文章编号:1007－6735(2018)03－0259－08　DOI: 10.13255/j.cnki.jusst.2018.03.008

# 基于文本挖掘的互联网教育课程主题发现与聚类研究

李梦杰[1]，　刘建国[2]，　郭　强[1]，　李仁德[1]，　汤晓雷[3]

（1. 上海理工大学 复杂系统科学研究中心，上海　200093；2. 上海财经大学 科研实验中心，上海　200433；
3. 沪江教育科技股份有限公司，上海　201203）

**摘要：**如何通过有效的数据挖掘对互联网教育平台中的课程主题进行挖掘、聚类是当前互联网教育亟待解决的问题之一。实验基于文本信息对某互联网教育平台的 1 472 门课程体系的主题分布及类别进行了分析。采集了某平台 1 472 门课程的描述信息，进而通过自建词典和停用词库对文本进行切词分词，并通过 TF-IDF 对词频权重进行处理。利用 LDA 主题模型对课程的主题分布进行识别，发现了 230 个主题，并得到了每门课程在这 230 个主题下的文档–主题分布以及主题–词分布。进一步基于分布相似性函数对课程进行层次聚类，发现基于不同抽象层次主题的课程相互关联。最后将 16 个主题信息进行了可视化，这些主题分别从内容和数量两个角度反映出了课程的主题特征以及课程的聚合分布情况。

**关键词：**主题发现；层次聚类；互联网教育；文本挖掘
**中图分类号：**N 32　　**文献标志码：**A

# Topic Discovery and Clustering Research for Online Courses Based on Text Mining

LI Mengjie[1], LIU Jianguo[2], GUO Qiang[1], LI Rende[1], TANG Xiaolei[3]

(1. *Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *Laboratory Center, Shanghai University of Finance and Economics, Shanghai 200433, China*; 3. *Hujiang Education &Technology Co., Ltd., Shanghai 201203, China*)

**Abstract:** How to dig out informations from courses and conduct cluster analysis through effective data mining for online education is one of the problems to be solved. The topic distribution and classification of 1 472 courses from an online education platform were analyzed experimentally based on the text description. The text informations of 1 472 courses from the platform were collected, a customized dictionary and stop word list were constructed to do the word segmentation, and then the TF-IDF was employed to calculate the word frequency weighting. The topic distribution was recognized by using LDA and 230 topics were discovered.Both the document-topic distribution and topic-word distribution

**收稿日期:** 2017−09−05
**基金项目:** 国家自然科学基金资助项目(61773248，71771152)
**第一作者:** 李梦杰(1992–)，女，硕士研究生．研究方向：文本分析、网络科学．E-mail：susan555@126.com
**通信作者:** 刘建国(1979–)，男，教授．研究方向：科学知识图谱分析、网络科学．E-mail：liujg004@ustc.edu.cn

for each course text were obtained under the 230 topics. The hierarchical clustering for courses was completed based on the distribution similarity function and it is found that the courses were interrelated based on different levels of abstract topics. In the end, informations of 16 topics were visualized. This discovery of topics hidden in the semantics reflects the topic feature and the aggregate distribution of massive courses.

**Keywords:** *topic discovery*; *hierarchical clustering*; *online education*; *text mining*

主题发现也被称为主题抽取或主题识别，目的是处理和分析大规模信息并且使用户能够以最快速有效的方式了解信息内容、发现信息中的主题[1-2]。主题发现[3]能在一定程度上挖掘出文本之间的关联信息[4]，从而得以被广泛地运用到非结构化数据的挖掘中。文献[5-8]对在线数据库和科技文献的研究热点以及研究进展进行了主题发现。文献[9-12]则分别基于新浪网页、雅虎新闻、百度知道等问答系统以及YouTube视频网站的Web文本完成了主题发现的工作。而文献[13-17]则基于用户自产生的文本如微博、博客、推文、用户评论等进行主题发现。在互联网教育领域中，文本作为教育大数据中一种特质的类型，最真实、直接地反映了学习者的学习动机、认知发展、情感态度、学习体验[18]。互联网教育平台海量课程的涌现使得课程的投放者和课程的学习者无法逐一对其认知完全，并且，课程丰富而复杂的描述信息多是自然语言的文本表述，这些都给我们认识和了解当下在线课程背后的知识分布以及结构特征带来了挑战。因此，对于在线课程海量描述性文本信息的挖掘就显得十分必要，而此时非结构化教育大数据的主题发现就会展示出它的高效性以及重要性[17, 19-20]。当前，互联网教育平台的课程种类繁多，大多用文本进行课程内容描述。无论是平台管理者，还是消费者，都很难通过阅读课程描述信息对所有的课程进行全面了解。

