面向社交媒体评论的子话题挖掘研究*

夏丽华[1] 韩冬梅[1,2]

(1.上海财经大学信息管理与工程学院 上海 200433;
2.上海市金融信息技术研究重点实验室 上海 200433)

摘 要:[**目的/意义**]在线用户在社交网络分享产品的体验,即便是同种产品的评论,往往包含不同的子话题(产品的不同方面)。面向在线评论的子话题挖掘能够分析参与者对产品的不同方面的关注及需求,为管理者提供更多的决策支持。[**方法/过程**]现有话题挖掘多采用分类、聚类、概率主题模型的方法,由于描述同一产品的文档往往十分相似,现有方法难以保证子话题的差异性。为此,将概率主题模型融合词共现关系,提出GPLSA方法,包括PLSA算法初步识别子话题、去除公共背景词、合并相似的子话题及更新子话题关键词等步骤。[**结果/结论**]知乎网站MOOCs数据集上的实验结果表明,GPLSA方法的主题凝聚性高于现有算法,能够有效提高子话题发现的质量。结合MOOCs子话题反馈的学习者需求,给出完善MOOCs管理的有效建议。

关键词:社交媒体;在线评论;话题识别;PLSA;词共现
中图分类号:G250.73 **文献标识码**:A **文章编号**:1002-1965(2020)04-0110-07
引用格式:夏丽华,韩冬梅.面向社交媒体评论的子话题挖掘研究[J].情报杂志,2020,39(4):110-116.
DOI:10.3969/j.issn.1002-1965.2020.04.016

Subtopic Mining Research Based on Social Media Reviews

Xia Lihua[1] Han Dongmei[1,2]

(1.School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433;
2.Shanghai Financial Information Technology Key Research Laboratory, Shanghai 200433)

Abstract:[**Purpose/Significance**]Online users share product experiences on social networks, and reviews of the same product often contain different subtopics (different aspect of the product). Subtopic mining for online reviews can analyze participants´concerns and their needs on different aspects of products, and provide more decision support for managers.[**Method/Process**]Existing methods for discovering topics are commonly based on classification, clustering and probabilistic topic model. However, as the documents describing the same product are often very similar, it is difficult for existing methods to ensure the diversity of subtopics. To tackle this problem, this paper proposes GPLSA method to integrate probabilistic topic model and word co-occurrence relationship, which includes subtopics detection by PLSA algorithm, removal of common background words, merging similar subtopics and updating subtopic keywords.[**Result/Conclusion**]The experiments on MOOCs dataset from zhihu website can prove that GPLSA has higher Topic Cohesion than the existing methods, and effectively improve the quality of subtopic discovery. According to the learners´needs for MOOCs subtopics feedback, effective suggestions for improving MOOCs management are given.

Key words:social media; online reviews; topics detection; PLSA; word co-occurrence

0 引 言

社交媒体评论包含各种有价值的信息[1-3],由于评论具有时效性、数量大的优势,有助于及时、全面地获取用户关注热点,从而了解用户的需求,满足当前企业决策的需要。当在线用户有需求时会对相关产品抱有期望,他们在选择产品时,就会特别关注产品的某些特定的方面,并通过社交媒体分享传播使用体验。这

收稿日期:2019-12-12 修回日期:2020-03-10
基金项目:全国教育科学规划教育部重点课题“新媒体环境对大学生学业情绪的影响及教学策略研究”(编号:DIA170369)研究成果之一。
作者简介:**夏丽华**(ORCID:0000-0003-4964-7422),女,1974年生,博士研究生,研究方向:数据挖掘、智能学习;**韩冬梅**(ORCID:0000-0001-6299-7246),女,1961年生,教授,博士生导师,研究方向:数据分析与挖掘、预测与决策、智能电子商务等。

些受用户关注的产品的方面统称为子话题,同种产品评论往往伴随着不同的子话题,每一个子话题都描述了产品的不同方面。传统的方法在分析产品的子话题时,多从文档角度或关键词的角度展开分析,采取分类、聚类、主题模型方法,由于描述同种产品的文档往往十分相似,现有方法难以保证子话题的差异性,影响子话题发现的效果。本文针对社交媒体评论的热点子话题进行研究,帮助管理者发现用户关注的问题及用户需求,在研究的过程中选用 MOOCs 领域的社交媒体评论作为研究语料。

1　话题识别技术相关工作

TDT(Topic Detection and Tracking)用于话题识别及其演化[4],关于话题发现的研究大致有以下几种思路:文档角度、主题模型建模和构建词图挖掘话题等。

