文 献 与 数 据 学报
第 2 卷　第 1 期　　2020 年 3 月

用户生成内容的图书主题标签研究

——以豆瓣读书用户生成评论为例

陈　婵
（中共武汉市委党校图书馆，武汉　430024）

摘　要：[**目的 / 意义**] 为社交网络平台管理者提供一种分析和优化主题标签的方法，以帮助用户在社交网络平台上准确地获取图书的相关信息，满足用户个性化检索的需求。[**方法 / 过程**] 本文提出了一种基于用户生成内容（UGC）的主题分析方法。以社交平台“豆瓣读书”为例，选取《平凡的世界》和《围城》两本经典图书，首先爬取用户对该书的评价数据，然后对数据进行清洗，基于 Latent Dirichlet Allocation（LDA）主题分析方法对数据进行分析，以获取图书的相关主题标签。[**结果 / 结论**] 通过基于用户生成内容对图书的主题进行分析，一方面完善了图书的标签，提高用户对书籍的查准率，另一方面用户生成内容中具有鲜明的主题性和情感倾向，因此豆瓣读书制作标签时可以考虑增加情感类词，提高社交网络平台的个性化推荐功能。

关键词：UGC　图书标签　LDA
分类号：G250
DOI：10.31193/SSAP.J.ISSN.2096-6695.2020.01.08

0　引　言

随着互联网的迅速发展，现代社会逐渐步入大数据时代。以互联网为依托的各种新媒体的出现较传统媒体而言给人们的生活带来了极大的便利，使得人们能够在新媒体环境下更加自由地生活。大量的用户在微博、豆瓣等新媒体平台发表大量其在工作、生活、学习中总结的经验、诀窍等知识内容，即用户生成内容（UGC），成为人们获取知识的主要来源。UGC 中有关某个实体的知识通常是由不同的用户创造出来，并以碎片化的形式发布在不同的 UGC 平台上，每个知识碎片都是用户个体经验的总结，具有主观片面性，是对某个实体的主观感受。豆瓣是一个集博客、

[作者简介] 陈婵（ORCID：0000-0003-3223-4824），女，助理馆员，硕士，研究方向为数字图书馆与知识服务，Email：263993314@qq.com。

交友、小组、收藏于一体的新型社区网络，已经被公认为中国 Web2.0 时代最纯粹、最精彩的先锋网站。豆瓣对用户标注频率较高的标签进行了分类，包括文学、流行、文化、生活、经管和科技六大类，大类下涵盖 143 个热门标签。其中，在豆瓣读书中随意搜索一本书，系统会出现 8 个常用的标签供用户参考。8 个标签的内容具有一定的规范性，标签通常为作者、书名、时代、题材、国别。然而，图书标签的制定存在很多问题。一方面，标签的繁杂使得豆瓣读书中存在标签冗余、语义重复、专指度不高、缺乏准确性等问题[1]。如张爱玲的《半生缘》同时具有“文学”“中国文学”“文学经典”3 个标签，而这三个不同的标签在语义上就有交集，存在包含关系，而导致标签的专指度不高；金庸的《笑傲江湖》出现了“小说”、“武侠”和“武侠小说”3 个标签，从含义上讲，这 3 个标签是重复的。外国文学《复活》用户常用标签有一个是“经典”，这个标签缺乏准确性，不能准确反映该小说的主题特征。不规范的标签会占用大量的资源，一方面给标签的管理者带来不便，另一方面也会造成用户不能找到合适的书籍，以及读者的评论内容和系统提供的主题标签很多时候是不相符的情况。而读者在选择图书的时候往往会参考其他读者的评论，但 UGC 是碎片化的知识，读者无法有效获取其他读者对书的评论。因此，对于读书爱好者来说，要从庞大的网络资源中选取出自己想看的书并非易事。大部分读者在读完书后都会有自己的想法，不可避免地产生大量对书的主观评论内容，但是用户所发表的评论是围绕同一本书的，其宗旨是不变的。

基于此，本文提出一种基于用户生成内容的图书主题标签研究方法，通过分析用户所发表的评论，基于 LDA（Latent Dirichlet Allocation）主题生成模型来提取图书主题标签，进而对图书标签进行有效管理，改善图书标签存在的问题，使其能以读者最需要的方式呈现。