本文爬虫抓取某日语特色社团所选的共1 479门课程的描述信息网页数据，采用文本挖掘的方法对课程的文本信息进行主题发现的实证研究。首先对爬虫抓取的文本信息进行数据的预处理，包括数据的清洗、中文的分词、特征的选择等；其次，利用主题模型对文本进行数值化的表示；然后，基于文本之间的距离进行无监督的文本聚类；最后，进行主题发现。实验结果发现这些课程基于不同距离阈值的聚类呈现出不同层次的共16个主题，并且可按主题被划分为9个类别。由于课程是来自于学日语的社团，所以有88.3%的课程主题都与语言学习相关，并且主题的内容也更多反映出了学习者的学习目的。

## 1 文本的预处理

### 1.1 数据准备

课程的描述信息主要包括课程名称、学习目标、适用人群、课程特色等字段，基于当前国内某互联网教育学习平台所提供的API，利用Python爬虫抓取数据，最终得到1 479篇文档作为本文的数据基础，数据的时间窗口为2015年7月1日至2015年12月31日。

Web文本数据由于存在非结构化、动态更新等问题，给数据的清洗和预处理带来了很多困难。例如中文描述文本与一些特殊字符“#、&”、HTML标记、各种超链接的URL等混合在一起，需要使用正则表达式进行过滤和删除；另外，Web 2.0的普及造成网络流行语的不断涌现，数据清洗的时候需要人工将网络语翻译成规范文本表达，如“小白”要被翻译成“初学者”等；而且，在语言类的知识里随处可见英文缩写和简写，这些同样需要被译成对应的中文表达，如“J.test”，“PETS”要被翻译成“实用日本语鉴定考试”，“全国英语等级考试”。本文的研究选用了来自5个不同专业背景的志愿者对这些数据进行清洗和规范化处理。

对清洗后的文本信息再进行下一步的预处理工作，文本的预处理工作围绕图1的预处理框架图展开。

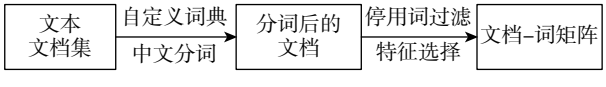


**图1 预处理框架图**

**Fig.1 Preprocessing frame diagram**

### 1.2　中文分词

因为中文词无间隙，语法结构特殊，所以为了提取非结构化的文本数据中的有效信息，首先要利用中文分词算法将这些非结构化的描述性文本转化成结构化的数据。现有的中文分词算法分为基于规则的分词方法、基于统计的分词方法和基于理解的分词方法[21]。Jieba 分词算法基于规则和统计相结合，本文采用 Python 的 Jieba 分词模块对语料进行分词。

由于使用的语料为互联网教育领域中以日语学习类为主的语料，专业性较强，这将会给分词工作带来如下困难：歧义的消除问题，比如“五十音图”是切分为“五十/音图”还是整体作为一个专有名词；未登录词的识别问题，实验中发现语料中有很多专业术语和网络新词，这些需要通过结合实际情况和专业背景，建立贴合语料的自定义词典，从而提高分词的准确性；分词过程中还有关于未登录词和登录词的识别顺序问题，比如“商务英语”是应该作为未登录词进行优先的识别还是应该按原有的登录词直接切分为“商务/英语”，这些也同样需要根据实际需求更改自定义词典中的词频，词频越大对应的词就能优先被切分出来。

比如，对于“零基础或者有日语五十音图基础、打算巩固并进一步学习的日语兴趣爱好者”这句话，不添加自定义词典的情况下，会被分成“零/基础/或者/有/日语/五十/音图/基础/、/打算/巩固/并/进一步/学习/的/日语/兴趣/爱好者”，可以看出“零”“五十”“音图”这些词都失去了在语境中的本来意义，而如果将“零基础”“五十音图”加入到自定义词典并且在分词的过程中读取自定义词典，这句话则会被切分成“零基础/或者/有/日语/五十音图/基础/、/打算/巩固/并/进一步/学习/的/日语/兴趣/爱好者”，显然后者的分词结果更有效。

基于对原始语料的初步了解，首先建立了含有 324 个词条的自定义词典，并不断地调整词频的大小和继续扩充自定义词典，使得分词结果尽可能准确。最后建立了总共含有 575 个词条的自定义词典。

数据的清洗和预处理排除掉了一些无效的测试子班和只有课程名称的描述文档，最终得到 1 472 篇有效的文本文档。基于已建立好的自定义词典，利用 Python 的 Jieba 分词包，对这 1 472 个有效的文档进行了分词。