1.1　基于文档的聚类和分类方法　最早的话题识别通常采用文档聚类方法[5],一组核心的关键词作为话题特征的代表,基于相似性的度量指标,包含这些关键词的文档具有更大的相似度,因此这些文档划分为同一话题;反之,不同话题的文档间的相似度则会较小。例如,利用 TF/IDF 计算基于关键词的特征值,用余弦相似度指标来对文档集合进行聚类,相似度较高的文档分到同一个结果簇内,划分好的每一个结果簇对应一个话题。

基于文本聚类技术研究者提出很多有效的算法,常用的代表性聚类算法有层次聚类 single pass incremental 聚类算法、分割聚类 K-means 算法等。其中 single pass incremental 聚类算法,对于每一篇新文档,计算新文档与已知话题模型向量之间的相似性,如果相似度大于某个阈值,则将该文档分配给相应的话题,否则将被视为一个新话题文档。K-means 算法是一种经典的分割聚类算法,K 值的预先设置有一定困难,初始聚类中心的选取比较敏感,这些因素极大地影响话题发现效果。

1.2　基于主题模型建模方法　主题模型,又称概率主题模型,PLSA(Probabilistic Latent Semantic Analysis)和 LDA(Latent Dirichlet Allocation)等主题模型广泛应用于文本挖掘、话题识别领域[6-7]。主题模型从词的角度入手,给定一组文档,生成基于词分布的一组语义相关话题,高概率的词被用来表示话题的特征,同时每一篇文档被视为这些话题分布的概率。主题模型研究迅速发展,并取得了丰富的研究成果[8-10]。但是这种主题模型最后生成的主题词为单个词的语义单元,丢失了原始文档中词间的语义关系,不可避免地,基于词的主题表示直接作为话题结果,会导致话题的可读性下降,难以被人们直观理解[4]。

为了改善主题语义的不明确问题,研究者在主题模型中引入先验知识,例如时间、社交网络数据及地理信息等[11-12],但这种先验知识仅局限于微博或科技文献特定格式,很难推广到更多领域。一些研究者[13]提出更为有效的 KBTM 主题模型(Knowledge-Based Topic Models),先验知识按半自动或自动方式提取,从而得到更高凝聚度的话题。Andrzejewski 等[13]提出 DF-LDA 模型,定义两种先验知识 must-links 和 cannot-links,其中 must-links 表示两个词属于同一个话题,cannot-links 表示两个词不应该属于相同话题。上述这些模型都假定加入的先验知识是正确的,忽略知识验证往往会导致不连贯的话题。为此,Wang 等提出在多个领域数据中动态挖掘先验知识[14],使用 PMI(Pointwise Mutual Information)等简单的度量方法排除不正确的知识,然后利用正确的知识指导模型的学习,从而产生更高质量的话题,这些研究方法使得特定领域子话题的发现效果依赖于丰富的多个领域数据,而且要求不同领域的数据要具备相似的特征。

1.3　基于图挖掘方法　基于词共现关系,研究者提出图分析方法进行话题挖掘。Sayyadi 等[15]提出 KeyGraph 方法,利用词共现关系将文本数据转换为词图结构,然后使用社区发现算法划分图,最后得到的每个社区代表一个话题,来自每个社区的关键字被视为话题的特征。这种构建词图结构的方法可以有效发现具有紧密共现关系的词,但是面临非线性趋势的计算复杂性的挑战,另外这种图挖掘方法只孤立计算两个词之间的关系,并没有充分考虑词的背景信息,即该词与多个邻居词的共现关系。

研究者提出将概率主题模型融合图挖掘的方法[4-5,16-17],综合考虑语义信息和词共现关系,从而避免不完整的信息检测,有效提高话题识别的质量。Zhang 等[5]提出 LDA-IG 方法,词作为图的结点,依据 LDA 主题模型的语义信息和词共现关系,计算图的每条边的权重,能够更有效地发现在不同领域的话题,但是并没有进一步讨论特定领域的子话题识别问题。Shams 等[17]提出 ELDA 方法,首先利用 LDA 模型生成初步子话题,然后基于词共现关系借助其它领域的相似子话题自动提取先验知识并迭代更新 LDA 模型,这些来自多领域的先验知识虽然能够帮助识别不同领域的共有特征,但是难以保证特定领域的子话题的区分度。周楠等[4]提出 ET-TAG 方法,使用 PLSA-BLM 生成初步子话题,借助两个词之间的共现次数更新关键词,从而获得高质量的子话题,但是该方法并没有充分考虑词的背景信息。