本研究通过将用户生成内容分析结果作为豆瓣读书标签信息的有效补充，有利于网站个性化推荐；同时可以帮助豆瓣读书用户了解某一本书的总体用户认同度，而不必大量浏览冗余的文本评论，为用户寻找图书节省时间精力，提高社会环境下搜索的结果。商家通过分析的数据可实现对读者的精准推荐，提高用户对小说网站的满意度，提高社会环境下搜索结果的准确性。另外，丰富了主题分析的理论体系，拓展了 UGC 的应用领域与范围。

1　相关研究现状

1.1　UGC 相关研究

UGC 即用户生成内容，是指社会化媒体平台的用户通过网络自发地进行文字、音频、视频等信息的创作并通过互联网分享和扩散，体现的是网络时代信息开放资源共享的时代精神，典型的 UGC 应用代表有微博、豆瓣、优酷、维基、分答、知乎等社交媒体平台。

有关 UGC 的研究受到相关研究学者的广泛关注并取得大量研究成果，本文主要对 UGC 应用方面的研究进行阐述。Chevalier 等人搜集了图书名、价格、评论数量、排名星级等数据，研究了消费者评论对书籍的相对销售量的影响，并运用模型对影响图书的各种因素进行分析。他们发现，图书评论的改进会导致该网站相对销售量的增加，且对于大多数研究样本，图书一星评价的影响大于五星评价的影响[2]。褚晓敏等人根据电影的简介和短评（UGC）进行电影标签自动推

荐，并使用 FudanNLP 进行分词处理，用分类器进行分类训练，最后融合基于不同类型文本的标签推荐的结果[3]。罗培铭以小红书为例，从技术、个人与社会三个维度探讨在小红书社区中对用户生成内容造成积极影响的因素[4]。

1.2　文本标签相关研究

标签，即标志目标的分类或内容。标签反映着用户对资源的认知，方便用户查找其所需的信息。M. Hu 等人使用词典等新方法挖掘客户评论过的产品特征，并在每个评论中识别观点句子以确定每个观点句子是正面的还是负面的[5]；Christoph Trattner 等人对基于标签的信息访问进行了研究，结果表明基于标签的浏览界面在性能和用户满意度方面都明显优于传统的搜索界面[6]；李丕绩等人通过对句法分析，对句子 K-means 聚类进行语义去重以及 LDA 主题分析，为每个实体抽取特征标签[7]；邓莎莎等人以淘宝商城上商品评论数据为例，通过 LDA 聚类方法提取商品主题并分析商品评论[8]；熊回香、叶佳鑫基于 LDA 主题模型对用户关注的人及用户粉丝的微博进行主题分析，生成微博用户标签，进而较为准确地描述用户的微博特征[9]。

当前，各新媒体平台上已经积累了海量的 UGC 资源，然而人们对其开发利用尚不充分，不利于提高用户使用满意度。本文根据用户生成内容的特点，提出了基于用户生成内容的主题分析研究，以豆瓣读书为例，根据用户的评价对图书标签进行完善，以提高用户对网站使用的满意度。

2　相关理论和技术

2.1　理论基础

文本分类是按照预先定义好的分类体系，根据文档的内容和属性，将文档集合中的每一个文档归入一个或多个类别的过程[10]。文本分类算法包括训练集和测试集。通常将大量已分类数据作为算法的训练集，得到一个分类器，然后用分类器对测试集进行分类。目前分类算法有 K- 近邻算法、决策树、Logistic 回归、朴素贝叶斯以及基于 TF-IDF 的算法等。

2.2　相关技术

2.2.1　中文分词方法

对于英文文档而言，词与词之间的分隔是通过特定的间隔标记符号实现的，例如空格和标点符号等，所以遍历文档，就能够实现英文文档的分词，并获得单词列表。但是汉字是一种象形会意语言，字与字之间，词与词之间的组合灵活多变，且字与字之间，词与词之间没有明确的分隔标记，加上汉语词汇存在一词多义、多词一义等现象，这使中文分词较为困难。比如：“发展中国家兔的饲养”一句，可能导致两种分隔结果：发展中国家 / 兔 / 的 / 饲养；发展 / 中国 / 家兔 / 的 / 饲养。无论将词语怎么划分，都会缺失部分语义信息。要实现比较好的分析结果，就要充分理解语句的真正含义。如果将句子切分成单汉字，则会丢失以词为基础的许多重要信息。一般来说，我们可以用概率分词法、语法分词法以及神经网络分词法对文本分词。以下简单介绍这三种分词方法的基本原理。