### 1.3　特征选择

在分词的过程中，类似“我们、在、的、是、太”等出现频率太高却没有太大意义或者类别色彩不强的词，需要被过滤掉从而减少存储空间和计算时间。并且一些类似于“掌握、学习、课程、熟练”等与主题不相关的词也需要被过滤以减噪。使用停用词库之前，比如“提升阅读能力、听力、翻译、口译等各项技能”这句话，会被分成“提升/阅读/能力/、/听力/、/翻译/、/口译/等/各项/技能”，而加入了停用词库以后的结果是“阅读/听力/翻译/口译”，可以看出剔除掉了“提升”、“各项”等无效信息后，分词结果内容显得精炼、集中。

实验除了加入通用的中文停用词库过滤无效信息以外，还针对语料的特点以及实际需求建立了贴合语料的停用词库，对初步的分词结果进行减噪，停用词库共包含 2 827 个词条。分词结果确定以后，将自然语言的文本信息转化为计算机能够理解和存储的数据类型。

对文本的表示有多种模型，如向量模型、布尔模型和概率模型。空间向量模型可以较好地实现文本的抽象[22]。基于统计的代表性方法——向量空间模型（vector space model, VSM）[23]根据词频建立向量，并使用 TF-IDF（term frequency and inverse document frequency）给每个向量的特征词计算权重。

TF-IDF 算法由 Salton 等[24]首先提出，用于评估某个字或某个词对于一个语料库中的一个文本的重要程度。对于给定的文本文档，词频表示为

$$tf_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

式中，分子表示词 $w_{ij}$（文档 $\boldsymbol{d}_j$ 中的第 $i$ 个词）出现的次数，分母表示文档 $\boldsymbol{d}_j$ 中所有词之和。反文档频率为

$$idf_i=\ln\frac{|D|}{\left|\left\{j:t_i\in\boldsymbol{d}_j\right\}\right|} \tag{2}$$

$idf_i$ 指总文档的数目 $D$ 与包含有词语 $w_{ij}$ 的文档的数目求商之后的自然对数值，$tfidf_{ij}=tf_{ij}\times idf_i$，$tfidf_{ij}$ 的值越大表示词 $w_{ij}$ 的重要性越大、越关键。

对分词后的 1 472 篇文档进行过滤停用词，然后用空间向量表示，对每篇文档选择权重较大的前 20 个词作为关键词，得到总共 1 472 个多维特征词向量。

## 2 主题表示

由于 TF-IDF 算法会忽略词项之间的语义相关性以及文本集合的内部结构，而文本分类的重要工具 LDA（latent dirichlet allocation）[25-26]加入了对语义信息的考虑，利用词袋模型针对文档的内部结构以及文档之间的结构进行建模[27]，并且相较于 pLSA（probabilistic latent semantic analysis）[28-29]，LDA 不仅进行了贝叶斯的改进，还能通过直接分析词的语义内涵，在不增加计算量的情况下对新加入的语料进行识别。

由于实验中课程的文本描述内容比较短，对语料进行了特征的选择后，很多文档向量的特征词不足 20 个，而 LDA 主题建模能一定程度上减少数据稀疏性问题。故采用 LDA 对文档集进行主题建模，将文档表示成不同比例题的混合，为后面文本文档的有效相似性测量奠定基础。

### 2.1 LDA 主题模型

Blei 等[30]在 2003 年提出三层贝叶斯的概率主题模型 LDA，该模型将文档集中的每一篇文档都视为若干个主题的混合分布，而每个主题是给定的词汇表中词语的一个分布[31]。

LDA 的模型表示如图 2 所示。在 LDA 的生成模型中，$M$篇文档对应于$M$个独立的狄利克雷–多项式共轭结构；$K$个主题对应于$K$个独立的狄利克雷–多项式共轭结构。

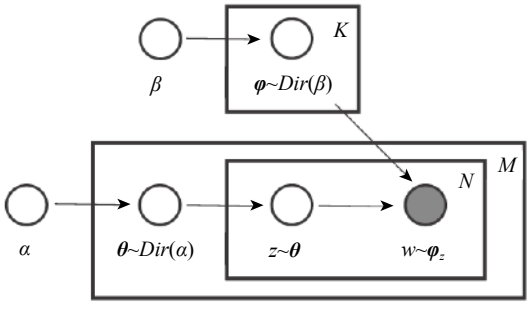


**图 2 LDA 的图模型表示**

**Fig.2 LDA graph model representation**

图中的空心点表示隐含变量，实心点表示可观察值，矩形表示重复过程。大矩形表示文档集（总共 $M$ 篇文档）中的每篇文档从狄利克雷分布中反复抽取主题分布 $\boldsymbol{\theta}$；小矩形表示从主题分布中反复抽样产生文档的词，一共抽取 $N$ 个词。一篇文档 $\boldsymbol{d}_j$ 的产生过程可以表示为两个过程：a.从狄利克雷分布 $p(\boldsymbol{\theta}|\alpha)$随机选择一个 $k$ 维的向量 $\boldsymbol{\theta}_j$，表示文档 $\boldsymbol{d}_j$ 中的主题混合比例；b.根据特定的主题比例对文档 $\boldsymbol{d}_j$ 中的每个词进行反复抽样，得到 $p(w_{ij}|\boldsymbol{\theta}_j,\beta)$。