2　GPLSA 话题挖掘方法

基于已有研究成果[4-5,16-17],本文将词共现关系融

合到概率主题模型,提出 GPLSA 方法,处理同种产品评论的子话题的差异性问题。本文方法与已有研究成果有以下两点区别:一是不需要借助多领域的先验知识,避免来自多领域数据的不正确知识以及由于多领域数据共有特征造成的子话题区分度不明显的限制性;二是通过词共现关系的定义,不仅考虑两个词的关系,还充分考虑词的背景信息,即该词与多个邻居词的复杂共现关系。本文采用的 GPLSA 方法包括五个主要步骤:使用 PLSA 算法生成初步子话题及相应关键词;定义词共现关系;公共背景词的噪声处理;合并相似冗余的子话题;子话题关键词更新。GPLSA 话题识别方法的具体实现步骤如算法 1 所示:

算法 1 GPLSA 子话题识别方法

Input: $D = \{d_1, d_2, \cdots, d_n\}$,来自社交媒体的数据集;Output: $T = T_1, T_2, \cdots, T_k$,子话题集合;

keywords $(T_i) = \{w_1, w_2, \cdots, w_v\}$,每个子话题 T_i 的关键词

1:预处理过程;

2:#采用 PLSA 生成初步子话题过程

3:执行 PLSA 算法生成语义子话题;

4:为每个子话题选择高概率的关键词;

5:#定义词图关系

6:基于条件概率定义词共现关系,具体定义如 Eq3.1 所示;

7:#去除公共背景词

8:For 每个词 $w \in T_i$

9:定义公共背景词权重,计算背景词与所有话题词簇的关系值,具体定义如 Eq3.2 所示;

10:End for

11:按着背景词权重大小排序,删除最高的前 K 个词;

12:#相似子话题合并

13:For 每个子话题 $T_i \in T$

14:For 每个子话题 $T_j \in T - T_i$ 15:计算两个子话题 (T_i, T_j) 之间的相似度,具体定义如 Eq3.3 所示;

16:End for

17:End for

18:将相似度最高的前 K 个值对应的子话题进行合并;

19:#更新子话题关键词

20:For 每个词 $w \in T_i$

21:计算每个关键词的权重大小,具体定义如 Eq3.4 所示;

22:End for

23:根据关键词权重选择最高的 K 个关键词。

2.1 预处理及基于 PLSA 的子话题识别 中文文本数据的预处理工作主要包括分词、词性标注、过滤停用词及特征抽取[17]。使用分词工具,根据词性标注(Part-of-Speech tagging)对帖子进行切词处理,考虑到动词的高嗓音,只保留名词、动名词、名词短语,生成新的文本语料库。所有帖子信息都被切分成句子,预处理后得到的数据集组成形式定义如下:

(第 1 行)词 1、词 2、词 3 项、词 4、词 5

(第 2 行)词 2、词 3、词 6、

(第 3 行)词 1、词 2、词 4 项、词 7

……

首先采用 PLSA 概率主题模型[6]进行子话题的初步识别,生成基于词分布的语义相关子话题,高概率的词被用来表示子话题的特征。为研究方便,设文档集合 D ,主题集合 Z ,词集 W , $P(d_i)$ 是从文档集合 D 选中文档 d_i 的概率, $P(z_k \mid d_i)$ 是主题 z_k 在文档 d_i 中的出现概率, $P(w_j \mid z_k)$ 是词 w_j 在主题 z_k 中出现的概率。PLSA 模型是针对文本中隐含的主题进行建模,通过 $P(z_k \mid d_i)$ 、$P(w_j \mid z_k)$ 两层概率对整个样本空间构建模型,即从文档 d_i 到词 w_j 的生成过程,具体过程如图 1 所示。

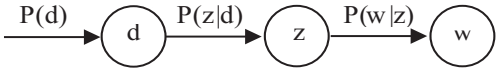

图 1 PLSA 概率图模型的示意图

图 1 所示的“文档-词项”的生成过程实际上是计算文档中每个词的生成概率:

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i) = P(d_i)\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \quad (1)$$

其中 $P(d_i)$ 可事先计算求出, $P(w_j \mid z_k)$ 和 $P(z_k \mid d_i)$ 未知,是本研究要估计的参数值。由于词和词之间是相互独立的,文档和文档之间也是相互独立的, $n(d_i, w_j)$ 表示词项 w_j 在文档 d_i 中的词频,所以整个文本语料库中词的生成概率为:

$$P(W \mid D) = \prod_{d_i \in D}\prod_{w_j \in W} P(d_i, w_j)^{n(d_i, w_j)} \quad (2)$$

