分类号_____ 论文选题类型 非师范类应用研究

U D C _____ 编号_____

华中师范大学

本科毕业论文（设计）

题 目 社会化问答平台用户兴趣演化研究
——以 Quora 为例

学 院 信息管理学院

专 业 信息管理与信息系统

年 级 2015 级

学生姓名 张鹤铭

学 号 2015214533

指导教师 陈烨

二〇一九年五月

华中师范大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下独立进行研究工作所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

学位论文作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士学位论文评选机构将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密　□ ，在_____年解密后适用本授权书。

2、不保密　□。

（请在以上相应方框内打“ √ ”）

学位论文作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

目 录

内容摘要.....1
关 键 词.....1
Title.....1
Abstract.....1
Key words.....2
1.引言.....3
2.相关研究.....4
2.1 主题挖掘相关研究.....4
2.2 主题演化相关研究.....5
3.研究框架.....5
3.1 研究技术路线.....5
3.2 研究基础理论.....6
3.2.1 BTM 理论.....6
3.2.2 频次计量法.....8
4.数据采集与预处理.....9
4.1 数据采集.....9
4.2 数据可视化与数据预处理.....10
5.用户兴趣主题识别.....11
5.1 兴趣主题挖掘模型构建.....11
5.2 兴趣主题识别结果.....14
6.用户兴趣演化分析.....15
6.1 兴趣主题热度计算.....15
6.2 兴趣演化趋势分析.....16
7.结束语.....17
附录.....19
参考文献.....38

内容摘要：[目的]通过分析社会化问答平台的用户兴趣及其演化特征，进而指导Quora中Film and Television话题下的个性化推荐和广告投放。[方法]对爬虫得到的数据按周期划分并进行清洗，结合用户行为数据对周期内的问答文本进行赋权并采用BTM（Biterm Topic Model）进行主题挖掘并对主题进行识别。最后通过热力图绘制了主题演化趋势图并进行了分析。[结果]结合用户行为数据的赋权问答文本的coherence值更高，其主题挖掘效果更佳。其中，Comedy（喜剧），SciFi（科幻），Criminal（犯罪）题材的影视剧出现的频率较高。[局限]本文没有对后续的主题演化趋势进行建模分析，从而得到可以量化的规律模型。[结论]通过结合用户行为数据对问答文本进行赋权，改进了BTM主题挖掘的效果；并且对Quora中Film and Television等影视剧话题的个性化推荐和广告投放提出了相应建议。

关 键 词：社会化问答平台　BTM(Biterm Topic Model)　用户兴趣　主题挖掘　主题演化

Title: Study on the Evolution of Users' Interests on Social Q&A Community: A Case Study of Quora

Abstract: [Purpose] This paper analyzes the users' interests and their evolution characteristics of on the social Q&A Community, so as to guide the personalized recommendation and advertising on Quora Film and Television topics.[Methods] The data obtained by crawler were divided and cleaned according to the period, and the question and answer text within the period was endowed with user behavior data, and BTM (Biterm Topic Model) was adopted to conduct Topic mining and Topic identification.Finally, the trend chart of theme evolution is drawn and analyzed by means of thermal map.[Results] The coherence value of Q&A text with user behavior data is higher, and the topic mining effect is better.Among them, comedy, SciFi, criminal theme of movies and TV plays appear more frequently.[Limitations] This paper did not conduct modeling analysis on the subsequent theme evolution trend, so as to obtain a quantifiable rule model.[Conclusion] The effect of BTM topic mining is improved by empowering the question answering text with user behavior data.In

addition, suggestions were made on the personalized recommendation and advertising of Quora Film and Television and other topics.

Key words: Online Social Q&A Community BTM(Biterm Topic Model) User Interest Topics Mining Topics Evolution

1 引言

社会化问答平台是依托 Web 2.0 发展起来的新兴的知识共享平台，这种平台没有明确的组织结构，允许用户根据自己的需求随时提出问题或解答[1]。社会化问答平台的核心是用户参与，用户的角色是模糊的，没有明确的界定，他们既可以是信息的产生者，也可以是信息的消费者。并且，随着互联网的普及，在线问答社区已经出现爆炸式地成长，成为重要的知识共享平台，例如美国的 Quora 和 StackOverflow，以及中国的知乎[2]。在线问答社区提供了一个平台，通过发布和回答问题来创建和分享知识。平台上的话题通过提问、意见、体验、评论等形式涵盖了广泛的主题。通过发布(或搜索)问题并收集答案，用户可以快速学习和采用与他们所关注的领域相关的知识，其中大部分是领域专家的第一手答案[3]。由此，人们将越来越习惯于从网络社交平台上获得其感兴趣的话题或消息，这也就使得用户个性化推荐成为重要的一项网络服务。因此，分析识别社会化问答平台用户的潜在兴趣话题并推荐相关信息具有重大的研究价值[4]。但是，一方面，目前针对社会化问答平台主题挖掘的精准度仍有待提高，另一方面，对于社会化问答平台某个话题下用户的兴趣演化的研究目前还相对较少。

目前，Quora 是国外较为盛行的社会化问答平台，于 2009 年成立，在创建之初采用邀请注册制，吸引了各行各业的精英人士[5]。一年以后正式对公众开放，用户可以使用社交网站账号登录 Quora，以此防止通过搜索引擎就能搜索到相应的内容。Quora 作为一个在线知识共享社区，其贯彻了 IT 扁平化的思想，通过在线社会网络将人们的实际社会生活映射到互联网上，人们在问答网站 Quora 合著内容并找到问题的满意答案[6]。其允许用户协同编辑和回答问题的方式汇聚了大量的问题和答案，给挖掘平台内用户兴趣和分析演化规律提供了良好的研究情境。

于是，本文选取当下流行的社会化问答平台 Quora，并希望从话题的角度切入，借助主题挖掘模型对 Quora 的用户兴趣进行识别。在此基础上，本文考虑了结合话题中问答的观看数等辅助数据给文本数据增加权重，加权后的文本数据可能会优化主题分类的结果。最后，本文分析处理出了话题下的主题演化趋势并采用内容分析法对演化趋势进行了分析，这对于了解当下用户的讨论热点以及进一步跟踪社会化问答平台内的热点话题有很大的帮助，对用户个性化推荐、广告投放和舆情监督具有很好的指导意义。

2 相关研究

目前，针对社会化问答平台的研究分布很广，相关研究分布在探究知识质量的增长模式[3]、用户互动机制研究[7]、意见领袖的产生[8]和检测[9]、观点提取[10]和兴趣偏好挖掘[11]等诸多方面。

2.1 主题挖掘相关研究

在社会化问答平台的主题挖掘和兴趣偏好获取的方面上，Jiang[12]等是通过赞同数（Upvote）等来研究气候变化（Climate Change）话题下的用户兴趣偏好。在这里，本文跟希望通过主题挖掘的方式来获取某个话题下的用户兴趣并研究其演化趋势。目前更多研究是通过对话题下的主题进行识别和热度分析的方法有从基于用户文本信息来展开的，例如由 Blei[13]等人在 2003 年提出的隐式狄利克雷分布（LDA）模型。针对不同的情境，一些学者对模型进行了改进并尝试应用于各种短文本主题挖掘领域，经典的改进模型包括 ATM[14]、Twitter-LDA[15]、Labeled-LDA[16]等模型。然而，LDA 模型忽视了文档之间文本主题的相关性，对于像 Quora 这类社会化问答平台的短文本，其稀疏的共现模式会导致较为严重的数据稀疏问题[17]。一个简单且流行的方式是 Weng[18]等提出的在训练 LDA 之前将单个用户发布的 tweets 聚合到一个文档中，将短文本重新聚合成为伪长文本。而 Zhao 等[19]人提供了另一种解决方案，即假设一篇短文本文档只有一个主题，这也导致模型丧失了在一篇文档中捕获多个主题的能力。于是，Yan[20]等提出了扩大假设的方式，即认为，既然短文本文档之间存在相关性，且会因为传统主题模型建模时不能很好地考虑短文本文档之间的相关性而遇到数据稀疏的问题，就可以考虑将文档之间相关性假设扩大到整个语料库空间中，建立 Biterm Topic Model(BTM)。BTM 通过假设增加了词汇隶属其它主题的可能性，从而使得 BTM 模型在短文本中的效果变好，而且在普通文本中的主题挖掘效果也相当不错。基于以上讨论，为了能更加准确的描绘出短文本文档话题下主题热度，本文考虑以 BTM 为基础，采用将用户的文本信息和用户在话题下的回答的被观看数（views）结合起来纳入词汇权重的考量标准来改进主题挖掘的效果。

