基于改进K均值算法的移动图书馆用户评论需求聚类研究*

郑德俊，朱婷婷，沈军威
（南京农业大学信息管理系，南京 210095）

摘要：对移动图书馆用户评论的自动聚类研究有助于更准确高效地获取用户需求。本文结合移动图书馆评论特征，在传统K均值算法的基础上，使用HT-LaD算法对初始聚类中心进行算法改进，并使用移动图书馆的用户评论数据进行实证。结果表明，利用改进后K均值算法完成移动图书馆用户评论文本的需求聚类是可行的，且聚类精度和稳定性得到提高。

关键词：移动图书馆；改进K均值聚类；用户评论；用户需求
中图分类号：G251
DOI：10.3772/j.issn.1673-2286.2017.10.005

1 引言

移动图书馆用户需求一直是移动图书馆的研究热点。已有研究主要聚焦于用户需求兴趣点的发现[1]、需求类型[2-3]和用户需求特征分析[4]、需求分析方法应用[5]、用户需求模型研究[6]等。目前，除基于问卷调查的需求获取方法被普遍采用外，移动图书馆平台的用户评论反馈也日益受到重视，研究者认为用户评论有助于发现用户的隐性需求及新需求[7]。移动网络信息服务环境下，与移动图书馆有关的用户评论数量增长迅速，依靠人工进行用户需求甄别与发现费时费力。因此，有必要借助一定的技术手段，引入文本挖掘方法，进行去重、筛选和分类以识别用户需求。

已有研究将文本分类应用于移动图书馆用户评论研究的成果较少，但相近领域的文本挖掘研究（涉及关联分析、文档分类、聚类和自动文摘[8]等）值得借鉴。如倪瑜泽等提出一种需求发现方法——DICM，对预处理后的用户评论文本进行基于信息增益的特征选择，利用朴素贝叶斯分类器对潜在演化需求进行分类[9]；崔建苓等通过采用本体和条件随机场模型融合的特征提取方法，对潜在软件需求进行汇总[10]。基于移动图书馆用户评论需求挖掘用户隐性需求（尤其是新需求），结果充满不确定性，因此本文拟采用文本聚类分析，为自动化识别和判定移动图书馆用户需求提供支持。

2 移动图书馆用户评论特征分析

2.1 数据采集与清洗

目前，国内移动图书馆主要有两种形式，一是购买商业公司的移动图书馆APP服务，另一种是自建移动图书馆服务平台。受限于图书馆自有技术团队和后期维护水平，国内绝大多数图书馆的移动图书馆服务以购买为主，相应的用户评论数据存储在商业公司服务器上，一般很难被公开查询。2014年以来，在国内某知名商业公司移动图书馆服务平台的支持下，本文获取了26 976条数据，并对评论数据进行错别字校正和繁简字体转换。

本文的目标是通过用户评论获取用户需求，需要对评论数据中包含的情感性评论进行过滤。目前在移动图书馆领域，尚没有完整的移动图书馆专属词典，而词典直接影响无效评论过滤和中文分词的效果，进

* 本研究得到国家社会科学基金项目“基于用户感知的移动图书馆服务质量评价及提升策略研究”（编号：13BTQ026）资助。

而影响聚类效果。本文首先构建移动图书馆领域专属词典，该词典的词汇涉及移动图书馆APP多方面服务，收词53 940条。具体构建思路：（1）对移动图书馆用户评论语料进行切分和词频统计，构造基础词典；（2）借鉴图书情报领域主题词表，经人工判别后添加到基础词典；（3）基于中国知网期刊数据库中与“移动图书馆”相关的文章摘要和关键词，通过citespace进行分析，添加到基础词典；（4）考虑到用户评论语言口语化、随意性强，适度融入搜狗输入法、紫光输入法等词库中的相关词汇；（5）基于哈尔滨工业大学《同义词林》对同义词、近义词进行扩展。移动图书馆领域专属词典结构形式为“词条+属性”，如“电子书 n”，其中n为名词。通过多渠道扩大收词来源，确保所构建的移动图书馆领域专属词典全面和实用。本文以专属词典中有实际意义的词汇作为有效词汇，计算各词汇在评论中的占比，设定合理阈值，进行无效评论的过滤工作。经过数据清洗，共得到18 869条有效用户评论数据，对移动图书馆用户评论数据的特征分析主要基于所获取的有效数据进行。表1列出清洗后的用户评论的长度数据。

