our evaluation, we used a text dataset extracted from the Learning Cell Knowledge Community consisting of learning cell documents and learners. Stop words were removed from each learning cell document. In the learner-topic model generation phase, we selected 3538 learners from the Learning Cell Knowledge Community, each of whom had created at least one Learning Cell or collected one Learning Cell. This selection process yielded 45512 Learning Cells linked to these learners.

Parameters estimation

The learner-topic model requires specification of the Dirichlet prior hyper parameters $\alpha, \beta, \alpha', \beta'$. According to a previous study (Blei, Ng & Jordan, 2003), the Dirichlet prior hyper parameters setup of the learner-topic model in the experiment is: $\alpha = 50/K$, $\beta = 0.01$; $\alpha' = 50/K$; $\beta' = 0.01$. In our research, we used the perplexity (Blei, Ng & Jordan, 2003) to choose the optimal K (number of topics). Latent topics were extracted from a single sample using the 2000th iteration of Gibbs sampling.

Learner knowledge interest topics

In Table 1, we list knowledge interest topics extracted by the learner-topic model for five learners, including the top ten words of each topic. These words appear with high probability in each topic. The specific meaning of each topic is based on analysis of the semantics of the representative words. From Table 1, it is easy to confirm that the top ten high-frequency words for each topic are closely related to that topic.

Table 1. Topics of knowledge interest for learners

Learner	Topic	Top ten most frequent words
Learner1	"Educational technology"	education (教育) informatization(信息化) integration(整合) information technology(信息技术) learning(学习) curriculum(课程) teacher(教师) technology(技术) development(发展) method(方法)
	"Digital education"	photography(摄影) report(报告) research(研究) education(教育) project(课题) frontier(前沿) digitization(数字化) culture(文化) module(模块) West(西区)

Learner2	“Mobile learning resources”	learning(学习) design(设计) mobile(移动) resource(资源) development(开发) education(教育) network(网络) model(模式) work(作品) teaching(教学) application(应用)
	“Computer technique”	method(方法) problem(问题) file(文件) introduction(入门). database(数据库) command(命令) server(服务器) configuration(配置) optimization(优化) language(语言)
Learner3	“Leaping teaching”	Chinese course(语文) instance(案例) learning cell(学习元) train(培训) share(分享) teaching(教学) English course(英语) leaping(跨越式) teacher(教师) learning(学习)
	“Digital education”	photography(摄影) report(报告) research(研究) education(教育) project(课题) frontier(前沿) digitization(数字化) culture(文化) module(模块) West(西区)
Learner4	“Computer technique”	method(方法) problem(问题) file(文件) introduction(入门) database(数据库) command(命令) server(服务器) configuration(配置) optimization(优化) language(语言)
	“Computer technique learning”	platform(平台) software(软件) system(系统) ontology(本体) knowledge(知识) monthly learning report(学习月报) operation(操作) open source(开源) building(建设) introduction(介绍)
Learner5	“Learning cell system”	learning cell(学习元) learning(学习) knowledge group(知识群) enterprise(企业) university(大学) learning community(学习社区) educational technology(教育技术) education(教育) introduction(介绍) page(页面)
	“Educational technology”	education(教育) informatization(信息化) integration(整合) information technology(信息技术) learning(学习) curriculum(课程) teacher(教师) technology(技术) development(发展) method(方法)

Learner collection interest topics

In Table 2, we list collection interest topics extracted by the learner-topic model for five learners, including the top ten words of each topic. Again, we analyze the semantics of the representative words to get the specific meaning of each topic. We find that the top ten high-frequency words for each topic are closely related to the topic.

Table 2. Topics of collection interest for learners

Learner	Topic	Top ten most frequent words
Learner1	“Educational technology”	educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁) basic problem(基本问题) instructional design(教学设计) teaching(教学) society (社会) learning theory(学习理论)
	“Educational technology in society”	learning cell(学习元) platform(平台) summary(概述) enterprise(企业) university(大学) method(方法) advice(意见) improvement(改进) usage(使用) learning(学习) group(小组)
Learner2	“Poetry and computer program”	employment(就业) common problem(常见问题) poetry(诗词) ancient(古代) interpret (解读) program(编程) notice(公告) editing(编辑) server(服务器) leapfrogging (跨越式)
	“Educational technology”	educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁) basic problem(基本问题) instructional design(教学设计) teaching(教学) society(社会) learning theory(学习理论)
Learner3	“Computer assisted teaching”	Chongwen(崇文) Minsheng(民生) experimental school(实验学校) strategy(策略) extension(推广) happy holiday(快乐的节日) Heilongjiang(黑龙江省) computer(电脑) child(幼儿) experience(心得体会)
	“Educational technology”	educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁) basic problem(基本问题) instructional design(教学设计) teaching(教学) society(社会) learning theory(学习理论)