2.2 主题演化相关研究

在社会化问答平台的主题演化研究上，Maity[21]等将 Quora 下的所有内容作为研究背景，采取爬取话题标签的方式来获取话题热度并使用回归分析对演化趋势进行了预测；Barua[22]等通过 LDA 模型对 Stackoverflow 中的主题进行分类后，对热点话题之间的关系进行了整理并分析了其演化趋势；Zou[23]等使用 LDA 作为主题挖掘工具对 Stackoverflow 社区中的 NFRs（Non-functional requirements）进行主题分类并获取到每个主题的出现率来表示其热度从而做出了演化趋势并进行了定性分析。上述研究主要是将整个平台内的所有话题作为研究角度的，本文则希望从平台中话题的角度来研究社会化问答平台用户兴趣及其演化趋势。此外，上述研究也没有采用更适合该平台的主题挖掘模型来研究用户兴趣及其演化趋势。

综上，现阶段针对社会化问答平台的用户兴趣演化存在主题挖掘方法选择和演化特征分析上的不足。本文将充分考虑到社会化问答平台的特征来选择较为合适的方法对平台中话题下的主题进行识别，在此基础上，再采用频次计量法对时间维度上主题演化的热度进行评估，然后采用内容分析法对其趋势进行了合理分析并提出了社会化问答平台的个性化推荐策略和广告投放方面的一些建议。

3 研究框架

3.1 研究技术路线

为了对社会化问答平台用户兴趣的识别和演化趋势进行分析，本文在研究思路上大致沿袭了数据采集、数据预处理、数据使用和数据分析的一般范式来设计研究技术路线（如图 1 所示）。

首先，在研究内容方面，本文选取了英文社会化问答平台 Quora 作为研究场景，但 Quora 下拥有众多的话题，选取合适的话题对于研究用户兴趣或主题演化有着深刻的影响。比较适合做演化分析的话题应该具备演化的速度较快，演化的特征较为明显的特征，而 Film and Television 这个话题比较符合这些条件。

随后，借助 Python 网络爬虫与数据库操纵技术，本文对 Quora 中 Film and Television 话题下的 2018 年 12 月 17 日至 2019 年 4 月 14 日的所有问答文本进行了采集和预处理。其中，预处理工作主要是采用 Python 中 NLTK 包对英文中的停用词、

数字标点进行剔除，并全部小写化。而后，还需要结合话题下的演化规律设置时间片划分的长度为用户兴趣演化分析做准备。

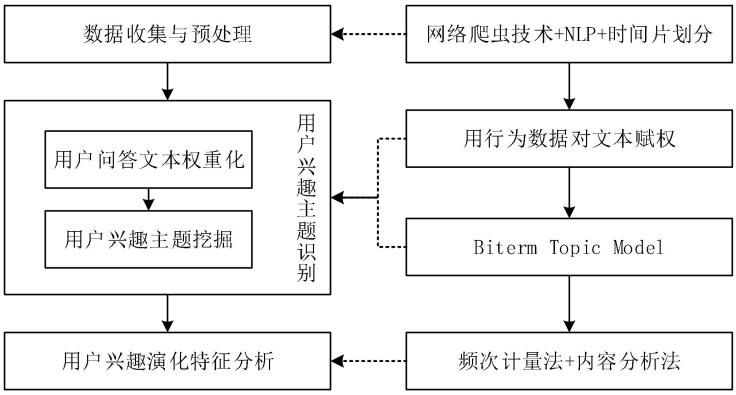

图 1 社会化问答平台用户兴趣演化研究技术路线图

在用户兴趣主题识别上，本文结合用户行为数据对问答文本进行赋权，并通过实验确定了赋权的参数和 BTM 需要设定的划分主题数，对每个时间切片下的数据依据影视剧题材种类进行了主题识别和归纳。最后，通过频次计量方法对每个时间切片下的用户兴趣主题进行了热度识别，之后使用了热力图绘制了每个主题在时间维度上的演化趋势并采用内容分析法对演化趋势进行了分析，对 Quora 在 Film and Television 话题等影视剧话题下的个性化推荐和广告投放提出了合理建议。

3.2 研究基础理论

3.2.1 BTM 理论

Yan[20]等的实验结果表明，针对用户兴趣识别上，采用 BTM 进行主题分类的效果在社会化问答平台这类短文本数据中比较优异。通过下图 2 中的演示，可以看出 LDA 模型对每一个文档都建立了一个主题分布，但这样的设定使得主题间关系被削弱，然而，前文 2 中已经阐述，BTM 是通过扩大假设的方式，针对每一个 biterm（词对）建立一个主题，这样就解决了文档文本较短、词汇稀疏等问题。

具体而言，BTM 认为每一个从语料库中抽取出来的 biterm（词对）都拥有一个指定的主题（如图 2 所示），而指定的某个主题会拥有一个关于语料库中词汇的分布，二者概率相乘便是这个词汇在抽取出来的 biterm（词对）中出现的概率，两个词汇相乘便是这个 Biterm（词对）出现的频率。

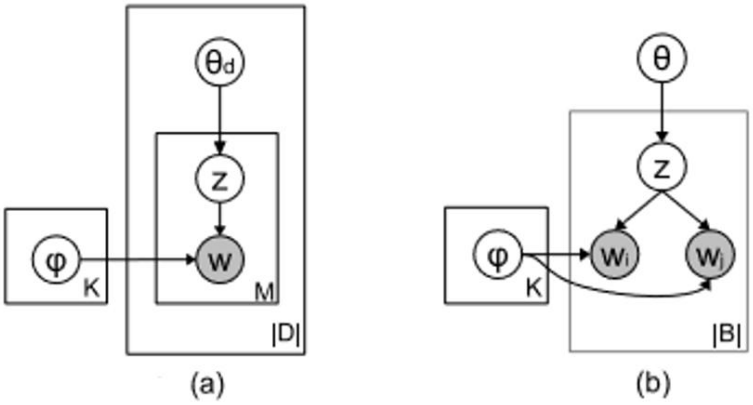

图 2　主题挖掘方法比较图（(a)LDA 模型 (b)BTM）

假设α和β是狄利克雷先验分布得到的参数，上述计算过程具体如下：

1. 对每个主题z，制定一个关于主题-词汇（topic-specific word）分布$\phi_z \sim Dir(\beta)$

2. 对整个语料库中的所有词对（collection）制定一个话题分布$\theta \sim Dir(\alpha)$

3. 针对词对集合B中个每个词汇b：

（a）分配一个指定的主题$z \sim Multi(\theta)$

（b）分配主题下的两个词：$w_i, w_j \sim Multi(\phi_z)$

接着上面的步骤，可以计算出指定 biterm（词对）的出现概率：

$$P(b) = \sum_z P(z)P(w_i \mid z)P(w_j \mid z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{1}$$

由此，整个语料库的概率就是：

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \varphi_{i|z} \varphi_{j|z} \tag{2}$$

接着，针对上文中已经假设得到的α和β的参数推断上，BTM 采取的是 Gibbs sampling（吉布斯采样）作为推断方式。Gibbs sampling 是一个简易且应用及其广泛的 MCMC 算法，通过 BTM 中的吉布斯采样算法可以计算出每个主题对于 biterm（词对）$b = (w_i, w_j)$分配的概率值：

$$P(z \mid \mathbf{z}_{-b}, B, \alpha, \beta) \propto (n_z + \alpha) \frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2} \tag{3}$$

在（3）式中$\mathbf{z}_{-b}$代表除了b之外所有 biterms（词对）的主题分配，B代表所有的 biterms（词对）集合，n_z代表 biterm（词对）b被分配给主题z的次数，$n_{w|z}$代表词汇w被分配给主题z的次数。这里的计算将延续 LDA 模型的传统，主题下的两个词汇w_i和w_j将被同时分配。最终，通过对主题在 biterms 中的分布次数和词汇在主题中的出现频次的计数，BTM 模型可以计算出主题-词汇（topic-word）分布和主题（global topic）分布：

$$\phi_{w|z} = \frac{n_{w|z} + \beta}{\sum_{w} n_{w|z} + M\beta} \tag{4}$$

$$\theta_z = \frac{n_z + \alpha}{|B| + K\alpha} \tag{5}$$

其中，$|B|$是所有 biterms（词对）的数量。

针对上述的模型，本文采用的是 Mimmo[24]等提出的评价主题分类模型的方式，即通过计算被 BTM 自动分类出的各个主题的 coherence 值来评估效果，其值越大说明主题分类效果越好，其具体计算过程如下：

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D(v_m^{(t)},v_l^{(t)})+1}{D(v_l^{(t)})} \tag{6}$$

在(6)式中，$D(v,v')$表示词汇v和v'的文档共现频率，$V^{(t)} = (v_1^{(t)},v_2^{(t)},...,v_M^{(t)})$代表主题$t$下概率分布最高的$M$个词汇，分子中多添加了 1 是为了防止分子为零的情况出现。