采用 EM 算法估计 PLSA 模型的未知参数 $P(w_j \mid z_k)$ 和 $P(z_k \mid d_i)$,估计出的参数值如下:

$$P(w_j \mid z_k) = \frac{\sum_{d_i \in D} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{d_i \in D}\sum_{w_j' \in W} n(d_i, w_j')P(z_k \mid d_i, w_j')} \quad w_j \in W, 1 \leqslant k \leqslant K \quad (3)$$

$$P(z_k \mid d_i) = \frac{\sum_{w_j \in W} n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(d_i)} \quad d_i \in D, 1 \leqslant k \leqslant K \quad (4)$$

其中 $n(d_i)$ 表示文档 d_i 中词的总数,即 $n(d_i) = \sum_{w_j \in W} n(d_i, w_j)$

2.2 词关系的定义及词的分类

2.2.1 词共现关系的定义 关系密切的词往往具有明显的共现关系,并且形成一个共现的词簇。任意两个词 w_i 与 w_j 间的共现关系[5, 15, 17]如公式(5)所示:

$$R(w_i, w_j) = p(w_i \mid w_j) = \frac{\sum_{d \in D(w_i, w_j)} 1/|d|}{\sum_{d \in D(w_j)} 1/|d|} \quad (5)$$

其中 $|d|$ 是文档长度, $D(w_j)$ 定义为包含词 w_j 的

文档集合,$D(w_i,w_j)$ 定义为既包含词 w_i 又包含词 w_j 的文档集合;在两个词关系的度量中要考虑文档长度 $|d|$,因为如果一个词出现在一个长文档中,那么该词与其它词共现的概率会更大,通过文档长度 $|d|$ 标准化公式(5)中的条件概率。在词共现关系 R 中,如果两个词在不同的文档中共现次数更多,并且词 w_j 在子话题中是一个更具体明确的词,那么这两个词之间的关系就会更强。

2.2.2　核心词和背景词的定义　PLSA 算法生成初步子话题及相应关键词,这些关键词分为核心词和背景词[4-5,17]。背景词,又称为桥词(bridge word),与多个子话题词簇有着最密切关系的词,因此它不适合代表任何子话题,并且这些词应该从子话题关键词中剔除[5,17]。核心词(hub word)是与一个子话题词簇密切相关的词,它经常与该子话题中的其它关键词共同出现在文档中,因此核心词是一个子话题关键词中的最佳代表。例如,如图 2 所示,在子话题词簇 T_1 中,词 w_1 是核心词,词 w_2 是背景词。

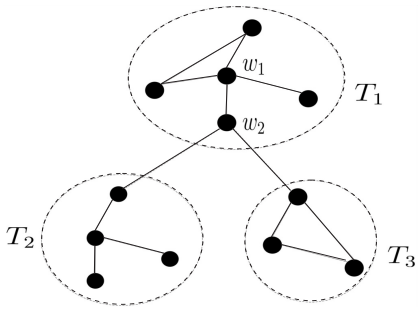

图 2　词与词关系示意图

根据背景词和核心词的特征,识别每个子话题的核心词要注意如下准则:核心词与一个子话题的其它关键词频繁共现;背景词与每个子话题词簇相联接,影响子话题核心词抽取的质量,因此要从子话题核心词中删除;每个核心词唯一代表一个子话题,不能与其它子话题核心词冲突。遵循上述准则,基于词共现关系的定义,后续进一步去除背景词,识别每个子话题的核心词。

2.3　**公共背景词的噪声处理**　大量的高频背景词出现在各个文档之中,背景词影响核心词抽取的质量,降低子话题的区分度[4,17]。现有概率主题模型无法有效去除背景词,本研究基于词共现关系,采用公式(6)计算背景词与所有子话题词簇的关系值,识别每个子话题中的公共背景词。

$$BG(w_i)=\sum_{T_j\in T-T_i} log(edg(w_i,T_j))\times \sum_{w_j\in T_j,w_j\neq w_i} R(w_i,w_j)\quad w_i\in T_i \tag{6}$$

其中,词共现关系 R 按公式(5)计算,$edg(w_i,T_j)$ 代表词 w_i 与子话题 T_j 中关键词有共现关系的词的个数,为了准确推断词 w_i 与该子话题 T_j 的关系,同时考虑下面两个因素:词 w_i 与子话题 T_j 中所有关键词关系的和;词 w_i 与子话题词簇 T_j 共现的词的个数,词 w_i 与子话题 T_j 有关联的词越多,代表词 w_i 与子话题 T_j 的关系更紧密。

计算每个词的 $BG(w_i)$ 值,按降序排列,去除每个子话题内最高的前 K 个背景词,Algorithm 1 中第 7~11 行描述了有效去除公共背景词的处理过程,从而提高子话题关键词抽取的质量。