Learner4	“Computer program technique”	learning cell(学习元) framework(架构) sort(排序) audit(审核) version(版本) database(数据库) cluster(集群) logon(登录) process(流程) language(语言)
	“Technique and architecture ”	regular expression(正则表达式) forcer(动力) evolution(进化) Lin Huiyin(林徽因) Liang Sicheng(梁思成) arrangement(整理) test(测试) Chinese(中文) button(按钮) Prometheus(普罗米修斯)
Learner5	“Educational technology”	educational technology(教育技术) network(网络) paradigm(范式) new comprehensions(新解) change(变迁) basic problem(基本问题) instructional design(教学设计) teaching(教学) society(社会) learning theory(学习理论)
	“Information technology and basic education reform”	information technology(信息技术) course(课程) integration(整合) deep(深层次) application(运用) deepen(深化) perspective(视角) reform(改革) basic education(基础教育) theory(理论)

Learner interest words and self-defined tags

For this part of the study, we chose learners who had self-defined interest tags, enabling us to evaluate the model by comparing learner interests with learner self-defined interest tags. Learner self-defined interest tags are sets of keywords defined by a learner and used to describe his/her specialties and interests in LCKC (as shown in Figure 1). Therefore, learner self-defined interest tags are an informal reflection of learner interests.

To study the difference between learner interest words and learner self-defined interest tags, we first combined knowledge interest words (ten words selected from two of the most popular knowledge interest topics) and collection interest words (ten words selected from two of the most popular collection interest topics) as interest words for each learner. We then compared the learner self-defined interest tags with the combined set of learner interest words as shown in Table 3. For example, Learner3 provides the tag words “Chinese course(语文),” “English course(英语),” “leapfrogging (跨越式)” and “teaching(教学).” Therefore, we might expect that Learner3 is also interested in the research of the “leapfrogging project (跨越式项目).”

Table 3. Interest tags and interest words for learners

Learner	Interest tags	Interest words
Learner1	educational technology(教育技术学) ubiquitous learning(泛在学习) web based education platform(网络教育平台) learning cell(学习元) leapfrogging project(跨越式项目) poetry(诗歌)	education(教育) informatization(信息化) integration(整合) information technology(信息技术) learning(学习) photography(摄影) report(报告) research(研究) education(教育) project(课题) educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁) learning cell(学习元) platform(平台) summary(概述) enterprise(企业) university(大学) method(方法)
Learner2	educational technology(教育技术) ubiquitous learning(泛在学习) web based education platform(网络教育平台) leapfrogging project(跨越式项目)	learning(学习) design(设计) mobile(移动) resource(资源) development(开发) method(方法) problem(问题) file(文件) introduction(入门). database(数据库) command(命令) employment(就业) common problem(常见问题) poetry(诗词) ancient(古代) interpret (解读) program(编程) educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁)
Learner3	leapfrogging project(跨越式项目)traveling(旅行) English course(英语) Chinese course(语文)	Chinese course(语文) instance(案例) learning cell(学习元) train(培训) share(分享) teaching(教学) photography(摄影) report(报告) research(研究) education(教育) project(课题) Chongwen(崇文) Minsheng(民生) experimental school(实验学校) leapfrogging (跨越式) English course(英语) educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁)