表 1 移动图书馆用户评论字数统计

字 数	评论数/条	占 比
≤200字	18 823	99.8%
>200字	46	0.2%
总计	18 869	100%

少于等于200字的用户评论比例高达99.8%，龚才春将少于等于200字文本定义为短文本[11]。从整体数据看，移动图书馆用户评论属于短文本，具备长度短，信息量少的特点，但不拘泥语法，存在拼写错误[12]。

2.2 特征分析

移动图书馆用户评论与一般APP评论存在差异。本文将移动图书馆用户评论与一般APP评论（起点读书、豆丁书房、手机知网）作出对比，借鉴李亚松的思路[13]，并结合APP自身服务性质和特点，从评论主体、评论自身分析移动图书馆评论的独有特征，如表2所示。具体来看，移动图书馆用户评论存在两点特征。

（1）基于评论主体的移动图书馆用户评论特征。一般APP评论主体更多元化（包括各行各业的人），用户群体的巨大差异性造成评论质量的参差不齐，其中不乏匿名评论，因而一般APP评论主体的隐匿性强，评论的真实性、可靠性有待商榷；移动图书馆评论主体多为在校师生和科研人员，且登录与学号、工号绑定，评论主体相对单一，因而所发表的评论差异性小，隐匿性小，评论内容的真实性、可靠性更高。

表 2 移动图书馆评论的差异性分析

角 度	特 征	一般APP评论	移动图书馆评论
评论主体	多元化	高（各行各业的人都有）	低（学生、教师、研究员为主）
	差异性	大	小
	隐匿性	大	小
评论自身	社交性	强	弱
	连续性	不同用户间交流	同一用户跟踪反馈
	形式	文字、图片、链接等	文字（中文为主）
	语言表达	良莠不齐	随性、简洁、不规范
	评论倾向	简单评价，表明态度	指出不足，提供建议

（2）基于评价内容的移动图书馆用户评论特征。①社交性。一般APP评论社交性较强，用户间有交流，可以回复、点“赞”或点“踩”，某些评论易导致跟风评论，还会出现因观点不同在评论中恶语相向的情况；移动图书馆服务平台的“意见反馈”模块是嵌入在APP中的，每个用户都是独立的个体，后台工作人员通过APP实现与用户的交流与反馈，因而，移动图书馆评论的用户间交流较少，社交性相对而言较弱。②连续性。一般网站用户评论的连续性体现在不同用户间的互动与交流；移动图书馆用户评论的连续性体现在同一用户在不同时间对APP的评价与心得。③评论形式。一般APP评论的形式多样化（文字、图片、URL链接等），移动图书馆评论形式单一（仅文字），且以中文为主。④语言表达。互联网环境下，一般APP的用户评论不乏语言表达粗俗，还有不法分子散布虚假信息、广告信息；移动图书馆用户评论环境相对更好，但是存在大量的语病、错别字，除使用文字外，用户还以特殊字符来表达内心情绪，有些评论还出现繁体字。受用户群影响，与一般APP的用户评论相比，移动图书馆用户评论倾向于指出不足，提供建议，更具参考性。⑤评论倾向。一般APP评论更多的是表达对产品的评价与态度，移动图书馆用户评论除表达用户对产品的评价、指出问题与不足外，还有用户对产品某方面较为具体的建议与改进策略。当然，移动图书馆的用户评论内容中也存在表达个人喜恶的泛泛评论，在面向用户需求识别时，可通过数