2.4　**相似子话题合并**　利用 PLSA 模型,可以在每个子话题下得到不同的词频分布,由于公共背景词的剔除,出现重复子话题的可能性增大[4],因此发现相似的子话题并对其进行合并可以去除冗余,可以改善子话题发现的效果。现有研究基于 KL 散度(KL Divergence)判断两个子话题的词频分布差异[4],KL 散度值越小,两个子话题分布差异越小,两个子话题应该合并。KL 散度计算是基于两个子话题下前 K 个关键词的词频分布,由于 PLSA 模型初步发现的子话题关键词质量不高,导致 KL 散度方法并不适合度量子话题相似度。

通常相似子话题具有更多的相关的核心词[17],因此基于词共现关系定义两个子话题的关系,如公式(7)所示。

$$TR(T_i,T_j)=\sum_{w_i\in hub(T_i)} log(edg(w_i,T_j))\times \sum_{w_j\in T_i,w_j\neq w_i} R(w_i,w_j) \tag{7}$$

其中词共现关系 R 按公式(5)计算,$hub(T_i)$ 只包括子话题 T_i 下的核心关键词,由于去除子话题下的公共背景词,保证核心关键词的抽取质量。文献[17]只计算了词 w_i 与子话题 T_j 中所有关键词关系和,本研究引入 $edg(w_i,T_j)$,同时考虑词 w_i 与子话题词簇 T_j 共现的词数。考虑到 $TR(T_i,T_j)$ 具有不对称性,可采用以下公式计算两个子话题之间的相似度 $SIM(T_i,T_j)=(TR(T_i,T_j)+TR(T_j,T_i))/2$。按降序排列 $SIM(T_i,T_j)$ 值,将最高 K 个值对应的子话题进行合并,Algorithm 1 中第 12-18 行描述相似子话题合并过程,避免重复子话题的出现,从而使子话题更为合理。

2.5　**子话题关键词更新**　目前大多数常见的子话题特征词抽取方法是基于统计的角度进行选择,例如,TF-IDF、词频等,但是当关键词与子话题词簇内多个词具有紧密关系时,其更适合作为子话题的最佳代表,因此具体的子话题关键词权重计算方法如公式(8)所示。

$$KEY(w_i)=log(edg_\mu(w_i,T_i))\times\sum_{w_j\in T_i-w_i} R(w_i,w_j) \tag{8}$$

其中,词共现关系 R 按公式(5)计算,$edg_\mu(w_i,T_i)$ 代表词 w_i 与子话题词簇 T_i 共现的词的个数,共现的关

键词可以保证词簇内部的语义关联度更为紧密[4]，本研究μ取值为3，子话题的关键词至少和3个词有共现关系。将每个子话题下的关键词按照$KEY(w_i)$降序排列，$KEY(w_i)$值越高说明这个词在当前子话题下越有代表性，Algorithm 1中第19~23行描述子话题关键词更新过程。

3 实验评估

3.1 实验数据集 知乎平台（WWW.ZHIHU.COM）是知识社交平台，用户分享着彼此的知识、经验和见解。MOOCs学习者倾向于在社交媒体共享信息和相互交流[18]，尤其对于年轻一代来说，社交媒体是交流、寻求信息和社会参与的主要手段，因此知乎“MOOCs”话题吸引很多学习者的讨论，这种社交媒体也成为学习者传播MOOCs资源的新途径。按提供课程的平台，知乎“MOOCs”话题下细分为学堂在线、中国大学MOOC、Coursera、edX、优达学城（Udacity）及MOOCs学院相应子话题。本研究通过python编程实现“MOOCs”话题帖子的采集，获得从2011-2018年共7年时间发布“MOOCs”话题下的所有帖子，包括1963个问题主题帖，每个问题帖下有若干个回答帖及评论帖，共收集29523个帖子。

3.2 评估标准 主题模型是一种无监督数据分析方法，通过评估标准不仅可以衡量话题发现的效果，还能优化模型参数，使得话题识别结果和人工标记更为匹配，识别的话题具有更好的可解释性。近年来，许多主题模型采用主题凝聚性度量方法（Topic Coherence measure）评估每个模型生成的话题[14,17]，主题凝聚性度量方法由Mimno[19]提出，其评估结果和专家人工标记一致，能够更好地发现有意义或高凝聚度的话题。主题凝聚性指标定义如下：

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{CODF(v_m^{(t)},v_l^{(t)}) + 1}{DF(v_l^{(t)})} \quad (9)$$