考虑到用户行为数据如回答被观看数（views）可以在一定程度上表达用户的喜好，本文这里也考虑将用户的行为数据回答被观看数（views）作为参数对文本数据进行赋值，通过实验结果中 coherence 值得反馈来选定 BTM 中的主题数（num_topics）参数和合适的参数来增加文本数据的权值，从而比较未赋权文本和赋权文本之间的实验效果。最后，本文还对识别出来的主题依据影视剧题材分类方法进行了识别和归纳。

3.2.2 频次计量法

用户兴趣演化特征描述则需要对每个时间段下的主题的热度进行识别，Zou[23]等提出过采用统计主题在各个文档中的出现频率来表示主题的热度，这里本文采用同样

的方式来计算。针对 BTM 中，计算该文档的主题分布需要借助如下公式：

$$P(z|d)=\sum_{b}P(z|b)P(b|d) \tag{7}$$

$$P(z|b)=\frac{P(z)P(w_i|z)P(w_j|z)}{\sum_{z}P(z)P(w_i|z)P(w_j|z)} \tag{8}$$

$$P(b|d)=\frac{n_d(b)}{\sum_{b}n_d(b)} \tag{9}$$

在（8）式中，$P(z)=\theta_z$，$P(w_i|z)=\phi_{i|z}$；在（9）式中，$n_d(b)$ 代表 biterm（词对）b 在文档 d 中的出现频率。之后，根据每篇文档中的主题概率分布最大值对应的主题数来决定文档所属的主题，并通过统计每个主题在所有文档中的出现频率来表示主题热度。然后根据主题在十七周中的变化趋势，本文将使用 heatmap 绘制出演化趋势图，并在此基础上，通过内容分析法来研究主题演化的特征，对社会化问答平台的话题演化趋势进行总结，以及对 Quora 该话题下的个性化推荐提出一些建议。

4 数据采集与预处理

4.1 数据采集

本文使用 Python 作为爬虫语言，借助 selenium 中的 webdriver 模拟登陆 Quora 后，选取 Film and Television 下的 All Questions 选项进入话题下的问题列表，其中问题绝大多数是按照时间顺序，从当下往过去排列的。本文通过 bs4 中的 Beautifulsoup 解析网页中 2018 年 12 月 17 日至 2019 年 4 月 14 日时段下（共十七周）所有问题的链接并保存到 mysql 数据库中。接着，本文通过访问 mysql 中的每一个问题链接，使用模拟登陆和解析网页的方法提取出网页中（1）问题的链接；（2）问题的标题；（3）问题的回答总数；（4）关注此问题的最后时间；（5）回答此问题的文本信息（图片、视频等信息不予考虑）；（6）回答的时间；（7）回答的观看数保存到 mysql 数据库中。

本文通过爬虫一共采集了 3849 条问题和 6212 条问答数据，之后，对问题回答数量分布进行统计分析（如图 3），可以发现大部分问题只有 0 至 1 个回答，少数问题可以得到 10 个以上的回答。这也说明，在该主题下会有很多个性化的想法和疑问。

但从拥有 6 至 20 个回答数的问题分布情况来看，下降趋势并不是很明显，这也说明话题下也存在着很多的共性问题。

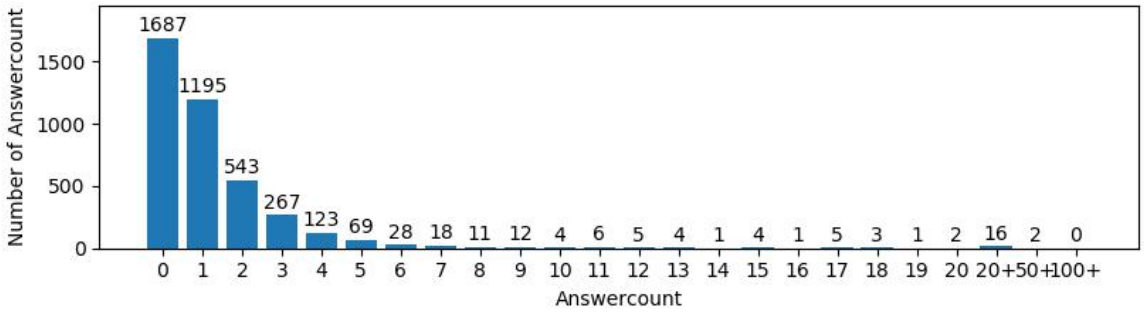

图 3　Film and Television 主题问题回答数量分布情况（2018.12.17-2019.04.14）

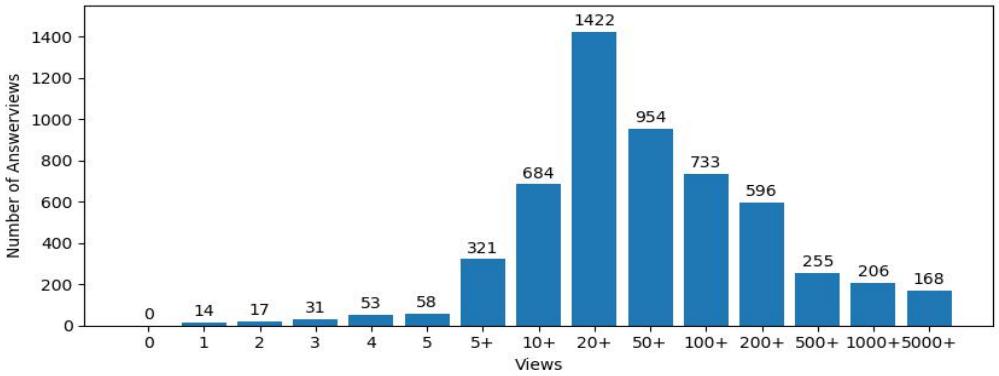

图 4　Film and Television 主题问题回答被观看次数分布情况（2018.12.17-2019.04.14）

通过图 4 中对十七周中问题回答被观看次数的分布情况，在刨除掉没有回答的问题后，可以发现从回答被观看 1 次到回答被观看 20 次以上之间一直处于增长状态，而且被观看超过 5 次以上的回答数在迅速增长。此外，从柱状图的分布情况来看，其大致符合右偏分布的情形，大部分问题会被浏览 20 至 50 次左右。此外，本文还发现，被观看超过 1000 次和 5000 次的回答数量依然很可观，这种现象说明话题下的某些问答引起较多用户的关注，答案也得到了众多用户的赞许，说明 Film and Television 话题存在较强的读者共鸣情况。由此，研究此话题下用户的主题偏好和主题演化趋势会对用户推荐效果的改善有较强的帮助和指导意义。

4.2 数据可视化与数据预处理

由于英美剧和电视节目的更新周期为一周，本文将时间窗口设置为一周来进行用户兴趣演化分析。于是，从 2018 年 12 月 17 日至 2019 年 4 月 14 日被划分成十七周，其每周问题的数量分布如图 5 所示，通过图 5 可以看出一周中的问题数量在 210 个上

下浮动，数据分布整体比较平稳。

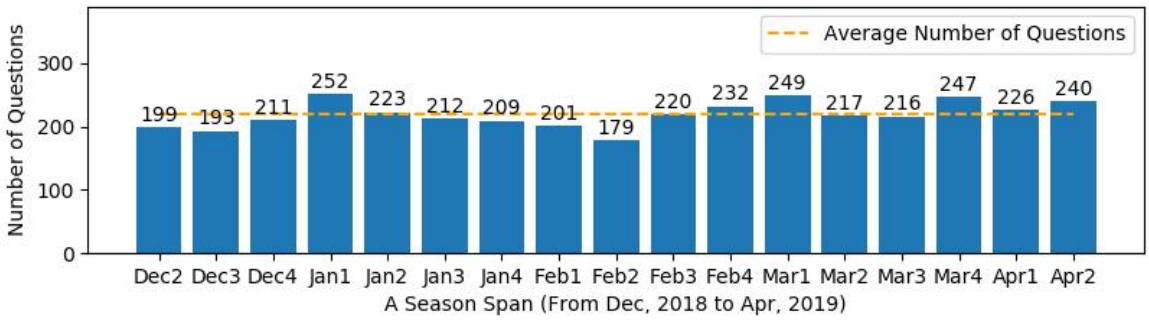

图 5 Film and Television 主题问题分布情况（2018.12.17-2019.04.14）

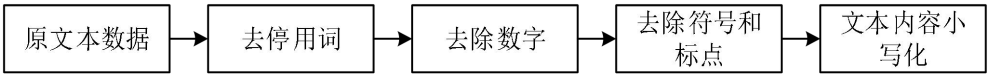

图 6 数据清洗实施步骤

接下来，数据清洗环节主要是通过将按周切分好的问答数据利用 python 中的 nltk 包对英文中的停用词、数字和标点符号进行清除，全部改成小写（如图 6 所示），并将整理好的文档保存下来以便主题挖掘的时候使用。