据清洗剔除价值较低的文本数据。

总之，由于移动图书馆评论内容的差异性，使其用户评论比一般APP的用户评论具有更高的用户需求识别价值，对用户评论的文本聚类算法提出更高要求。

3 传统K均值聚类方法局限及改进设想

Flury认为同一类簇内的实体是相似的，不同类簇的实体是相异的[14]。文本聚类根据“同类文本相似度高，不同类文本相似度低”的假设，利用无监督的机器学习方法将相似度高的文本聚合到一个簇得到聚类结果[15]。文本聚类有多种方法，如李伟等介绍了常用的文本聚类算法，并从算法适用范围、初始参数的影响、终止条件以及对噪声的敏感性等方面对各类方法进行分析比较[16]。在多种文本聚类方法中，K均值聚类算法凭借原理简单、收敛高效，应用最广泛。K均值算法是Macqueen提出的一种基于划分的聚类方法，其基本思想是在给定的数据集中，随机选择k个数据对象作为k个类的初始中心点[17]。传统K均值聚类存在3点缺陷：聚类个数k需要人工赋值、初始聚类中心选择存在随机性[18]、孤立点对聚类效果有影响[19]。

对于K均值聚类方法中初始聚类中心选择随机性的缺陷，很多研究者尝试进行改进。如Tzortzis[20]和张志祥[21]等设计minmax K均值算法，但并不能保证排除全部可能的孤立点；傅德胜等选择k个高密度区域的点作为初始聚类中心，但增加了时间复杂度[22]；Bradley选取部分数据作为样本，选择不同的初始中心点分别执行K均值算法，但是样本选择不确定且容易造成局部最优[23]；朱晓峰等提出一种基于文本平均相似度的K均值算法[24]；左进等在数据紧密的地方均匀选择k个初始中心，此法增加了算法的复杂度[19]。

现有的K均值聚类对初始中心点的改进方案与移动图书馆的应用场景有很多不同，尤其是移动图书馆评论数据的独有特征对聚类算法提出更高要求，因此需设计有针对性的改进算法，具体提出两点设想。

（1）引入紧密性参数。移动图书馆用户评论内容不仅涵盖用户简单的评价、心愿和态度，还包含用户提供的改进建议，但少数简单评论所含信息量过少，不利于分析用户的潜在需求，属于低价值评论文本。因此，在使用K均值聚类时，有必要引入紧密性参数，将具备高文本价值的评论和用户简单低价值的评论区分。

（2）调整初始中心点计算方法。由于移动图书馆用户评论多是短文本，语言表达随性，在衡量评论间相似性时，若想降低高维空间带来的影响，须将距离公式进行标准化，同时计算平均值，选取高平均值对应的文档作为初始中心点，保证初始聚类中心点的分布均匀，降低结果的波动性，提高算法稳定性。

因此，将高紧密性（High Tight，HT）与低平均距离（Low average Distance，LaD）相结合，基于HT-LaD的K均值改进算法有助于对移动图书馆用户评论进行聚类并获取用户需求。