其中$V^{(t)} = (v_1^{(t)},\cdots,v_M^{(t)})$是子话题$t$下最具有代表性的$M$个词，$DF$代表包含词$v_l^{(t)}$的文档数，$CODF$代表包含词$v_l^{(t)}$和词$v_m^{(t)}$的文档个数，为了避免对零取对数，所以在分子上加1。主题凝聚性C值越高，代表生成话题质量越高，并且话题具有更好的可解释性。

3.3 实验结果 LDA是一种经典的无监督主题模型[7]，广泛应用于话题或子话题发现的相关研究中；ELDA方法要求有共同特征的多领域数据，当数据集只是某个特定领域时，相当于执行LDA方法，并不能改进子话题识别效果；ET-TAG方法基于PLSA进一步优化子话题结果，借助词共现次数有效提高子话题关键词发现质量。基于此，本文选择LDA、ET-TAG方法作为基准算法，与GPLSA方法作对比，采用主题凝聚性指标来评价算法性能。实验结果如表1所示，主题凝聚性值越高，代表生成子话题质量越高，ET-TAG方法基于概率主题模型考虑了两个词的共现关系，因此其性能好于传统LDA算法；相对于基准方法，本研究提出的GPLSA方法生成的子话题质量最高，在经过PLSA算法初步识别子话题后，在公共背景词噪音处理、合并相似冗余子话题、更新关键词等优化步骤中，都充分考虑词的背景信息，即该词与多个邻居词的共现关系，显著提高子话题发现的效果，从而有效保证子话题的差异性。

表1 不同算法在MOOCs子话题的发现效果

算法	主题凝聚性
LDA	-1972.98
ET-TAG	-1966.32
PLSA	-2043.91
PLSA+P1	-1951.85
PLSA+P1+P2	-1905.04
GPLSA（PLSA+P1+P2+P3）	-1881.01

注：P1是去掉公共背景词，P2是合并相似子话题，P3按词共现关系更新关键词权值

3.4 参数灵敏度评价 PLSA算法初步识别15个子话题，每个子话题抽取30个最可能的词进行特征描述，采用主题凝聚性指标度量不同目标参数值，GPLSA参数灵敏度的评价如图3所示。在图3（a）、（b）中，当去除公共背景词的个数、合并相似子话题的个数取很小的值时，主题凝聚性值急剧下降，这是因为在每个子话题中选择了不恰当的公共背景词作为关键词，

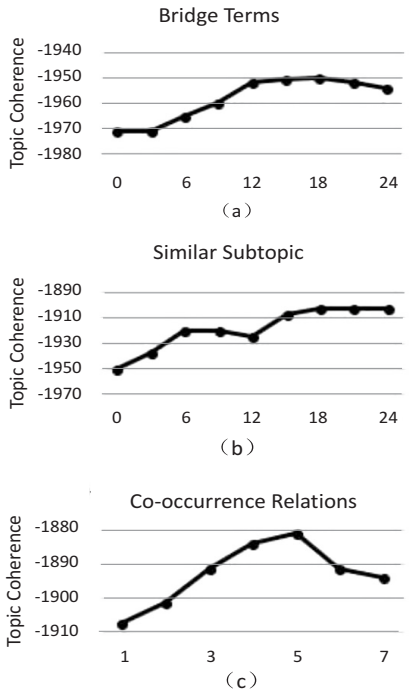

图3 GPLSA参数灵敏度

同时存在重复的子话题,这些步骤影响关键词抽取的质量;随着公共背景词及冗余子话题的去除,主题凝聚性值增大,并趋向平稳,子话题质量得到提高。在图 3(c)中,随着关键词共现阈值的增大,主题凝聚性值先增大后降低,这说明关键词共现次数逐步增大,有助于改善子话题的凝聚度,但是当关键词共现次数过高时,由于共现关系要求提高,导致子话题下的关键词个数变少,影响关键词抽取的质量。

3.5　**子话题结果的讨论**　MOOCs 实验语料总计 29 523篇文档 9 000 余词汇,基于该实验语料采用 GPLSA 方法进行 MOOCs 学习者话题挖掘,依次经过一系列的优化处理后,无监督地生成若干子话题和关键词,MOOCs 话题识别结果如表 2 所示。