5 用户兴趣主题识别

5.1 兴趣主题挖掘模型构建

在使用 BTM 等主题挖掘模型对文本数据进行主题兴趣识别时，如何设置分类的主题数通常是一个棘手的问题，通过实验结果反馈并结合实际情况进行调试是一个有力的方式。此外，考虑到文本内容的观看数应该能够反映文本的重要程度，而且 Film and Television 话题下的很多回答都被观看了超过 1000 次（见前文图 4）。于是，本文考虑将文本信息重新编辑评估。这里，本文对文本数据的赋权方式是通过回答被观看数来展开的，即先设置一个浏览值，话题下的回答内容被观看超过一次该浏览值则将被在文档中复制一次。针对赋权文本的权重参数，也需要通过实验的结果反馈进行调试，并判断采用赋权文本的 BTM 效果是否有明显改进。

如何选取合适的主题数是一个比较关键的因素，其值大小会对最终主题挖掘的 coherence 值和实际挖掘效果产生一定的影响。这里，初步暂定模型的迭代次数为 20 次，同时，为了比较利用未加权文本数据和加权后文本数据的训练效果并预调试 BTM 中 num_topic（主题数）的参数值，本文分别设置被观看了超过 50,100,1000,2000，

5000 次的问答数据即被复制一次。使用 2019 年 4 月 8 日至 2019 年 4 月 14 日的问答数据作为训练样本，借助 BTM 内部的 coherence 计算方式得到了每个主题的 coherence 值，并对每次主题挖掘后的各个主题的 coherence 值取平均值来表示此次主题挖掘的训练效果，并得到了如图 6 所示的初步实验结果。

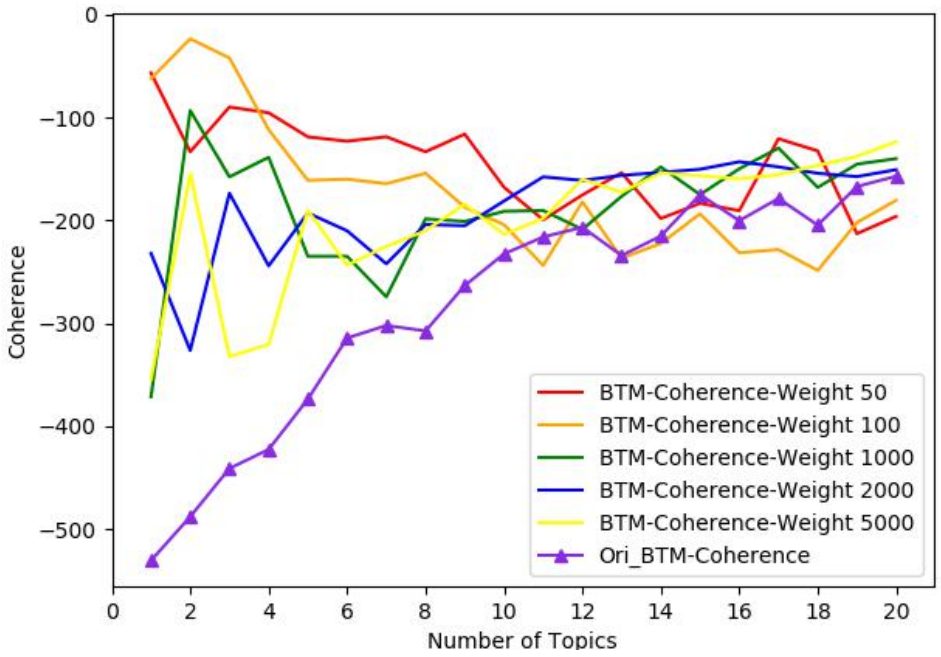

图 7 有无权重文本数据训练效果比较

通过实验对比（如图 7 所示）发现，对未加权文本进行主题分类的 coherence 值一开始就不是很高，随着话题数量的增加在逐步提升。而赋权之后的文本被 BTM 训练过之后进行主题分类。这里还可以发现，随着权重值的增加，赋权文本的 coherence 值会接近未赋权文本，造成这样的原因是当权重阈值设置得较高时，能得到赋权的文本数量会比较少，接近于未赋权文本的情况。但令人惊喜的是，其 coherence 值在话题数低于 10 个的时候无论权重设置为多少都会优于未赋权文本的值。

考虑到在 Quora 这类短文本文档 Film and Television 话题下一周的问题数量在 200 个左右，问答数量在 700 个左右的现实情况，这里的模型不宜设置较多的话题数，应该设置在 10 个以内。这样以来，本文认为该实验结果可以初步说明考虑用户行为数据的重要性，权重的阈值应当适当降低，这样有利于文本数据的主题挖掘效果的提升。结合以上分析，可以认为通过对回答被观看数这一用户行为数据的应用，较好地提高了 BTM 的主题挖掘训练效果。

此外，在图 7 中本文发现对权重阈值设置为 50 和 100 的加权文本数据进行 BTM

主题分类的 coherence 值在 number of topics（主题数）超过 5 个之后下降趋势比较明显。所以，根据图 6 中反馈的结果，本文会将主题数设置为 3-6 之间，即接近使得主题挖掘模型 coherence 值迅速下跌的 number of topics（主题数）的取值，具体值需要根据话题的分类结果来波动。

为了进一步验证加权文本后对主题分类模型 BTM 在 Quora 中 Film and Television 话题下的训练效果，本文对权重值得设置有进行了如下的实验。本文将话题分类数选定在 6 个，并设置循环值为 5000，步长为 50，即从回答被观看 1 次即被复制一次至回答被观看 5000 次以上再被复制一次。通过图 8 的实验结果可以看出 BTM 的 coherence 值在被线性函数拟合后呈现一个下降的趋势，而且在权重阈值在 0-500 这个区间内的下降速度较快，但总体效果是要优胜于未加权文本的效果的。这说明，在设置权重阈值的时候应该考虑现实情况，需要选择一个合理的筛选机制，在这里权重阈值设置为 100 左右比较好。

结合实验 coherence 结果的变化，本文认为产生这种结果的主要原因是 Biterm Topic Model 的模型是针对 biterm 词对而建立的，本文采用的赋权模式是采用简单的复制模式展开的。而根据图 5 中的 Film and Television 话题下的回答被观看数的分布情况来看，被观看次数超过 1000 次的文档是比较可观的，这也就急剧增加了用户比较感兴趣的高频 biterm（词对）的出现概率，从而使得一些词汇的概率值得到了较好的提升，与此同时也使得主题内部排名靠前的词汇分布更加聚合。

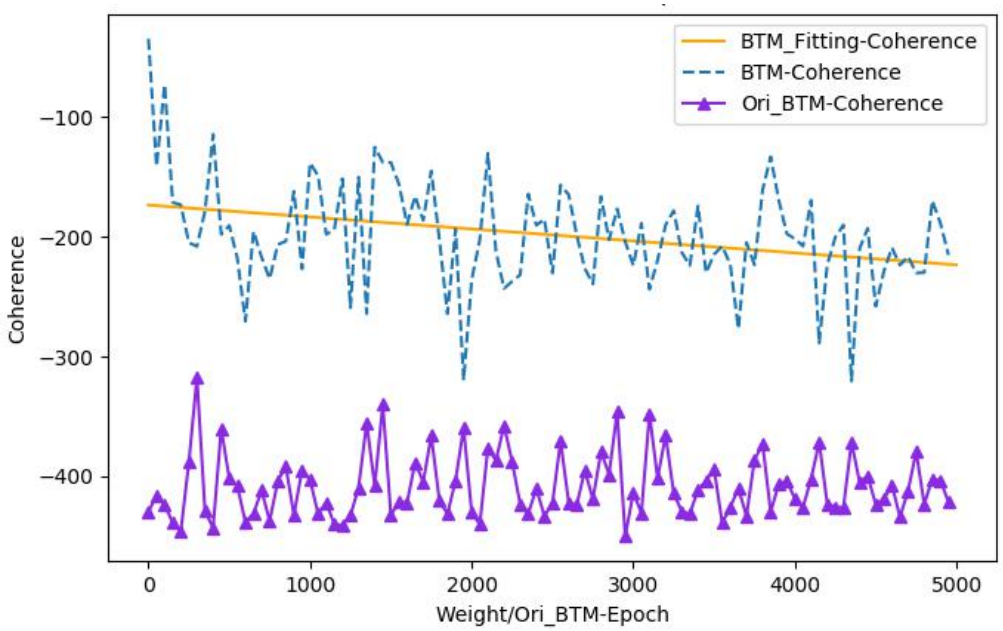

图 8 文本权重赋值变化与无权重文本 coherence 值比较

5.2 兴趣主题识别结果

借助改进文本内容之后的 BTM，本文对十七周数据的主题进行了分类，并得到了按照概率值排名得到的每个主题下排名前 10 位的热词。通过英美剧和电视节目知识的常识和 Wikipedia 中的词条匹配，本文结合影视剧题材分类方法，将 Film and Television 话题下的主题分为剧情（Feature）；爱情（Affectional）；战争（War）；喜剧(Comedy)；科幻(SciFi)；动画(Animation)；惊悚(Thriller)；犯罪(Criminal)；纪录（Documentary）；情节（Plot）；影视艺人(Actor)；电视节目（TV Show）共 12 个主题。之后，就可以对每周的主题分类结果添加相应的标签，这里，本文选取 2019 年 4 月 8 日至 2019 年 4 月 14 日的主题分类结果（如表 1 所示）来展示分类效果（详细数据见附录）。