Learner4	ubiquitous learning(泛在学习) ontology(本体) Semantic annotation(语义标注) natural language processing(自然语言处理)	method(方法) problem(问题) file(文件) introduction(入门) database(数据库) platform(平台) software(软件) system(系统) ontology(本体) knowledge(知识) learning cell(学习元) framework(架构) sort(排序) audit(审核) version(版本) regular expression(正则表达式) forcer(动力) evolution(进化) Lin Huiyin(林徽因) Liang Sicheng(梁思成)
Learner5	Web development(网站开发) educational technology(教育技术) personal learning(个性化学习) personal recommendation(个性化推荐) leapfrogging project(跨越式项目)	learning cell(学习元) learning(学习) knowledge group(知识群) enterprise(企业) university(大学) education(教育) informatization(信息化) integration(整合) information technology(信息技术) learning(学习) educational technology(教育技术) network(网络) paradigm(范式) new solutions(新解) change(变迁) information technology(信息技术) course(课程) integration(整合) deep(深层次) application(运用)

User evaluation

We invited 25 LCKC users to participate in the experiment. Each user evaluated their generated knowledge interests, collection interests and the overall learner interest result quantitatively, by counting the number of the discovered top ten words that accurately reflected their interests. The precision is equal to the number of the discovered top ten words relevant to an individual's interests divided by ten. Similar metric has been used in evaluating tasks (Michelson & Macskassy, 2010). Cronbach's alpha (.701) indicates an acceptable internal consistency estimate of reliability of evaluation scores.

Table 4. Interest mean and standard deviation

Interest	Mean (average precision)	*SD*	*N*
Knowledge Interest	.820	.076	25
Collection Interest	.764	.075	25
Learner Interest	.884	.068	25

The average levels of precision for knowledge interests, collection interests and overall learner interests are presented in Table 4. The percentages of precision for knowledge interests, collection interests and overall learner interests are presented in Figure 4.

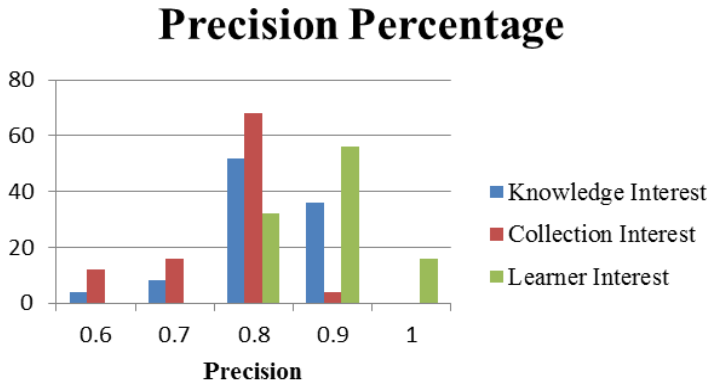

Figure 4. Percentages of precision for interests

Discussion

This study uses a learning content mining approach to automatically discover learner interests, using the combination of content generated by each learner and the dynamic interactions of learners with the online

resources to reflect learner interests. In this section, we discuss the results of our three research questions. In considering the learner knowledge and collection interests extracted in the experiment, we judge and analyze their relationships and differences from each learner's self-defined interest tags through observation.

Differences between learner knowledge interests and collection interests

In comparing the result of extracted knowledge interests (Table 1) with collection interests (Table 2), we note that, for each learner, knowledge interests and collection interests are similar. For example, "educational technology" is present in both the created learning content and the collected learning content for Learner1. Similarly, "computer techniques and educational technology" feature in both the created learning content and the collected learning content for learner 2. Visual analysis of the differences between knowledge interests and collection interests for each user suggests that these areas of interest are related and consistent in their subject matter. The results shown in Tables 1 and 2 address the research question regarding similarities and differences between learner knowledge interests and collection interests.

Differences between the mined interest words and self-defined interest tags

In comparing the result of mined interest words with self-defined interest tags (Table 3), we find that, for each learner, the set of mined interest words contains the learner's self-defined interest tags. For example, Learner1's self-defined tag of "educational technology" is among the interest topics (topic of "educational technology" and topic of "educational technology in society") extracted for Learner1. Similarly, Learner2's self-defined tag of "ubiquitous learning" is contained in the interest topics (topic of "mobile learning resources" and topic of "educational technology") elicited for Learner2. The results shown in Table 3 address the research question regarding the relationship between mined interest words and learner self-defined interest tags. We find that, for a particular learner, the set of mined interest words contains the learner's self-defined interest tags, but is not limited to the scope of the learner's self-defined interest tags. For example, Learner4 is also interested in the topic of "architecture," which is not found in Learner4's self-defined interest tags.