4 算法改进的关键技术描述

4.1 算法框架

图1为本文提出的用户需求聚类研究算法框架。算法输入为待分析的用户评论源数据，主要分为用户评论预处理模块、结构化表示模块、需求聚类模块。

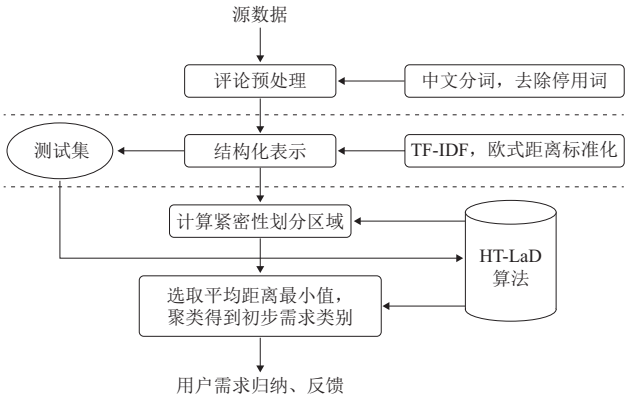

图 1 用户需求聚类研究算法框架

（1）用户评论预处理模块。基于Python 2.7平台过滤无效评论，进行繁简转换，考虑到移动图书馆的用户评论表达口语化，易产生歧义，因此调用结巴分词包对用户评论进行分词，最后基于中文停用词表完成停用词剔除工作。（2）结构化表示模块。用户评论表示模块对预处理后的用户评论进行结构化表示，本文拟利用TF-IDF计算特征词的权重方法[25]，利用标准化后的欧氏距离度量文本间相似性，以适应高维空间下的数据集。（3）需求聚类模块。通过计算文档紧密性，合理设置阈值，将用户评论划分到两个集合中，即高紧密性区域和可能的孤立值区域。在高紧密性区域计算文档标准化欧氏距离的平均值并进行排序，得到最具代表性的文档作为初始聚类中心，采用HT-LaD算法进行文本聚类，得到用户需求。

4.2 HT-LaD算法构建过程描述

本文对聚类初始中心的选择进行改进，归纳为HT-LaD算法，即通过每个文本的距离进行平均值计算和排序，将落在紧密性区域的文档作为初始聚类中心。

（1）计算评论中的特征词权重。根据TF-IDF值相乘得到移动图书馆用户评论特征词的权重（w）。

（2）计算标准化欧氏距离。文本欧氏距离将定义为：假设有文本d_1和文本d_2，一般用$dist(d_1, d_2)$表示文本间距离，两条评论数据的距离越小，说明二者越相关。如果移动图书馆的用户评论数量庞大，就容易形成高维矩阵，加入标准差的衡量可以降低文本向量长度的影响，间接起到降维作用。用公式表示即标准化后的值等于标准化前的值与分量的均值的差，再除以分量的标准差。

（3）计算平均距离集合U。在计算移动图书馆用户评论数据集中一个文本与其他所有文本间的距离后，计算其平均值作为该文本的平均距离，并进行排序，将距离较小的文本作为初始聚类中心。定义文档i与文档j的距离为d_{ij}，构建文档对角矩阵M，即$d_{ij}=d_{ji}$，$d_{ii}=0$。a_i指一个文档与其他$n-1$个文档间的距离平均值，U指平均距离集合，即a_i集合。

$$a_i=\frac{d_{i1}+d_{i2}+\cdots+d_{in}}{n} \quad (1)$$

（4）引入“数据紧密性参数”进行区域限定。数据紧密性参数计算参考左进等提出的算法[19]，公式（2）中D为移动图书馆用户评论文档集，n代表评论个数，$G_t(i)$为i的t个最近邻数据点集合。本文中有效评论为18 869条，t取值为0—18 868，t取值为0表示该评论与其他评论都不相关，该评论很显然是孤立的一条评论，即聚类中的异常值；t取值为18 868表示所有的评论数据都集中在一起，即所有评论数据都相关，分布的紧密性最高。以文档紧密性$Tigh$值作为降维依据，经过多次尝试与分析，后续的实验可设定参数t为100。进一步计算得到文档紧密性的平均值，所有小于平均值的数据点，被认为是稀疏数据点，予以删除，最后得到紧密数据点集合U1。

$$Tigh<\frac{1}{n}\frac{1}{\sum_{j\in G_t(i)}D(i,j)} \quad (2)$$

（5）对新集合U∩U1中的数据集根据平均距离排序，选择最小值作为中心点并删除与之相关的文本，多次重复上述步骤，直到有k个中心点。当k个中心点选取完成时，聚类过程也随之结束。

综上，HT-LaD算法的改进重点是初始中心的选择。移动图书馆用户评论中，紧密性的考量可避开需求识别价值小的用户评论，根据距离平均值排序后选出的中心点能更优地代表一部分数据，保证中心点分布均匀。

5 实证研究

5.1 数据来源

实证数据仍使用经过清洗后的移动图书馆用户评论数据。随机选取9 250条有效评论，邀请2名情报学硕士依据自身使用体验和已有知识积累进行人工自由标注和聚类，1名信息资源管理博士对标注结果进行审核，标注一致度超过90%。同时基于Python 2.7平台，利用改进后的算法进行聚类测试，并将机器聚类结果与人工标注结果进行对比分析。