表 1　2019.04.08-2019.04.14 主题分类结果

主题标签	Topic1		Topic2		Topic3		Topic4	
（主题标签）	（Comedy）		（Criminal）		（SciFi & The Avengers）		（SciFi）	
主题下热词及词权	episode	0.04	people	0.045	thanks	0.015	standard	0.028
	guy	0.039	career	0.044	**world**	0.014	**fiction**	0.028
	family	0.038	episode	0.042	frank	0.014	examine	0.028
	mantain	0.035	city	0.041	**paul**	0.013	structure	0.028
	quo	0.028	**weapon**	0.041	harris	0.013	informative	0.028
	adultoriented	0.028	**cop**	0.041	invisible	0.013	drama	0.028
	change	0.026	**firing**	0.041	**hemsworth**	0.013	cast	0.028
	phenomenon	0.026	**shooting**	0.041	bana	0.013	**chikills**	0.017
	simpson	0.026	**criminal**	0.041	liam	0.013	**jessica**	0.017
	favourite	0.026	**police**	0.041	jordan	0.013	**hero**	0.017
主题出现频率	55		182		262		95	
主题热度占比	9%		31%		44%		16%	

对于主题属于多个题材的情况，则需要通过借助主题下那些表征这个主题的词汇热度去选择热度较高的主题词设置为这个主题下的关键词。当遇到特殊话题例如艺人

韵事，影视首发和集中讨论某些经典电影的时候，会在相应的主题后面加上热点主题，就像表 1 中的 Topic3 一样，但可能并不在以上的 12 个主题标签中。

6 用户兴趣演化分析

6.1 兴趣主题热度计算

本文对每周话题进行演化分析的时候，还需要对相似话题进行合并简化。在表 1 中，实验环境中 BTM 将两个 SciFi 话题分开是因为一个主题是单纯谈论一些 SciFi 电影，另一个话题在讨论 SciFi 是因为 The Avengers 即将上映而掀起了一波讨论的小高潮。但为了方便绘图和主题演化分析，本文在这里选择将两个话题合并为 SciFi & The Avengers，类似的处理会也会存在于前面数周中的主题分类的标签工作中。最终，本文绘制出了 2018 年 12 月 17 日至 2019 年 4 月 14 日共计十七周的主题演化趋势热力图（如图 9 所示）。在图 9 中表征的是每周时段下各个主题在时间维度上的演化情况，每一行代表每个主题在这十七周中的热度演变情况。

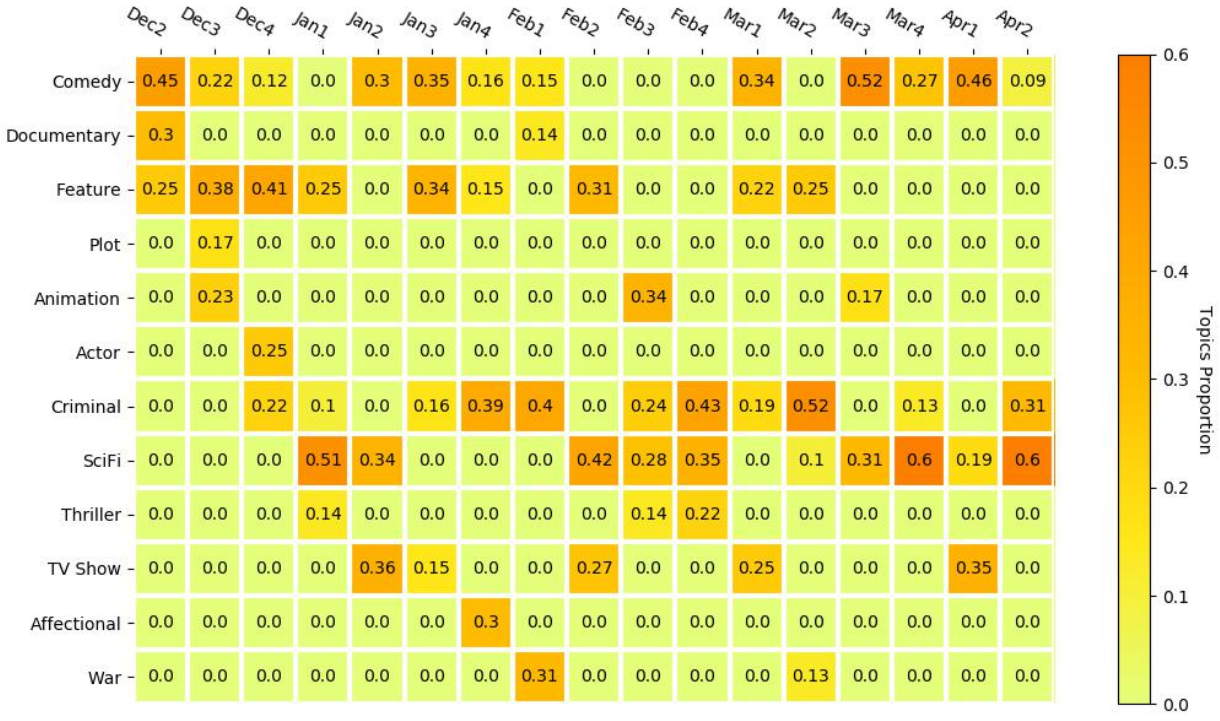

图 9　主题演化趋势热力图（2018.12.17-2019.4.14）

此外，借助前文 3.2 中提到的频数计量法，可以得到表 1 中的每个主题的出现频率统计，从而得到每个主题在所有文档中所占的比例。这里，本文通过这个比例来表

示每个主题在此周所有讨论内容中的热度，其数值将在表 1 中的主题热度占比中体现，并在图 10 中通过数值和热力图中的颜色深浅来表征。

6.2 兴趣演化趋势分析

结合对图 10 中各个主题在 2018 年 12 月 17 日至 2019 年 4 月 14 日的演化分析，本文发现 Quora 中 Film and Television 话题的如下规律：

话题更新速度较快，虽然在题材上可能仍然属于一个类别，但具体的内容让会变化较大，一般生存周期就在一周到两周左右；本文认为造成这样的原因是英美剧和电视节目的更新频率较快和事件发生具有一定的偶然性，例如某些艺人的一些新闻可能会产生该周的一个热点主题。

话题从触发到成为热点需要一定的时间，一般在一周左右，随之又会被新的话题所取代；此外，从现实中热点事件的发生到成为话题热点也会存在一定的延迟，比如在 Deadpool PG13 版 2018 年 12 月 12 日在北美上映之后，与 Deadpool 相关的 SciFi 成为讨论热点则是在两周之后的话题中，这正好处于 Deadpool 在北美上映档期中，且处于中国大陆上映之前的时间段。本文认为造成这样的原因是，用户对话题的讨论需要时间的积累，当前话题下讨论的热点主题可能在早于该时间段的时候已经成为热点话题，这也说明如果需要对用户进行内容推荐的时候，需要结合一些现实生活中的热点问题来提前预测。

在十七周的演化主题中，Comedy（喜剧）、SciFi（科幻）和 Criminal（犯罪）话题的出现次数较多，是该话题下的讨论主流，由此，系统可以对 Film and Television 话题下关注相应话题的用户进行主推相关内容。因为当下的在更新的情景喜剧不是很多，大部分 Comedy 内容可能会偏向于经典喜剧和电视节目。这样针对 Comedy 话题的内容推荐就很简单，在有关于过去经典的喜剧内容中选取出推送的内容，其受到喜爱的概率会相应增大。Criminal（犯罪）话题下多会讨论影视情境和情节等问题，话题较有可能在下一周演化成为 Thriller（惊悚）和 Character（影视角色）等话题。SciFi（科幻）话题则会讨论一些经典的科幻影视以外的新话题，因为当下在更新的影视作品非常多，在对用户内容进行推荐的时候，就需要考虑当下正在更新热点作品和即将放映的新品来进行推荐。

在这里，本文虽然是从 Quora 中的 Film and Televiosn 话题作为切入点来研究

用户兴趣演化的，但这里的经验可以推广到类似的话题中去，例如 Movies，Television Series 等影视剧话题中，对影视剧话题的个性化推荐和广告投放有着很好的借鉴意义。