表 2　话题发现结果

ID	子话题标签	子话题关键词	评论样本
1	优质平台	平台,大学,course,online,open,学堂在线,Stanford,AI,公司,science	学堂在线的数据结构,用的课本就是清华大学 cpp 版的数据结构
2	实践课程	基础,编程,python,入门,部分,语言,算法,数学,计算机,设计	由密歇根大学开设的零基础 Python 入门专项课程
3	老师信息	老师,需要,学校,上课,了解,史记,秦始皇,机器学习,讲课,台大,吕世浩	台大吕世浩老师的《秦始皇》《史记》
4	证书	认证,付费,提供,申请,signature track,选择,值得,注册,旁听,拿到	很多课程可以加入 signature track
5	互动交流	作业,东西,知道,大学,课堂,同学,讨论,网络,自学,答案	最后就是互动的质量了,老师或者助教最好能多在课程的论坛里和学生一起讨论
6	学习困惑	视频,东西,资源,找到,不了,电脑,建议,概念,谢谢,出现	怎么我电脑看不了视频 YOURTUBE 的 Coursera
7	MOOCs 运营	知识,工作,在线教育,门槛,微专业,运营,制作,模式,经验	MOOCAP 课程设计团队有来自清华大学数学系、物理系、化学系和生命学院的教授

MOOCs 发帖者中有求助者、援助者,分析 MOOCs 话题发现结果,第一类至第四类子话题涉及 MOOCs 平台、课程、授课老师以及证书方面,反映了 MOOCs 学习者对这一类信息的关注偏好。MOOCs 平台为学习者提供了开放的自主学习环境,同时学习者也需要独立完成相关学习活动,例如制定自己的学习目标,寻找合适的学习资源,安排自己的学习进度,在传统课堂这些学习活动完全由教师负责确定,由于学习者能力的差异,面对丰富的海量资源,很多求助者不清楚应该学习哪些内容,他们需要借鉴别人的经验,做出类似的学习选择。因此从学习者关注点出发设计课程,有助于满足用户需求,从而进一步改善学习者体验。第五、六类子话题涉及学习者的情感方面,他们面对学习中的各种困惑,渴望与老师、伙伴的互动交流,反映了 MOOCs 平台教学支持服务并没有满足 MOOCs 学习者的需求,MOOCs 平台互动沟通存在障碍,学习者只能通过社交平台寻找帮助,也说明学习者需要更多的情感支持与激励,缩短在线学习的距离感,因此改进 MOOCs 教学平台支持服务也是亟待要解决的问题。第七类子话题涉及 MOOCs 运营及制作问题,这说明社交媒体已经成为 MOOCs 设计团队获取学习者反馈的途径,以期改善教学质量和平台支持服务。这些研究结果可以 MOOCs 管理者及设计者提供有益的启示。

4　结　论

本文面向社交媒体对子话题挖掘展开研究,基于主题模型融合词共现提出 GPLSA 方法,有效提高子话题识别的质量。根据模型,采集知乎 MOOCs 帖子,算法评估结果表明, GPLSA 方法相比于 LDA 方法及 ET-TAG 方法具有更好的性能,可以获得更高凝聚度的子话题。GPLSA 方法拓展子话题识别方法,该方法应用于社交媒体数据的挖掘,突破了传统调查和访谈样本数量的限制,不仅能够为教育管理者提供更多的决策支持,也对诸如在线电子商务等其它领域也具有重要的管理意义。

参 考 文 献

[1] Bolat E,O′ Sullivan H.Radicalising the marketing of higher education:Learning from student-generated social media data [J]. Journal of Marketing Management,2017(2):1-22.

[2] 安　璐,王小燕,李　纲.恐怖事件情境下微博信息组织与关联可视化[J].情报杂志,2019,38(12):157-163,199.

[3] 吴志远.喧哗与躁动:南海争端事件中的网络社会心态研究——基于南海仲裁案的网络舆情分析[J].情报杂志,2018,37(5):103-110.

[4] 周　楠,杜　攀,靳小龙,等.面向舆情事件的子话题标签生成模型 ET-TAG[J].计算机学报,2018,41(7):1490-1503.

[5] Zhang C,Wang H,Cao L,et al.A hybrid term-term relations analysis approach for topic detection[J].Knowledge-Based Sys-