对于给定的文档集，$w_{ij}$（第 $j$ 篇文档中的第 $i$ 个词）表示可以被观察到的已知变量，$\alpha$ 是每篇文档下主题的多项分布的狄利克雷先验参数，反映了文档集合中主题的相对强弱；$\beta$ 是每个主题下特征词的多项分布的狄利克雷先验参数，代表了所有主题自身的概率分布；而第 $j$ 篇文档中的第 $i$ 个词的主题变量 $\boldsymbol{z}_{ij}$、第 $j$ 篇文档下的主题分布隐含变量 $\boldsymbol{\theta}_j$（$k$ 维向量，$k$ 为主题个数）和第 $k$ 个主题下的特征词的分布 $\boldsymbol{\varphi}_k$（$v$ 维向量，$v$ 为词的总数）都是需要根据观察到的变量来学习估计的。主题和特征词的联合概率为

$$p(\boldsymbol{\theta},\boldsymbol{z}_j,w|\alpha,\beta)=p(\boldsymbol{\theta}_j|\alpha)\prod_{i=1}^{N}p(\boldsymbol{z}_j|\boldsymbol{\theta}_j)p(w_{ij}|\boldsymbol{z}_j,\beta), \quad (3)$$

整合 $\boldsymbol{\theta}$ 和 $\boldsymbol{z}$，得到文档的边缘分布为

$$p(w|\alpha,\beta)=\int p(\boldsymbol{\theta}_j|\alpha)\left(\prod_{i=1}^{N}\sum_{\boldsymbol{z}_i}p(\boldsymbol{z}_i|\boldsymbol{\theta}_j)p(w_i|\boldsymbol{z}_i,\beta)\right)\mathrm{d}\boldsymbol{\theta} \quad (4)$$

在实验过程中，整个文档集（即 1 472 个特征词向量的集合）作为输入内容进行 LDA 训练。主题数 $K$ 需要在模型训练前指定。实验中，选取使得模型的困惑度 $Q$ 取最小值时的主题数[31-33]。

$$Q=\mathrm{e}^{-\sum_{j=1}^{M}\log(p(w))/V} \quad (5)$$

式中：$V$ 为文档集中的总词数；$p(w)$表示单词 $w_{ij}$ 在所有主题的分布值与该词所在文档的主题分布的乘积。

### 2.2 LDA 训练结果

通过反复多次的 LDA 训练，得到困惑度随不同主题个数的变化如图 3 所示。图中，困惑度最小时对应的主题个数为 230，故最后设定 $K$=230。对于经验参数 $\alpha$ 和 $\beta$，参照文献[32]设置为 $\alpha$=50/$k$，$\beta$=0.01。LDA 最重要的是得到对两组参数的估计，分别是各主题下词语的概率分布 $\boldsymbol{\varphi}$ 和各文档的主题概率分布 $\boldsymbol{\theta}$[34]，这些估计值也就是 LDA 训练之后通过采样得到的输出结果。实验中采用吉布斯抽样（Gibbs sampling）[32]的方法，每次选取概率向量的一个维度，给定其他维度的变量值抽样确定当前维度的值，不断地迭代，直至收敛，输出待估参数[35]。