7 结束语

对社会化问答平台用户兴趣演化研究，本文是以 Quora 作为研究案例，采用话题作为切入点展开的。通过 Python 网络爬虫技术，数据库操纵技术和数据清洗工具包，依据英美剧和电视节目的更新周期将时间窗口设置为一周，对从 Quora 中的 Film and Television 话题中采集出来的文本数据进行了数据预处理工作。基于文章对 LDA 模型和 BTM 的理解和讨论，本文最终选择了 BTM 作为主题挖掘的方法，并结合用户的回答被浏览数作为参数对文本数据进行了赋权。实验结果证明赋权文本的效果要优于为赋权文本，BTM 主题数确定在 3-6 个较好，赋权文本的参数设置为回答被浏览超过 2000 次左右即被复制一次效果较好，这也是本篇文章的一个重要改进和创新点。之后，通过对十七周每周赋权文本数据的主题挖掘，本文最终得到了每周的用户兴趣主题和热点词汇，并对主题进行了归纳分类。

通过每个主题下的热点词汇，可以对主题的讨论内容进行评估和分类。之后，依据每周中主题在各个问答文档中的出现频次占比，可以对主题的热度进行标识并得到相应数值。将每周讨论的主题热度通过热力图用颜色深浅来表示，再将每周讨论的主题依据横轴按时间顺序从过去至将来就可以得到一个主题演化趋势图。

最终，本文结合 2018 年 12 月 17 日至 2019 年 4 月 14 日中的主题演化趋势，对 Film and Television 话题下的演化趋势进行了定性分析并认为 Film and Television 话题演化速度比较快。此外，通过对十七周数据的具体分析，本文还认为话题热点的产生从现实中到平台中会有一定的延迟，而且话题的讨论主流题材主要是 Comedy（喜剧）、SciFi（科幻）和 Criminal（犯罪）题材。由此，系统可以对 Film and Television 话题下关注相应话题的用户进行主推相关内容。针对 Comedy 话题的内容推荐可以在关于过去经典的喜剧内容中选取出推送的内容，其受到喜爱的概率会相应增大。Criminal（犯罪）话题下多会讨论影视情境和情节等问题，话题较有可能在下一周演化成为 Thriller（惊悚）和 Character（影视角色）等话题。SciFi（科幻）话题在对用户内容进行推荐的时候，则可以考虑当下正在更新热点作品和即将放映的新品来

进行推荐。

综上，本文对 Quora 中 Film and Television 话题下的用户兴趣进行了主题挖掘，并通过对文本数据的赋权优化了 BTM 的主题识别效果。在随后的用户兴趣演化趋势分析中，通过对主题的定性分析为 Quora 在该话题下进行个性化推荐和广告投放提出了一些建议。但本文没有对后续的主题演化趋势进行定量建模分析得到可以量化的规律，这也将成为用户兴趣演化的下一个重要突破方向。

附录

表 2 2018.12.17-2018.12.23 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Feature & Criminal	Topic1	girl	0.018	69	16%
		character	0.017		
		murderer	0.017		
		favourite	0.016		
		obscure	0.013		
		historical	0.013		
		event	0.013		
		breaking	0.012		
		pinkman	0.012		
		jesse	0.012		
Feature & Adventure	Topic2	sure	0.059	39	9%
		imposed	0.058		
		silverlake	0.055		
		tonto	0.05		
		actor	0.049		
		man	0.047		
		silted	0.046		
		workshop	0.046		
		character	0.043		
		real	0.043		
Documentary	Topic3	movie	0.048	126	29%
		throne	0.043		
		great	0.041		
		delegate	0.03		
		polish	0.03		
		important	0.03		
		wikipedia	0.03		
		henry	0.03		
		explain	0.03		
		france	0.03		
Comedy	Topic4	piece	0.025	75	17%
		parent	0.022		
		ben	0.022		
		plan	0.022		
		greg	0.022		
		byrnes	0.022		
		propose	0.022		
		wikipedia	0.022		
		focker	0.022		
		learn	0.022		

Comedy	Topic5	time	0.017	84	22%
		sound	0.015		
		age	0.015		
		home	0.012		
		experience	0.01		
		guy	0.009		
		prarie	0.009		
		foley	0.009		
		horse	0.009		
		companion	0.009		

表 3　2018.12.24-2018.12.30 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Comedy	Topic1	time	0.015	84	22%
		drama	0.012		
		comedy	0.012		
		big	0.011		
		book	0.009		
		sherlock	0.009		
		day	0.008		
		suggest	0.008		
		think	0.008		
		scenario	0.008		
Plot	Topic2	**film**	0.032	67	17%
		stroytelling	0.016		
		challenge	0.014		
		visual	0.014		
		director	0.014		
		speaking	0.014		
		device	0.014		
		fun	0.014		
		recent	0.014		
		creative	0.014		
Animation & Character	Topic3	**actor**	0.014	87	23%
		character	0.014		
		thing	0.013		
		called	0.012		
		person	0.01		
		mind	0.008		
		cartoon	0.008		
		girl	0.008		
		great	0.008		
		real	0.007		

Feature & Criminal	Topic4	episode	0.018	148	38%
		bad	0.017		
		face	0.016		
		moment	0.012		
		scene	0.011		
		died	0.011		
		role	0.01		
		beraking	0.009		
		romance	0.009		
		expecting	0.009		

表 4 2018.12.30-2019.01.06 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Criminal & Actor	Topic1	time	0.015	137	21%
		drama	0.012		
		comedy	0.012		
		big	0.011		
		book	0.009		
		sherlock	0.009		
		day	0.008		
		suggest	0.008		
		think	0.008		
		scenario	0.008		
Actor	Topic2	featured	0.025	163	25%
		woman	0.023		
		thought	0.022		
		young	0.02		
		penny	0.017		
		shirley	0.017		
		laverne	0.017		
		marshall	0.017		
		recent	0.017		
		news	0.017		
Feature & Character	Topic3	rick	0.018	265	41%
		time	0.017		
		episode	0.016		
		question	0.011		
		actor	0.01		
		jackson	0.01		
		meant	0.01		
		percy	0.01		
		mean	0.01		
		teach	0.01		

Comedy & Actor	Topic4	happy	0.032	78	12%
		old	0.032		
		star	0.031		
		think	0.027		
		winkler	0.027		
		student	0.026		
		high	0.026		
		fonzie	0.026		
		wit	0.026		
		henry	0.026		

表 5 2019.01.07-2019.01.13 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Thriller	Topic1	home	0.034	115	14%
		american	0.032		
		horror	0.032		
		long	0.032		
		toddler	0.032		
		mom	0.032		
		bloody	0.032		
		hidden	0.032		
		belly	0.032		
		corpse	0.032		
Feature & Cult	Topic2	human	0.048	201	25%
		godfrey	0.048		
		reggio	0.048		
		experience	0.047		
		produced	0.047		
		concentrate	0.047		
		philip	0.047		
		glass	0.047		
		musical	0.047		
		koyanisqaasti	0.047		
SciFi & Deadpool	Topic3	fred	0.07	193	24%
		recreate	0.068		
		savage	0.067		
		bedroom	0.064		
		hero	0.063		
		deadpool	0.062		
		version	0.06		
		kidnap	0.048		
		movie	0.044		
		princess	0.043		

SciFi & Starwar	Topic4	best	0.039	217	27%
		movie	0.033		
		adlibbed	0.028		
		carbonite	0.028		
		empire	0.027		
		strike	0.027		
		leia	0.027		
		choose	0.027		
		han	0.027		
		frozen	0.027		
Criminal & Thriller	Topic5	scene	0.055	79	10%
		upset	0.04		
		needle	0.04		
		murdered	0.04		
		jesse	0.04		
		andrea	0.04		
		breaking	0.04		
		bad	0.04		
		brock	0.04		
		inside	0.04		

表 6 2019.01.14-2019.01.20 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
SciFi & Deadpool	Topic1	movie	0.025	324	34%
		say	0.016		
		scene	0.016		
		pg	0.016		
		version	0.015		
		deadpool	0.014		
		princess	0.014		
		break	0.014		
		bedroom	0.014		
		savage	0.014		
TV Show & Comedy	Topic2	sex	0.015	290	30%
		written	0.013		
		character	0.012		
		main	0.011		
		stern	0.01		
		prank	0.01		
		potter	0.01		
		wizard	0.01		
		porkies	0.01		
		convince	0.01		

TV Show	Topic3	kind	0.011	343	36%
		read	0.01		
		say	0.01		
		bengi	0.01		
		kya	0.01		
		fav	0.01		
		karna	0.01		
		ka	0.01		
		bhi	0.01		
		mai	0.01		