（1）概率分词法是基于基本分词词典的统计分词方法。这种方法要先构建词典，利用字与

字相邻共现的频率或概率来反映成词的可信度，也就是对语料中相邻共现的各个字的组合进行统计，计算共现频率，超过一定阈值的，则认为此字组可能构成了一个词。这种方法切分速度快，而且效率高。

（2）语法分词法通过对自然语言文法或句型文法的分析来抽取主题词加以标引。比如，从专门文献中挑选形如“本文讨论了……”这样的特征句型。识别出这类特征句子后，由自动抽词处理器对句子抽词并进行加权处理。这种分词方法大多受限于科技文献的标题与文摘，但易于描述与归纳。

（3）神经网络分词法是在模拟人脑结构和行为的基础上，用大量简单的处理单元广泛连接组成复杂网络。基于目前这种方法的不成熟，神经网络分词方法研究采用的样本集大多都是小范围的文本，知识表示规则不全。

2.2.2 停用词过滤

本文挖掘到 UGC 上千条，其包含大量的词项。评论内容中并不是每个字都有研究含义。文档中包含了大量的停用词。这里所谓的停用词，就是那些非常常见，但是没有信息含量的助词、介词、连词以及数字符号等。如“啊”“在”“的”“和”等，这些词也可称作虚词，包含副词、冠词、代词等，在文档构成中使用十分广泛，却难以对文档分类提供帮助，进行文本分析，我们经常需要对停用词进行剔除。剔除停用词首先建立停用词表。停用词表需要手动建立，因为文本内容的不同，以及想要生成的结果不同，所以在其他文本中不常用的词可能在本篇文档很重要。其次根据不同的分词方法将文档与停用词表中的词匹配，匹配成功则继续下一步，不成功则输出词，这就是去停用词后的结果。

2.2.3 文本主题词提取与分析

本文采用 LDA 提取主题词。LDA 主题模型是一种基于概率型的主题模型发现，能够提取文本隐含的非监督学习模型，是主体模型中的代表[11]。LDA 模型假设，文档中存在 K 个潜在主题，每篇文章都可以看作所有主题的混合概率分布。文档主题服从 Dirichlet 分布，LDA 主题模型将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。因此，本文选用 LDA 模型挖掘读者评论的主题信息。当有 N 篇文档、M 个主题和 L 个单词时，在一篇文档中的第 i 个单词属于主题 N 的概率可以表示为公式（1）[12]：

$$P(Li) = \sum k=1MP(Li \mid ni=j)P(ni=j) \quad (1)$$

其中，参数 n 代表主题，i 代表单词。

3 实 证

本文提出的基于 UGC 的主题分析研究方法，以豆瓣读书为例，首先爬取图书的用户评论内容，即 UGC 数据，然后对 UGC 数据进行分词预处理，过滤停用词，删除无意义的词以降噪，然后基于 LDA 进行主题词抽取，最后利用 matplotlib 对获得的主题词进行词云展示以获取图书主题词标签，使得分析结果更加清晰。其研究方法如图 1 所示。

图 1　基于 UGC 的主题分析研究方法流程

3.1　UGC 数据来源

为了获取大量的 UGC，本文选用豆瓣读书中《平凡的世界》和《围城》两本经典图书的 UGC 作为数据源，分别采集到 5344 条和 7540 条 UGC 数据。通过对无效数据的重复清洗和整理，得到有效评论 4142 条评论和 6222 条。本文主要以《平凡的世界》为主线进行分析。

3.2　UGC 数据的分词预处理

本文选择 jieba 分词工具进行语料的分词处理。其中 jieba 分词自带了一个叫作 dict.txt 的词典，里面有 2 万多条词，包含了词条出现的次数（这个次数是基于人民日报语料等资源训练得出来的）和词性。其算法实现：基于字典树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法抽取出的数据如图 2 所示，以《平凡的世界》为例。