设定迭代次数为 2 000，最后得到的参数估计结果如下：

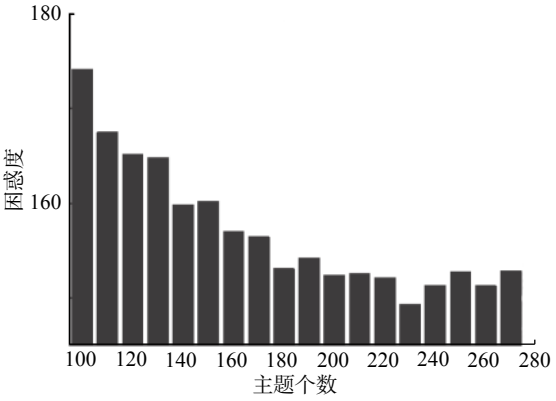


**图 3　不同主题个数的困惑度**

**Fig.3　Perplexity of different topics**

**a**. 一个 $K\times V$ 的矩阵 $\boldsymbol{\varphi}$，表示每个主题下生成每个词的概率。$K$ 表示主题的个数为 230，$V$ 表示文档集中所有不同词的个数为 2 173。

**b**. 一个 $M\times K$ 的矩阵 $\boldsymbol{\theta}$，表示文档集中每个文档生成主题的概率，$M$ 表示文档的总数为 1 472，$K$ 是主题数。

表 1 列出了 LDA 训练之后得到的 230 个主题中的前 4 个主题（分别为 Topic 0th, Topic 1th, Topic 2th, Topic 3th），Freq（Frequency 的缩写）对应的数值越大表示该主题产生该词的概率越大，如前 4 个主题产生“能力考”一词的概率为 0.270 6，“思维”一词的概率为 0.109 1，“应试”一词的概率为 0.334 8，“考试”一词的概率为 0.522 6，即可认为前 4 个主题与“能力考”、“思维”、“应试”、“考试”这 4 个词最相关。

**表 1　Top 4 个主题**

**Tab.1　Top 4 topics**

| Topic 0th | Freq | Topic 1th | Freq | Topic 2th | Freq | Topic 3th | Freq |
|---|---|---|---|---|---|---|---|
| 能力考 | 0.270 6 | 思维 | 0.109 1 | 应试 | 0.334 8 | 考试 | 0.522 6 |
| 测试 | 0.090 3 | 创意 | 0.065 5 | 语法 | 0.145 6 | 词汇量 | 0.082 6 |
| 随堂 | 0.090 3 | 思维能力 | 0.054 6 | 设计 | 0.014 7 | 申请 | 0.013 9 |
| 考察 | 0.090 3 | 自我 | 0.043 7 | 拼唱 | 0.014 7 | 逻辑推理 | 0.013 9 |
| 逐项 | 0.067 7 | 激活 | 0.043 7 | 洞悉 | 0.014 7 | 中医 | 0.013 9 |
| 重点难点 | 0.056 5 | 导图 | 0.032 8 | 融合 | 0.014 7 | 父母 | 0.013 9 |
| 传达 | 0.022 7 | 逻辑 | 0.032 8 | 流畅地 | 0.014 7 | 日语考试 | 0.013 9 |
| 口译 | 0.011 4 | 惯性 | 0.032 8 | 文 | 0.014 7 | 速记 | 0.013 9 |
| 健康 | 0.011 4 | 作者 | 0.032 8 | 从业者 | 0.014 7 | 科班 | 0.013 9 |
| 五十音图 | 0.011 4 | 创始人 | 0.032 8 | 网络 | 0.014 7 | 中短篇 | 0.000 1 |

## 3　课程聚类

### 3.1　距离的测量

对 1 472 篇课程描述信息文档进行 LDA 训练后，得到了每篇文档在 230 个主题上的分布，即 1 472×230 的主题概率分布矩阵 $\boldsymbol{\theta}$，可看作是每篇文档在 230 维主题空间上的分量值。因为每篇文档在 230 个主题上的分布都是从一个分布中通过吉布斯采样得来的[30]，且衡量两个概率分布之间距离的指标——Kullback-Leible 散度不仅考虑了分布的数值大小，还能反映分布的特征，且测量精度通常要优于余弦距离等，故采用 KL 距离来度量文档之间的距离。当两个随机分布相同时，它们的 KL 距离为零，当两个随机分布的差别增大时，它们的 KL 距离也会增大。又因为 KL 距离不具有对称性，于是本文采用更平滑的 KL 距离的变形式 JS（Jensen-Shannon divergence）距离[36]计算文档之间两两的距离。相较于 KL 距离，JS 距离的计算结果是对称的，并且结果落在区间[0, 1]之间，更适合作距离计算。JS 距离的计算公式如下：

$$JSD(\boldsymbol{\theta}_p\|\boldsymbol{\theta}_q)=\frac{1}{2}D(\boldsymbol{\theta}_p\|\boldsymbol{\theta}_m)+\frac{1}{2}D(\boldsymbol{\theta}_q\|\boldsymbol{\theta}_m) \quad (6)$$

$$JSD(\boldsymbol{\theta}_p\|\boldsymbol{\theta}_q)=\sum_{x\in X}\boldsymbol{\theta}_p(x)\ln\frac{\boldsymbol{\theta}_p(x)}{\boldsymbol{\theta}_q(x)} \quad (7)$$

式中，$\boldsymbol{\theta}_m=\frac{1}{2}(\boldsymbol{\theta}_p+\boldsymbol{\theta}_q)$，$\boldsymbol{\theta}_p$和$\boldsymbol{\theta}_q$分别是第 $p$ 篇文档和第 $q$ 篇文档的概率分布。

### 3.2　层次聚类

层次聚类可以对文本聚类实行无监督的学习，能够在不事先指定聚类数的情况下对样本进行分析。本文采用基于矩阵、分步进行的凝聚层次聚类方法[35]对 1 472 门课程的描述性文本进行聚