表 7 2019.01.21-2019.01.27 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Comedy & Character	Topic1	green	0.047	72	16%
		chandler	0.041		
		schwimmer	0.041		
		cast	0.04		
		joey	0.039		
		leblanc	0.036		
		aniston	0.035		
		kudrow	0.035		
		rachel	0.035		
		best	0.035		
Scene & Criminal	Topic2	movie	0.016	71	16%
		story	0.013		
		dark	0.011		
		criminal	0.01		
		people	0.01		
		scene	0.01		
		believe	0.01		
		english	0.009		
		thing	0.009		
		time	0.009		
TV Show	Topic3	lot	0.016	68	15%
		used	0.016		
		actor	0.013		
		sinatra	0.012		
		ncis	0.012		
		superbowl	0.012		
		football	0.012		
		loop	0.012		
		downtown	0.012		
		parking	0.012		

Feature	Topic4	movie	0.017	153	40%
		series	0.016		
		actor	0.014		
		people	0.012		
		tom	0.01		
		funny	0.01		
		bombadil	0.01		
		certain	0.009		
		juvenil	0.009		
		giant	0.009		
Comedy	Topic5	member	0.031	91	20%
		fear	0.02		
		near	0.019		
		day	0.016		
		sitcom	0.016		
		way	0.016		
		run	0.016		
		cast	0.016		
		loyal	0.016		
		left	0.016		

表 8　2019.01.28-2019.02.03 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Affectional	Topic1	time	0.047	205	30%
		ancient	0.046		
		love	0.045		
		culture	0.045		
		language	0.045		
		platonic	0.045		
		understanding	0.045		
		code	0.045		
		space	0.045		
		exist	0.045		
Feature & Criminal	Topic2	watch	0.016	268	39%
		follow	0.015		
		like	0.013		
		set	0.012		
		breaking	0.01		
		plenty	0.009		
		coffee	0.009		
		listen	0.009		
		lion	0.009		
		bad	0.009		

Feature	Topic3	ancient	0.019	105	15%
		life	0.016		
		series	0.015		
		game	0.013		
		replaced	0.012		
		throne	0.011		
		nostalgia	0.011		
		capture	0.009		
		failure	0.009		
		cristo	0.009		
Comedy & Actor	Topic4	acting	0.06	113	16%
		role	0.06		
		tarrak	0.06		
		metha	0.06		
		ka	0.06		
		ooltah	0.06		
		chashmah	0.06		
		skill	0.06		
		experssion	0.06		
		great	0.06		

表 9 2019.02.04-2019.02.10 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Comedy	Topic1	work	0.014	101	15%
		right	0.014		
		entertainment	0.014		
		bad	0.013		
		mood	0.013		
		fantasy	0.012		
		happen	0.009		
		comedy	0.009		
		advert	0.009		
		broadcast	0.009		
Criminal	Topic2	death	0.029	260	38%
		villian	0.026		
		line	0.026		
		slaughtere	0.025		
		potty	0.025		
		trigger	0.025		
		cop	0.025		
		action	0.025		
		caused	0.025		
		demand	0.025		

War	Topic3	**artillery**	0.052	202	30%
		intense	0.052		
		soviet	0.052		
		american	0.052		
		massive	0.052		
		say	0.052		
		affair	0.052		
		number	0.052		
		gun	0.052		
		russian	0.052		
Documentary & Comedy	Topic4	**sitcom**	0.083	95	14%
		documentary	0.077		
		mood	0.067		
		picturization	0.066		
		style	0.066		
		unlike	0.065		
		make	0.064		
		trust	0.062		
		precise	0.059		
		jiffy	0.056		

表 10 2019.02.11-2019.02.17 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Comedy	Topic1	**shatner**	0.029	41	7%
		skywalker	0.029		
		audrey	0.029		
		hepburn	0.029		
		william	0.029		
		dorothy	0.029		
		indiana	0.029		
		jones	0.029		
		luke	0.029		
		lucy	0.029		
Criminal	Topic2	**batman**	0.013	196	35%
		completely	0.013		
		film	0.013		
		birdman	0.012		
		typical	0.012		
		freeze	0.012		
		terminator	0.012		
		arnold	0.012		
		superhero	0.012		
		lantern	0.012		

War	Topic3	cute	0.057	154	27%
		banoo	0.054		
		dulhann	0.053		
		divyanka	0.05		
		dabmn	0.049		
		rose	0.047		
		acting	0.045		
		mohabbatein	0.044		
		household	0.042		
		favourite	0.042		
Documentary & Comedy	Topic4	thing	0.022	173	31%
		country	0.014		
		orphan	0.014		
		chumlum	0.014		
		ron	0.014		
		rice	0.014		
		week	0.013		
		aesthetically	0.013		
		beautiful	0.013		
		example	0.013		

表 11 2019.02.18-2019.02.24 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
SciFi	Topic1	carrie	0.034	159	28%
		actor	0.03		
		kirk	0.029		
		neil	0.029		
		patrick	0.029		
		fisher	0.029		
		skywalker	0.029		
		luke	0.029		
		friend	0.029		
		reef	0.029		
Thriller & Criminal	Topic2	mother	0.048	82	14%
		lake	0.046		
		birth	0.046		
		new	0.046		
		jersey	0.046		
		caystal	0.046		
		voorhees	0.046		
		allow	0.046		
		introduce	0.046		
		jason	0.046		

Animation	Topic3	best	0.017	195	34%
		subtitle	0.016		
		enjoy	0.016		
		netfilx	0.012		
		anime	0.012		
		attention	0.012		
		understand	0.012		
		wat	0.012		
		actual	0.012		
		tongue	0.012		
Criminal	Topic4	movie	0.027	139	24%
		mind	0.026		
		deep	0.026		
		matthau	0.022		
		crook	0.021		
		audience	0.021		
		death	0.02		
		bank	0.02		
		robbery	0.02		
		resulting	0.02		

表 12 2019.02.25-2019.03.03 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
SciFi & Adventure	Topic1	series	0.058	195	25%
		replaced	0.051		
		stargate	0.05		
		half	0.05		
		farscape	0.05		
		tv	0.05		
		cast	0.05		
		sg	0.05		
		death	0.05		
		scene	0.05		
Thriller	Topic2	character	0.057	170	22%
		rewatch	0.05		
		slayer	0.05		
		vampire	0.05		
		firefly	0.05		
		whedon	0.05		
		think	0.05		
		school	0.05		
		creator	0.05		
		joss	0.05		

Criminal	Topic3	work	0.026	73	9%
		jake	0.026		
		prove	0.026		
		detective	0.026		
		talented	0.026		
		cop	0.026		
		peralta	0.026		
		ray	0.026		
		holt	0.026		
		brooklyn	0.026		
Criminal & Thriller	Topic4	actor	0.018	266	34%
		role	0.017		
		netfilx	0.016		
		crime	0.011		
		stalker	0.011		
		new	0.011		
		dirty	0.011		
		perfect	0.01		
		success	0.01		
		thriller	0.01		
SciFi & Captain Marvel	Topic5	**man**	0.024	75	10%
		started	0.019		
		imperium	0.016		
		captain	0.015		
		jr	0.014		
		iron	0.013		
		marvel	0.012		
		peak	0.012		
		screen	0.012		
		success	0.012		

表 13 2019.03.04-2019.03.10 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
SciFi	Topic1	time	0.021	316	34%
		question	0.017		
		clean	0.016		
		april	0.016		
		sabachthani	0.016		
		anatomy	0.016		
		jesus	0.016		
		died	0.016		
		cross	0.016		
		eloi	0.016		

Thriller & Criminal	Topic2	affair	0.027	232	25%
		urinate	0.025		
		club	0.016		
		carson	0.012		
		wife	0.011		
		johnny	0.011		
		friar	0.011		
		roast	0.011		
		camera	0.011		
		tonight	0.011		
Animation	Topic3	looking	0.02	177	19%
		try	0.019		
		case	0.019		
		drama	0.018		
		lost	0.017		
		feel	0.014		
		day	0.012		
		hwang	0.011		
		serial	0.011		
		uncover	0.011		
Criminal	Topic4	california	0.021	207	22%
		way	0.013		
		fun	0.011		
		beautiful	0.011		
		violence	0.011		
		uk	0.011		
		battle	0.011		
		throne	0.011		
		war	0.011		
		boring	0.011		

表 14 2019.03.11-2019.03.17 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Criminal & Sherlock Holmes	Topic1	**holmes**	0.028	181	29%
		school	0.026		
		produced	0.026		
		fan	0.026		
		law	0.025		
		world	0.024		
		benedict	0.024		
		longer	0.024		
		wife	0.024		
		cumberbach	0.024		