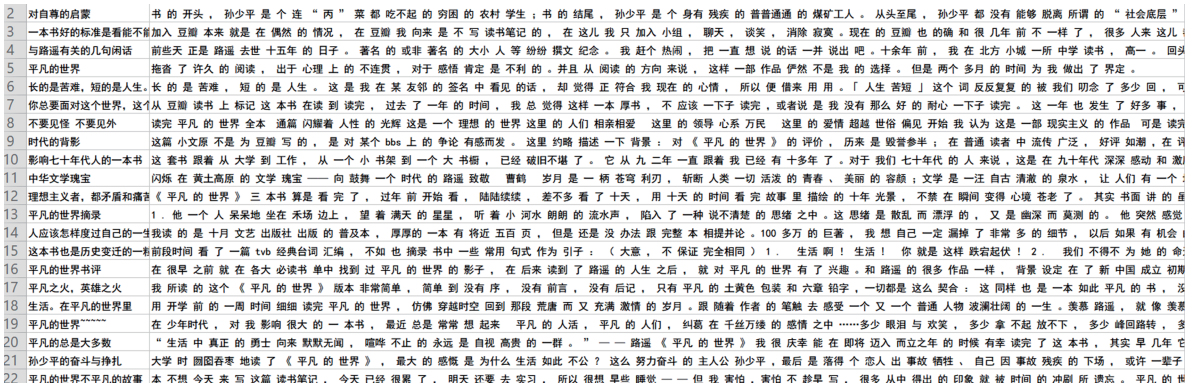

图 2 《平凡的世界》分词处理结果

3.3　降噪处理

对爬取的 UGC 数据分词后，为保证数据的质量，降低无关数据或噪声数据对结果的影响，需要对抓取到数据进行预处理[13]。首先剔除包含特殊字符与数字字符的评论，只保留重复评论中的一条。通过 TF-IDF 算法对《平凡的世界》提取的 50 个主题词如图 3 所示。

由图 3 可知有很多数字字符或是没有意义的词，如“1978”“现在”等。因此数据提取后需要构建停用词表，通过对挖掘到的用户评论的分析，确定评论文本分析结果的主题，通过停用词表删除不符合条件的词语。构建好的停用词以及字符有 2655 个。用该停用词表作为初始停用词

表，根据多次主题分析结果，对初始停用词表进行扩展，增加主题分类实验中出现的对于主题分类无意义的高频词，如：“孙少平”“田晓霞”“平凡的世界”“世界”等。这些词出现频率很高，但对于分析文中的其他主题词会有影响，所以需要经过不断删除。

```
田润叶  孙少安  平凡  爱情  女方  追求  文革时期  因为  润叶  我们  中国  门当户对  家室  男
方  小说  年收入  世界  心爱  精神  金钱  幻想  悲剧  情感  自己  富有  制约  兴趣  享受  感
情  现实  28  释卷  对润叶  1978  12  18  22  0.4  50  30  万过  现在  路遥  愿意  家庭
物质  女大当嫁  基尼系数  穷小子  可能
```

图 3　TF-IDF 算法提取《平凡的世界》主题词

3.4　主题词抽取

基于 LDA 算法对《平凡的世界》的用户评论数据抽取主题词，分成 15 个主题（Topic），每个主题抽取 10 个关键词，共 150 个关键词。从图 4 中可以发现，主题词中的情感词有“痛苦”、“苦难”、“幸福”及“善良”等；小说类型有“爱情”、“文学”、“历史”和“现实主义”等；小说描写的时代以及地区的词有“中国”、“改革开放”和“双水村”等。

```
Topic #0:
孙家 兄弟 优秀 女儿 田家 塑造 爱上 老婆 感动 儿子
Topic #1:
生命 苦难 幸福 经历 命运 美好 活着 痛苦 内心 也许
Topic #2:
爱情 现实 选择 润叶 幸福 喜欢 美好 也许 亲情 放弃
Topic #3:
爱情 润叶 家庭 金波 姑娘 双水村 农民 兰香 工作 命运
Topic #4:
中国 历史 语言 读过 体验 主题 有人 喜欢 描写 事实
Topic #5:
社会 青年 现实主义 中国 苦难 书评 美好 劳动 眼睛 爱情
Topic #6:
u3000 爱情 田小霞 女子 男人 是因为 红梅 贺秀莲 女人 金波
Topic #7:
劳动 奋斗 社会 痛苦 普通人 生命 苦难 命运 爱情 精神
Topic #8:
记得 大学 喜欢 高中 朋友 时间 同学 工作 老师 家里
Topic #9:
中国 社会 历史 农民 文学 这部 描写 改革开放 写作 城市
Topic #10:
精神 自尊 独立 尊重 高贵 欣赏 高尚 世俗 环境 想起
Topic #11:
喜欢 书中 真实 这部 感动 描写 情节 感受 也许 经历
Topic #12:
命运 农民 精神 改变 思想 社会 努力 家庭 选择 经历
Topic #13:
少年 美好 爱情 善良 忘记 苦难 花朵 沉重 生命 坐在
Topic #14:
真的 电视剧 主角 悲剧 好看 情节 剧情 好像 电影 发现
```