类，并采用JS距离测量通过主题模型得到的文档–主题矩阵中文档之间的距离。

具体的算法步骤为:

**a**. 将每个课程都划分为一个簇；

**b**. 计算课程的文本文档两两之间的JS距离；

**c**. 当课程 $i$ 的文本文档和课程 $j$ 的文本文档之间的JS距离小于等于阈值时，开始第一步的合并；

**d**. 依次执行直到把所有满足该条件阈值的课程合并完；

**e**. 增大阈值，按上述方法继续合并，直到最后一步所有的课程合并为一个簇。

而对于合并过程中两个簇之间距离的取值，采用平均链的方法，取两个簇中两两课程的文本文档的JS距离的平均值。

图4为根据邓恩指数($DVI$)[37]对聚类的全过程裁剪之后的聚类过程图。黑色的横线是由$DVI$确定的距离阈值，即距离阈值为0.71时，聚类最有效。图4展现了最后15步的聚类过程，实心点分布的密集性表明了样本量的多少。样本容量为 $m$ 和 $n$ 的两个簇进行$DVI$的计算，$DVI$的计算公式为

$$DVI=\frac{\min\limits_{0<m\neq n<K}\left\{\min\limits_{\substack{\forall x_i\in\Omega_m\\ \forall x_j\in\Omega_n}}\left\{\left\|x_i-x_j\right\|\right\}\right\}}{\max\limits_{0<m<K}\ \max\limits_{\forall x_i,x_j\in\Omega_m}\left\{\left\|x_i-x_j\right\|\right\}} \tag{8}$$

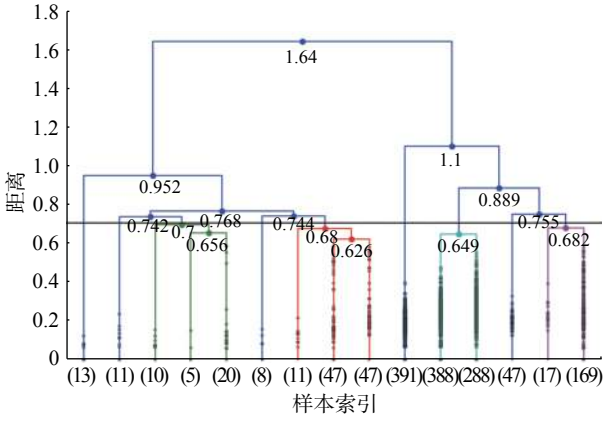


**图4 最后15步的聚类结果**

**Fig.4 Last 15 steps of the clustering result**

分子表示两个簇的最短距离(类间距离)，分母表示任意簇中的最大距离(类内距离)，$DVI$的值越大，即类间距离小的同时类内距离大，表示相应的聚类效果越好。

实验中的1 472门课程根据文本文档的JS距离进行聚类，可被划分为2类，分别为图4中距离为0.952和1.1时的合并点。从图4可以看出，随着距离阈值的增大，簇与簇之间的合并变得迟缓，距离的跨度越来越大。

从采用的凝聚层次聚类方法的原理看，距离阈值的不断增大才使得原本可能不被归为一个簇的样本归属于同一个簇。聚类的有效性是建立在合适的类内距和类间距的基础上。于是从最后一步聚类倒推，寻找最佳聚类效果的距离阈值。

将距离阈值小于0.71的聚类过程剪裁掉，得到最后9步的一个聚类结果，如图5所示。至此，完成了对1 472门课程的一个有效聚类。

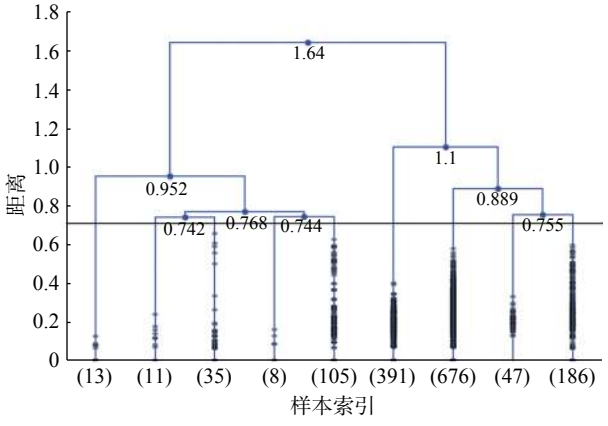


**图5 最后9步的聚类结果**

**Fig.5 Last 9 steps of the clustering result**

随后对照图5的结果，可视化每个簇内的所有课程对应的特征词信息。即对于图5中表现出的样本量分别为13, 11, 35, 8, 105, 391, 676, 47, 186的9个簇，使用关键词提取的方法筛选出簇内课程的特征词中频率最大的关键词作为该簇的主题；而对于横线(距离阈值大于0.71)以上的聚类点，分别再以涵盖不低于80%的特征词信息的词汇人工概括为新的主题，主题的拟定由来自7个不同专业背景的志愿者完成。最后，得到主题发现的结果如图6所示。