tems,2016,93:109-120.
[6] Hofmann T.Probabilistic latent semantic indexing[C].Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.Berkeley,California,USA,1999:50-57.
[7] Blei D M,Ng A Y,Jordan M I.Latent dirichlet allocation [J].Journal of Machine Learning Research,2003(3):993-1022.
[8] 韩忠明,张梦玫,李梦琪,等.面向复杂主题建模的流式层次狄里克雷过程[J].计算机学报,2019,42(7):1539-1552.
[9] 石 敏,刘建勋,周 栋,等.基于多重关系主题模型的Web服务聚类方法[J].计算机学报,2019,42(4):820-836.
[10] 马 红,蔡永明,信风芹.关联突发权重的主题模型:以共享单车法律问题学术文献为例[J].情报杂志,2019,38(4):181-186,193.
[11] 刘少鹏,印 鉴,欧阳佳,等.基于MB-HDP模型的微博主题挖掘[J].计算机学报,2015(7):1408-1419.
[12] 孙 锐,郭 晟,姬东鸿.融入事件知识的主题表示方法[J].计算机学报,2017,40(4):791-804.
[13] Andrzejewski D,Zhu X,Craven M.Incorporating domain knowledge into topic modeling via dirichlet forest priors[C].Proceedings of the International Conference On Machine Learning.United States:Morgan Kaufmann Publishers,2009:25-32.
[14] Wang S,Chen Z,Liu B.Mining aspect-specific opinion using a holistic lifelong topic model [C].International Conference on World Wide Web.2016:167-176.
[15] Sayyadi H,Raschid L.A graph analytical approach for topic detection [J].Acm Transactions on Internet Technology,2013,13(2):1-23.
[16] Colace F,Casaburi L,Santo M D,et al.Sentiment detection in social networks and in collaborative learning environments [J].Computers in Human Behavior,2015,51:1061-1067.
[17] Shams M,Baraani-Dastjerdi A.Enriched LDA (ELDA):Combination of latent dirichlet allocation with word co-occurrence analysis for aspect extraction [J].Expert Systems with Applications,2017,80:136-146.
[18] Veletsianos G,Collier A,Schneider E .Digging deeper into learners´ experiences in MOOCs:Participation in social networks outside of MOOCs,notetaking and contexts surrounding content consumption [J].British Journal of Educational Technology,2015,46(3):570-587.
[19] Mimno D,Wallach H M,Talley E,et al.Optimizing semantic coherence in topic models[C].Proceedings of the Conference on Empirical Methods in Natural Language Processing.Edinburgh,United Kingdom:Association for Computational Linguistics,2011:262-272.

(责编/校对:王平军)

(上接第89页)

略[J].湖北社会科学,2016(12):54-59.
[3] 张 静.“一带一路”战略下的湖北发展举措研究[J].湖北社会科学,2016(3):67-73.
[4] 陈继勇,蒋艳萍,陈大波.构建内陆开放新高地——基于湖北深度融入“一带一路”建设视角[J].江汉论坛,2018(1):34-39.
[5] 张金星.关于抢抓“一带一路”发展机遇助推湖北企业“走出去”的实践与思考[J].武汉金融,2016(6):20-23.
[6] 周谷平,阚 阅.“一带一路”战略的人才支撑与教育路径[J].教育研究,2015(10):4-22.
[7] 刘殿刚,毛和荣,顾 赤.“一带一路”战略视野下湖北中医药文化对外传播研究[J].时珍国医国药,2015,26(8):1961-1963.
[8] 黄芙蓉.“万里茶道”申遗与区域发展传播路径研究——湖北融入“一带一路”的战略思考[J].学习与实践,2016(11):129-134.
[9] 陈 潭.公共政策变迁的理论命题及其阐释[J].中国软科学,2004(12):10-17.
[10] 王 骚,靳晓熙.动态均衡视角下的政策变迁规律研究[J].公共管理学报,2005(4):31-35,97-98.
[11] 陈 芳.政策扩散、政策转移和政策趋同——基于概念、类型与发生机制的比较[J].厦门大学学报(哲学社会科学版),2013(6):8-16.
[12] 王宏新,付 甜,张文杰.中国易地扶贫搬迁政策的演进特征——基于政策文本量化分析[J].国家行政学院学报,2017(3):48-53,129.
[13] 黄 萃,赵培强,李 江.基于共词分析的中国科技创新政策变迁量化分析[J].中国行政管理,2015(9):115-122.
[14] 黄凯丽,赵 频.“一带一路”倡议的政策文本量化研究——基于政策工具视角[J].情报杂志,2018,37(1):53-58,46.
[15] Hogwood W Brian,Peters B.Guy,policy dynamics [M].New York:St.Martin´s Press,1983.
[16] Jack L Walker.The diffusion of innovations among the American States[J].The American Political Science Review,1969,63(3):880-899.
[17] 张 克.西方公共政策创新扩散:理论谱系与方法演进[J].国外理论动态,2017(4):35-44.
[18] Frances S Berry,William D Berry.Stata lottery adoptions as policy innovation:An event history analysis[J].The American Political Science Review,1990,84(2):395-415.
[19] 田志龙,陈丽玲,顾佳林.我国政府创新政策的内涵与作用机制:基于政策文本的内容分析[J].中国软科学,2019(2):11-22.
[20] 白 洁.湖北应加快对接长江经济带建设[J].学习月刊,2015(11):43-44.

(责编/校对:王平军)