图 4　LDA 算法提取《平凡的世界》主题词

3.5　词云展示

词云将词语按照一定顺序和规律进行排列，如按照频度递减或者字母顺序排列，并以文字的大小代表词语重要性[14]。通过词云分析文本，帮助用户简明扼要地了解文本内容。同时利用信息可视化技术，使内容表达的形式更为具体、生动。本文通过 matplotlib 将数据预处理后的词以词云的形式展现出来，方便读者了解文章的内容，图 5、图 6 分别对《平凡的世界》《围城》做词云展示。从图 5 中可以看出重要的词有“生活”“爱情”“农民”“社会”等，反映出用户对小说的关注角度。从图 6 中我们可以发现“钱钟书”“围城”“婚姻”“爱情”等主题词。

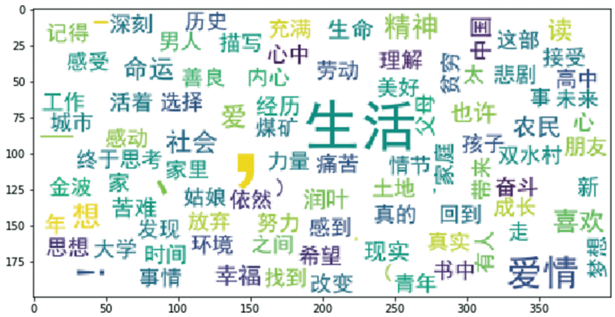

图 5 《平凡的世界》的主题词词云

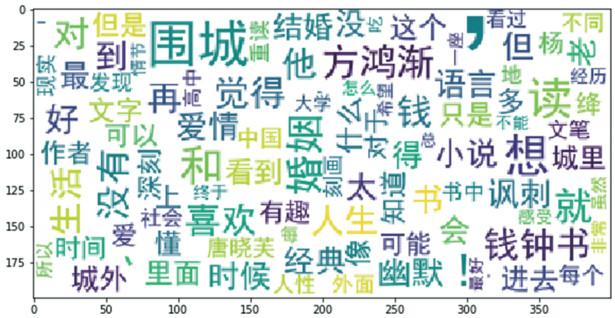

图 6 《围城》的主题词词云

从图 5 中搜索的主题词中可以发现，《平凡的世界》这本书的标签包含“文学”“中国”“农民”“爱情”“悲剧”。由此可以推断用户对文章的评论点大多基于书中主人公情感的触动。而且根据主题词可以推断出这篇文章可能描写的是中国农民或是普通人的爱情题材的小说，文章的结尾为悲剧。豆瓣读书对这本书的标签为“路遥”“平凡的世界”“小说”“中国文学”“人生”“经典”“当代”和“文学”等。通过比较发现，用户评论的主题词比豆瓣读书提供的增加了情感标签，对小说题材描述更为贴切。豆瓣读书增加了书名与作者，因为本文在做《平凡的世界》词云时，删除了作者、书名和书中人名等，所以词云中这部分内容没有体现。但本文做《围城》的词云，这部分内容并没有剔除。

通过图 6 发现，《围城》这本书的标签有“钱钟书”“围城”“方鸿渐”“婚姻”“中国”“经

典”“幽默”等词。《围城》这本书的标签为“钱钟书”“围城”“小说”“中国文学”“经典”“婚姻”“现代文学”“文学”等。根据展示结构，豆瓣读书标签为书名、作者、地区、类型、经典以及名著等词，除去书名及作者外，用户评论主题词更为具体，且增加了情感类的词。豆瓣读书在制作图书标签时可以做得更为具体，比如《平凡的世界》时代背景可以表述成“七八十年代”，小说类型可以加入“现实主义”，加入小说情感词“悲剧”“喜剧”等。