Criminal & Actor	Topic2	**series**	0.044	143	23%
		watch	0.022		
		favourite	0.021		
		holmes	0.02		
		solomin	0.019		
		fan	0.019		
		music	0.019		
		watson	0.019		
		vasily	0.019		
		youtube	0.019		
Feature & Actor	Topic3	movie	0.047	159	25%
		best	0.029		
		role	0.029		
		version	0.028		
		talkie	0.028		
		eagels	0.028		
		paramount	0.028		
		letter	0.028		
		nominated	0.028		
		jeanne	0.028		
SciFi	Topic4	ca	0.05	61	10%
		urge	0.05		
		people	0.05		
		meet	0.05		
		convention	0.05		
		attendee	0.05		
		firefly	0.05		
		stuff	0.05		
		resist	0.05		
		really	0.05		
Feature & War	Topic5	world	0.018	83	13%
		series	0.015		
		like	0.013		
		actress	0.012		
		sky	0.012		
		scene	0.01		
		normandy	0.008		
		brother	0.008		
		company	0.008		
		hbo	0.008		

表 15 2019.03.18-2019.03.24 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Animation	Topic1	good	0.045	70	17%
		instantly	0.043		
		maid	0.043		
		datin	0.042		
		programmer	0.042		
		geek	0.042		
		intorvert	0.042		
		miss	0.042		
		kobayashi	0.042		
		dragon	0.042		
Comedy & War	Topic2	character	0.067	81	19%
		popular	0.05		
		actor	0.031		
		homepage	0.031		
		holocust	0.029		
		toner	0.029		
		video	0.025		
		pop	0.022		
		vudu	0.022		
		base	0.022		
Documentary & SciFi	Topic3	watch	0.12	130	31%
		time	0.1		
		speed	0.096		
		aspect	0.093		
		mankind	0.093		
		future	0.09		
		commercial	0.016		
		hulu	0.016		
		documentary	0.008		
		ai	0.008		
Comedy & SciFi	Topic4	work	0.045	141	33%
		decribed	0.045		
		comedian	0.045		
		screen	0.045		
		battle	0.045		
		monster	0.045		
		seller	0.045		
		peter	0.045		
		depression	0.045		
		manic	0.045		

表 16　2019.03.25-2019.03.31 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Criminal & Character	Topic1	experience	0.056	96	13%
		tv	0.052		
		depicted	0.052		
		bad	0.052		
		prison	0.051		
		pain	0.051		
		violence	0.051		
		separated	0.051		
		heatache	0.051		
		witnessed	0.01		
Comedy & Actor	Topic2	hazzard	0.045	201	27%
		episode	0.045		
		john	0.045		
		ran	0.045		
		salary	0.045		
		wopat	0.045		
		tom	0.045		
		dispute	0.045		
		pretty	0.045		
		thing	0.045		
SciFi & Actor	Topic3	information	0.021	216	29%
		happens	0.016		
		character	0.014		
		watch	0.014		
		fictional	0.012		
		nature	0.012		
		khardshian	0.012		
		use	0.012		
		project	0.012		
		software	0.012		
SciFi & Star Trek	Topic4	gravity	0.026	226	31%
		star	0.022		
		movie	0.019		
		trek	0.018		
		science	0.018		
		artificial	0.016		
		production	0.014		
		commercial	0.013		
		fiction	0.011		
		cast	0.01		

表 17 2019.04.01-2019.04.07 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
TV Show & Comedy	Topic1	thing	0.031	102	19%
		comedy	0.027		
		distant	0.026		
		smile	0.021		
		simple	0.021		
		tera	0.021		
		monkey	0.021		
		sharma	0.021		
		mimicry	0.021		
		kapil	0.021		
SciFi & Thriller	Topic2	movie	0.03	104	19%
		like	0.018		
		think	0.018		
		stranger	0.018		
		want	0.017		
		real	0.014		
		thing	0.014		
		body	0.011		
		supes	0.011		
		demodogs	0.011		
Comedy & Affectional	Topic3	movie	0.026	85	16%
		actor	0.025		
		mcluhan	0.017		
		woody	0.016		
		line	0.016		
		theory	0.016		
		annie	0.016		
		alvy	0.016		
		singer	0.016		
		hall	0.016		
TV Show & Host	Topic4	question	0.017	85	16%
		missed	0.015		
		example	0.014		
		jeopardy	0.014		
		diagonsed	0.013		
		probabaly	0.013		
		routine	0.013		
		host	0.013		
		alex	0.013		
		stage	0.013		

Comedy & Thriller	Topic5	movie	0.037	166	30%
		thing	0.021		
		star	0.02		
		miss	0.019		
		based	0.019		
		rodriguez	0.017		
		nanjano	0.017		
		bala	0.017		
		gina	0.017		
		comedy	0.016		

表 18 2019.04.08-2019.04.14 主题分类结果

主题标签	主题号	主题下热词	热词词权	主题出现频率	主题热度占比
Comedy	Topic1	episode	0.04	55	9%
		guy	0.039		
		family	0.038		
		maintain	0.035		
		quo	0.028		
		adultoriented	0.028		
		change	0.026		
		phenomenon	0.026		
		simpson	0.026		
		favourite	0.026		
Criminal	Topic2	people	0.045	182	31%
		carrer	0.044		
		episode	0.042		
		city	0.041		
		weapon	0.041		
		cop	0.041		
		firing	0.041		
		shooting	0.041		
		criminal	0.041		
		police	0.041		
SciFi & The Avengers	Topic3	thanks	0.015	262	44%
		world	0.014		
		frank	0.014		
		paul	0.013		
		harris	0.013		
		invisible	0.013		
		hemsworth	0.013		
		bana	0.013		
		liam	0.013		
		jordan	0.013		

SciFi & Fantastic Four	Topic4	standard	0.028	95	16%
		fiction	0.028		
		examine	0.028		
		structure	0.028		
		informative	0.028		
		drama	0.028		
		cast	0.028		
		chikills	0.017		
		jessica	0.017		
		hero	0.017		

参考文献：

[1] Harper F M, Raban D, Rafaeli S, et al. Predictors of answer quality in online Q&A sites[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008: 865-874.

[2] Liu G, Wei Y, Li F. Understanding Consumer Preferences---Eliciting Topics from Online Q&A Community[J]. 2018.

[3] Wang G, Gill K, Mohanlal M, et al. Wisdom in the social crowd: an analysis of quora[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 1341-1352.

[4] 屠守中,闫洲,卫玲蔚,朱小燕.异构社交网络用户兴趣挖掘方法[J/OL].西安电子科技大学学报:1-6[2018-12-27].http://kns.cnki.net/kcms/detail/61.1076.TN.20181217.1102.002.html.

[5] 沈波,赖园园.网络问答社区“Quora”与“知乎”的比较分析[J].管理学刊,2016,29(05):43-50.

[6] 姚鹏燕. 基于社会网络分析的用户协作行为研究[D].北京邮电大学,2013.

[7] Rughiniş R, Marinescu-Nenciu A P, Matei Ş, et al. Computer-supported collaborative questioning. Regimes of online sociality on Quora[C]//2014 9th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2014: 1-6.

[8] Paul S A, Hong L, Chi E H. Who is authoritative? understanding reputation mechanisms in quora[J]. arXiv preprint arXiv:1204.3724, 2012.

[9] Patil S, Lee K. Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors[J]. Social network analysis and mining, 2016, 6(1): 5.

[10] Kumar A, Praveen S, Goel N, et al. Opinion Extraction from Quora Using User-Biased Sentiment Analysis[M]//Information Systems Design and Intelligent Applications. Springer, Singapore, 2018: 219-228.

[11] Liu G, Wei Y, Li F. Understanding Consumer Preferences---Eliciting Topics from Online Q&A Community[J]. 2018.

[12] Jiang H, Qiang M, Zhang D, et al. Climate Change Communication in an Online Q&A Community: A Case Study of Quora[J]. Sustainability, 2018, 10(5): 1509.

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[14] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.

[15] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[C]//European conference on information retrieval. Springer, Berlin, Heidelberg, 2011: 338-349.

[16] 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. Association for Computational Linguistics, 2009: 248-256.

[17] Hong L, Davison B D. Empirical study of topic modeling in twitter[C]//Proceedings of the first workshop on social media analytics. acm, 2010: 80-88.

[18] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.

[19] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[C]//European conference on information retrieval. Springer, Berlin, Heidelberg, 2011: 338-349.

[20] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 1445-1456.

[21] Maity S K, Sahni J S S, Mukherjee A. Analysis and prediction of question topic popularity in community Q&A sites: a case study of Quora[C]//Ninth International AAAI Conference on Web and Social Media. 2015.

[22] Barua A, Thomas S W, Hassan A E. What are developers talking about? an analysis of topics and trends in stack overflow[J]. Empirical Software Engineering, 2014, 19(3): 619-654.

[23] Zou J, Xu L, Guo W, et al. Which non-functional requirements do developers focus on? an empirical study on stack overflow using topic analysis[C]//2015 IEEE/ACM 12th Working Conference on Mining Software Repositories. IEEE, 2015: 446-449.

[24] 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. Association for Computational Linguistics, 2011: 262-272.