图6的主题发现结果表明，来自于日语社团的1 472门课程最终被划分为了9个小类，所属的主题依次为股票投资、语文素养、职业技能、音乐、实用外语、语言考试、语言基础、日语考级、日韩英。其中“股票投资”、“语文素养”等主题与“日语”并不相关。并且所有的主题最终归属于两个大的类别，一个是能力素养类，一个是语言素养类。语言类课程的主题并没有与类似“词汇”、“语法”、“作文”等反映知识内容

的词强相关，而是与“考试”、“基础”、“考级”等字样重合较多，所以更多地表达了学习者的学习目的。另一方面，从聚类簇中样本量的角度看，88.3% 的课程都是关于语言学习类的，这一特点符合日语社团这个数据背景。还值得关注的是，从课程数量的百分率上看，9 个子类中语言考试类和语言基础类的课程相对多而集中。

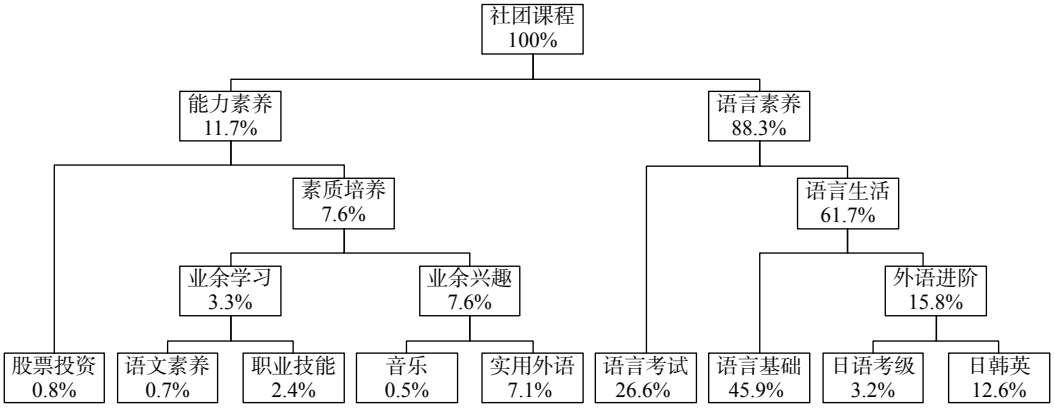


**图 6　主题发现结果**

**Fig.6　Results of topic discovery**

## 4　结束语

本文针对某互联网教育平台的课程进行了主题发现和聚类研究。实验以自主爬取的日语社团课程的详情介绍页面为研究基础，并通过文本预处理、主题表示、层次聚类等步骤完成了对 1 472 门课程潜藏在语义中的主题发现工作以及基于课程描述性文本的主题分布聚类。结果表明，1 472 门课程可被划分为主题不同的 9 个类别，并且这 1 472 门课程隶属于不同层次的共 16 个主题，从这些语义中发现的主题更多反映出了学习者的学习目的。而其中 88.3% 的课程主题都与语言学习类相关，这也契合了数据来自于日语学习社团这个大背景。

本文通过停用词库、自定义词典等大量的文本减噪处理，基于 LDA 的文本表示方法增强文本主题信息的使用，完全基于距离的探索性无监督层次聚类，对互联网教育平台课程潜藏在语义中的主题进行了发现，直接、客观地挖掘出了课程的特征以及学习者的学习目的。并且基于日语学习社团的课程所发现出的除日语学习以外的多样主题也能帮助平台更好地了解学习者选择课程背后隐含的主题关注点和兴趣点[38]所在，能够指导平台进行下一步课程的投放和课程的个性化推荐，以及为日后产品的设计提供科学的依据。

研究的不足主要为实验中预处理阶段的一些数据的清洗，以及自定义词典和停用词库的建立对实验结果的影响程度无法定量衡量。如何对文本进行更好的表示和更有效的聚类，值得进一步研究。

**参考文献：**

[1] 赵琦, 张智雄, 孙坦, 等. 主题发现技术方法研究[J]. 情报理论与实践, 2009, 32(4): 104–108.

[2] XU G X, QIU L R, LIU H F. Study on hot topic discovery from Chinese texts[J]. Journal of Digital Information Management, 2014, 12(4): 267–273.