4 结 语

越来越多网站提供了标签评论功能，如京东商城、国美在线、苏宁易购等，用户既可以自定义评论标签，也可以直接使用热门的评论标签。针对豆瓣读书评论存在的不足，本文认为根据用户生成内容完善实体标签具有一定的研究价值，因此提出了从用户生成内容的角度定义标签。本文主要是通过爬虫工具爬取《平凡的世界》的用户评论，对 UGC 数据进行预处理，然后用 LDA 算法提取主题词并以词云的形式展示结果，利用信息可视化技术让内容表达的形式更为简明、直观。研究发现，通过分析用户生成内容，一方面完善了图书的标签，提高了用户对书籍的查准率，另一方面用户生成内容具有鲜明的主题性和情感倾向，因此豆瓣读书制作标签时可以考虑增加情感类词，提高网站平台的个性化推荐功能。

本文的不足在于停用词表建得不完善，去停用词后的效果不理想，缺少情感分析模块，对于用户所关注的角度缺乏进一步研究。笔者在今后的研究中，将引入情感分析，这也是一个有价值的方向。

【参考文献】

［1］叶继元 . 信息组织［M］. 北京：电子工业出版社 , 2015.

［2］CHEVALIER J A, MAYZLIN D. The effect of word of mouth on sales: Online book reviews［J］. Journal of Marketing Research, 2006, 43(3): 345-354.

［3］褚晓敏 , 王中卿 , 朱巧明 , 等 . 基于简介和评论的标签推荐方法研究［J］. 中文信息学报 ,2015,29(6):179-184.

［4］罗培铭 . 虚拟社区用户生成内容的影响因素——以小红书为例［J］. 新闻研究导刊 ,2018,9(12):60-61.

［5］HU M, LIU B. Mining and summarizing customer reviews［C］//Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2004: 168-177.

［6］TRATTNER C, KAPPE F. Social stream marketing on Facebook: a case study［J］. International Journal of Social and Humanistic Computing, 2013, 2(1-2): 86-103.

［7］李丕绩 , 马军 , 张冬梅 , 等 . 用户评论中的标签抽取以及排序［J］. 中文信息学报 , 2012, 26(5): 14-19.

［8］邓莎莎 , 袁菱 . 商品评论主题分析研究［J］. 上海电力学院学报 ,2013,29(6):549-552+567.

［9］熊回香 , 叶佳鑫 . 基于 LDA 主题模型的微博标签生成研究［J］. 情报科学 ,2018,36(10):7-12.

［10］薛春香 , 张玉芳 . 面向新闻领域的中文文本分类研究综述［J］. 图书情报工作 ,2013,57(14):134-139.

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

［12］唐晓波 , 向坤 . 基于 LDA 模型和微博热度的热点挖掘［J］. 图书情报工作 ,2014,58(5):58–63.
［13］唐晓波 , 邱鑫 . 面向主题的高质量评论挖掘模型研究［J］. 现代图书情报技术 ,2015(Z1):104–112.
［14］唐家渝 , 孙茂松 . 新媒体中的词云 : 内容简明表达的一种可视化形式［J］. 中国传媒科技 ,2013(11):18–19.

Research on Book Topic Tags Based on UGC:Taking the Social Network Platform Douban Reading as an Example

CHEN Chan

(Library of Party School of the C.P.C Wuhan Municipal Committee, Wuhan 430024, China)

Abstract: ［**Purpose/meaning**］This paper provides a way to analyze and optimize book topic tags for social network platform manager, in order to help users to obtain relevant information of books accurately on the social network platform, and meet the needs of users for personalized retrieve. ［**Method/process**］This paper proposes a topic analysis method based on user generated content (UGC),takes the social platform *Douban Reading* as an example and selects two classic books, *Ordinary World* and *The Besieged City*. The process includes: (1) climbing the UGC data of the book;(2)cleaning the data;(3) analyzing the data based on the Latent Dirichlet Allocation (LDA) theme analysis method to obtain the relevant topic labels of books.［**Result/conclusion**］By getting and analyzing the topic tags of the books based on the UGC, on the one hand, the tags of the books are improved, and the precision of the user's book information retrieve is improved. On the other hand, the UGC has a distinct theme and emotional tendency. Therefore, when making topic tags, *Douban reading* can consider adding emotional words to improve the personalized recommendation function of the social network platform.

Keywords: UGC; Book tags; LDA

（本文责编：周　霞）