Effectiveness of learner-topic model

Experimental results show that the mean value of the learner interests is higher than the mean value of knowledge interests and collection interests separately. Generating learner interests from the knowledge interests and collection interests to acquire learners' interest characteristics is more accurately. User experimental results in Table 4 and Figure 4 indicate that the learner-topic model for mining learner interests is appropriate and effective, and support the use of the topic-modeling approach for discovering learner interests in open learning environments. In comparing user self-defined interest tagging with the results of the learner-topic model, it suggests that the learner-topic model represented in Figure 5 not only automatically discovers learner interests in detail, but also can be used to update learner interest tags when their interests change. This model thus enables learners to annotate their interests dynamically without manual intervention.

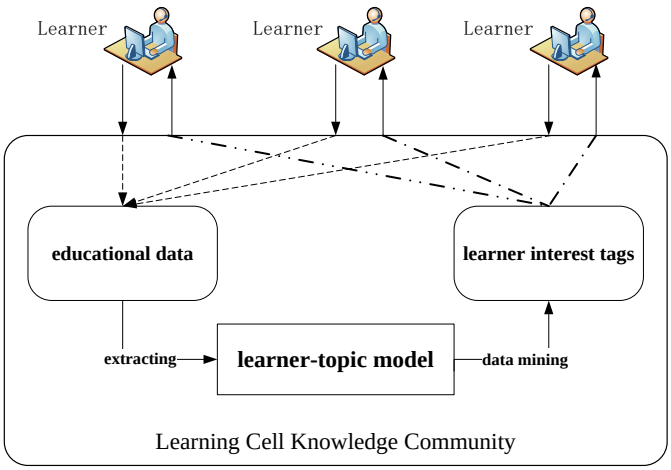

Figure 5. Feature of the learner-topic model in LCKC system

Conclusions

Building a learner interest model has been an important research topic in open learning environments. This paper discussed and demonstrated a method of using a LTM to solve the problem of mining learner interests and automatically generating learner interests in open learning environments. Experiments on a dataset from the Learning Cell Knowledge Community demonstrate that the method is able to discover learner interests effectively. Moreover, we find that knowledge interests and collection interests are related and consistent in their subject matter. Finally, the learner interest words discovered by the LTM method reflect self-defined interest tags, but cover a broader range of interests. In open learning environments, different interactions between learners and various sources of content indicate learners' attention to diverse topics of information. Learners' varied interests are revealed by the combination of knowledge interests and collection interests, which we derived from learner generated content plus data on the dynamic interactions between learners and online resources. Different semantic behavior actions such as "create," "like" could also provide insight into online learners' topic interest profiles (Peng, Liu, Liu, Gan, & Sun, 2016; Zhao, Cheng, Hong, & Chi, 2015). Our findings offer an approach to automatically discover learner interests in detail from learner generated content and dynamic interactions and relationships in open learning environments. Learning resource content and learners' behavioral features ("create" and "collect") are merged to more accurately acquire the learners' interests. Other useful operational behaviors will be utilized in the process of educational data mining into the learner-topic model, e.g., "comment," "share."

Further research could be conducted to address additional relationships between learners and online resources of interest. In our future work, we plan to use the mined interest words for personal resource recommendation, learning peers discovery and experts finding.

Acknowledgements

This research is funded by the project "the Research on Internet plus Educational System" (Project No. 16JZD043), major research subject of philosophy and social science of the ministry of education.

References

Abdous, M. H., He, W., & Yen, C.-J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology & Society*, *15*(3), 77–88.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Brusilovsky, P. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The Adaptive Web* (pp. 3-53). Berlin, Germany: Springer-Verlag.

Gong, S. (2012). Learning user interest model for content-based filtering in personalized recommendation system. *International Journal of Digital Content Technology & Its Applications*, *6*(11), 155-162.

Goda, K., & Mine, T. (2011). Analysis of students' learning activities through quantifying time-series comments. In *Proceedings of Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 154-164). Berlin, Germany: Springer.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, *29*(1), 90-102

Hwang, G. J. (2005). A data mining approach to diagnosing student learning problems in sciences courses. *International Journal of Distance Education Technologies*, *3*(4), 35-50.