[3] KAO A, POTEET S R. Natural language processing and text mining[M]. London: Springer, 2007.

[4] 程志, 黄荣怀. 文本挖掘及其教育应用[J]. 现代远距离教育, 2008(2): 71–73.

[5] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking[C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008: 3–12.

[6] 刘洪涛, 肖开洲, 吴渝, 等. 带舆论评价的引文网络构建与主题发现[J]. 情报学报, 2011, 30(4): 441–448.

[7] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书情报工作, 2014, 58(22): 70–77.

[8] 王连喜, 曹树金. 学科交叉视角下的网络舆情研究主题比较分析——以国内图书情报学和新闻传播学为例[J]. 情报学报, 2017, 36(2): 159–169.

[9] 王小华, 徐宁, 谌志群. 基于共词分析的文本主题词聚类与主题发现[J]. 情报科学, 2011, 29(11): 1621–1624.

[10] VACA C K, MANTRACH A, JAIMES A, et al. A time-based collective factorization for topic discovery and monitoring in news[C]//Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea: ACM, 2014: 527–538.

[11] ZHANG Z F, LI Q D. QuestionHolic: hot topic discovery and trend analysis in community question answering systems[J]. Expert Systems with Applications, 2011, 38(6): 6848–6855.

[12] GARGI U, LU W J, MIRROKNI V, et al. Large-scale community detection on Youtube for topic discovery and exploration[C]//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI, 2011.

[13] CHEN Y, AMIRI H, LI Z J, et al. Emerging topic detection for organizations from microblogs[C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland: ACM, 2013: 43–52.

[14] 路荣, 项亮, 刘明荣, 等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能, 2012, 25(3): 382–387.

[15] 罗春海, 刘红丽, 胡海波. 微博网络中用户主题兴趣相关性及主题信息扩散研究[J]. 电子科技大学学报, 2017, 46(2): 458–468.

[16] CATALDI M, DI CARO L, SCHIFANELLA C. Emerging topic detection on twitter based on temporal and social terms evaluation[C]//Proceedings of the 10th International Workshop on Multimedia Data Mining. Washington, DC: ACM, 2010: 4.

[17] 陈友, 程学旗, 杨森. 面向网络论坛的高质量主题发现[J]. 软件学报, 2011, 22(8): 1785–1804.

[18] 刘三女牙, 彭晛, 刘智, 等. 基于文本挖掘的学习分析应用研究[J]. 电化教育研究, 2016(2): 23–30.

[19] 阮光册. 基于 LDA 的网络评论主题发现研究[J]. 情报杂志, 2014, 33(3): 161–164.

[20] VICIENT C, MORENO A. Unsupervised topic discovery in micro-blogging networks[J]. Expert Systems with Applications, 2015, 15(17/18): 6472–6485.

[21] TAO W. Implementation of Chinese word segmentation algorithm based on bidirectional maximum matching in Police Affairs[J]. Application of Computer Technology, 2016: 153–155.

[22] BHARTI K K, SINGH P K. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering[J]. Expert Systems with Applications, 2015, 42(6): 3105–3114.

[23] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613–620.

[24] SALTON G, YU C T. On the construction of effective vocabularies for information retrieval[J]. Acm Sigplan Notices, 1973, 10(1): 48–60.

[25] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Singapore: ACM, 2009: 248–256.

[26] AGGARWAL C C, ZHAI C X. A survey of text classification algorithms[M]//AGGARWAL C C, ZHAI C X. Mining Text Data. Boston, MA: Springer, 2012: 163–222.

[27] KRESTEL R, FANKHAUSER P, NEJDL W. Latent dirichlet allocation for tag recommendation[C]// Proceedings of the 3rd ACM Conference on Recommender Systems. New York: ACM, 2009: 61–68.

[28] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1/2): 177–196.

[29] HOFMANN T. Latent semantic models for collaborative filtering[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 89–115.

[30] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993–1022.

[31] RIAHI F, ZOLAKTAF Z, SHAFIEI M, et al. Finding expert users in community question answering[C]// Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM, 2012: 791–798.

[32] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1): 5228–5235.

[33] 曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报, 2008, 31(10): 1780–1787.

[34] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423–1436.

[35] 王鹏, 高铖, 陈晓美. 基于 LDA 模型的文本聚类研究[J]. 情报科学, 2015, 33(1): 63–68.

[36] MAJTEY A P, LAMBERTI P W, PRATO D P. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states[J]. Physical Review A, 2005, 72(5): 052310.

[37] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of Cybernetics, 1973, 3(3): 32–57.

[38] 刘建国, 吴蓓蕾, 王超, 等. 基于改进用户兴趣点度量方式的推荐算法研究[J]. 情报学报, 2011, 30(11): 1158–1162.

（编辑：丁红艺）