Ishii, T., Mizoguchi, S., Kimita, K., & Shimomura, Y. (2015). A Topic model for clustering learners based on contents in educational counseling. In Yamamoto S. (Ed.), *Human Interface and the Management of Information. Information and Knowledge in Context* (pp. 323-331). Cham, Germany: Springer.

Kuswara, A.U., & Richards, D. (2011). Realising the potential of Web 2.0 for collaborative learning using affordances. *Journal of Universal Computer Science*, *17*, 311–331.

Li, M., Sagl, G., Mburu, L., & Fan, H. (2016). A Contextualized and personalized model to predict user interest using location-based social networks. *Computers Environment & Urban Systems*, *58*, 97-106.

Liu, Q., Chen, E., Xiong, H., Ding, C. H. Q., & Chen, J. (2012). Enhancing collaborative filtering by user interest expansion via personalized ranking. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, *42*(1), 218-33.

Minami, T., & Ohura, Y. (2013). Lecture data analysis towards to know how the students' attitudes affect to their evaluations. In *Proceedings of the 8th International Conference on Information Technology and Applications* (pp. 164-169). Sydney, Australia: ICITA.

Minami, T., & Ohura, Y. (2015). How student's attitude influences on learning achievement? An Analysis of attitude representing words appearing in looking-back evaluation texts. *International Journal of Database Theory and Application*, *8*(2), 192-144.

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: A First look. In *Workshop on Analytics for Noisy Unstructured Text Data* (pp. 73-80). New York, NY: ACM.

Ni, X., Lu, Y., Quan, X., Liu, W., & Hua, B. (2012). User interest modeling and its application for question recommendation in user-interactive question answering systems. *Information Processing & Management*, *48*(2), 218-233.

Peng, X., Liu, S., Liu, Z., Gan, W., & Sun, J. (2016).Mining learners' topic interests in course reviews based on like-LDA model. *International Journal of Innovative Computing, Information and Control*, *12*(6), 2099-2110.

Pennacchiotti, M., & Popescu, A. M. (2011). A Machine learning approach to Twitter user classification. In *International Conference on Weblogs and Social Media* (pp. 281-288). Barcelona, Spain: DBLP.

Sekiya, T., Matsuda, Y., & Yamaguchi, K. (2010). Development of a curriculum analysis tool. In *Proceedings of the 9th international conference on Information technology based higher education and training* (pp. 394-399). Piscataway, NJ: IEEE Press.

Sorour, S. E., Goda, K., & Mine, T. (2015). Evaluation of effectiveness of time-series comments using machine learning techniques. *Journal of Information Processing*, *23*(6), 784-794.

Sorour, S. E., Goda, K., & Mine, T. (2017). Comment data mining to estimate student performance considering consecutive lessons. *Educational Technology & Society*, 20(1), 73–86.

Tobarra, L., Robles-Gómez, A., Ros, S., Hernández, R., & Caminero, A. C. (2014). Analyzing the students' behavior and relevant topics in virtual learning communities. *Computers in Human Behavior*, *31*, 659-669.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1-35.

Wu, P., & Yu, S. (2015). Design of a novel intelligent framework for finding experts and learning peers in open knowledge communities. *EAI Endorsed Transactions on Future Intelligent Educational Environments*, *15*(2):e4.

Xu, Z., Ru, L., Xiang, L., & Yang, Q. (2011). Discovering User Interest on Twitter with a Modified Author-Topic Model. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 422-429). Washington, DC: IEEE Computer Society.

Yang, X., Qiu, Q., Yu, S., & Tahir, H. (2014).Designing a trust evaluation model for open-knowledge communities. *British Journal of Educational Technology*, *45*(5), 880–901.

You, Q., Bhatia, S., & Luo, J. (2016). A Picture tells a thousand words -- about you! User interest profiling from user generated visual content. *Signal Processing*, *124*(C), 45-53.

Yu, S., Yang, X., Cheng, G., & Wang, M. (2015). From learning object to learning cell: A Resource organization model for ubiquitous learning. *Educational Technology & Society*, *18*(2), 206–224.

Zhang, N., Zhu, M., Zhao, L., Gu, R., & Ting, I. (2008). Interest mining in virtual learning environments. *Online Information Review*, *32*(2), 133-146.

Zhao, Z., Cheng, Z., Hong, L., & Chi, E. H. (2015). Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web* (Vol.87, pp. 1406-1416). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.