5.2 实验结果

根据人工标注结果，发现用户评论反映的问题可概括为8个方面，因此设定聚类算法的K值为8，迭代次数为500次。根据表3可见，聚类类别1至聚类类别6反映了用户的关注点分别集中在资源丰富性、功能多样性、平台稳定性、登录快捷方便性、人性化设置和平台宣传推广等方面；聚类类别7和聚类类别8反映了用户仍关注移动客户端运行对网络流量资源的占用和消耗，移动阅读对人眼健康的影响。

表 3 聚类结果分布表

类 别	类 名	数量/条	占 比
1	资源丰富性	3 865	41.8%
2	功能多样性	1 453	15.7%
3	平台稳定性	1 027	11.1%
4	登陆方便快捷性	806	8.7%
5	人性化设置	952	10.3%
6	平台宣传推广	573	6.2%
7	流量资费考虑	361	3.9%
8	用眼健康	213	2.3%

类别1和2反映资源丰富性和功能多样性的评论数共计5 318条，占比57.5%，充分说明改进功能层面需求的广泛性；类别3和4是用户对移动图书馆服务平台表

达了更高的需求，平台稳定性和登录方便快捷需求占比19.8%，可概括为技术层面的需求；类别5和6分别反映用户在人性化设置和宣传推广方面的需求，共占比16.5%，可概括为用户关怀视角的需求。以上聚类结果分布与倪峰等研究结果相接近[7]，类别7和8反映的需求是郑德俊等研究者之前调查所未能得到的[3]。

利用BlueMC在线工具分析各类别评论数据，基于词频统计，各类别中排在前3位的有意义的实词如表4所示。

表 4 用户评论数据词频分布表

类别1		类别2		类别3		类别4		类别5		类别6		类别7		类别8	
词	词 频	词	词 频	词	词 频	词	词 频	词	词 频	词	词 频	词	词 频	词	词 频
资源	1 062	功能	453	快点	315	绑定	216	人性化	370	宣传	276	流量	363	护眼	94
图书	965	搜索	421	速度	309	门户	195	便捷	272	力度	254	节省	129	模式	82
书籍	801	下载	374	卸载	280	学校	164	个性化	179	推广	245	浪费	51	眼睛	48

从整体看，移动图书馆用户评论所反映的功能需求、技术需求和用户关怀需求，与移动图书馆服务质量测评模型保持一致[26]。本文聚类实践为以后用户需求的自动化识别提供可能。根据本文聚类结果，移动图书馆服务平台除持续在资源丰富、平台稳定、用户使用支配权方面不断努力外，还应在平台可用性方面多下功夫，通过改进系统技术减少所需流量。

5.3 算法效果的实验评价

为验证改进算法的科学性，本文采用综合评价指标加权调和平均值F[27]进行评价，F值由查全率（R）与查准率（P）计算可得。其中，聚类正确评论数与聚类评论总数的比值为P值，聚类正确评论数与数据集评论总数的比值为R值，F值有助于准确衡量聚类结果和对比聚类效果[28]。

$$F=\frac{2\times P\times R}{P+R} \qquad (3)$$

为更好地观察不同算法的性能差异，本文将9 250条有效评论平均分为5组分别计算各组的F值（见图2）。

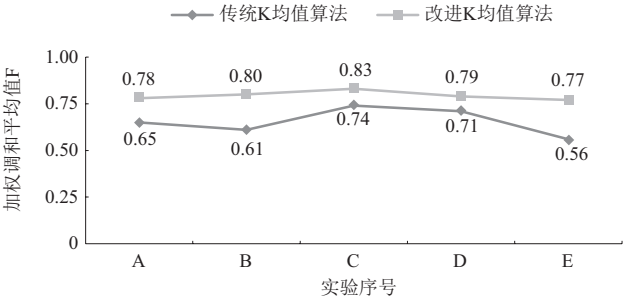

图 2 实验结果F值对比图

5组实验结果中，HT-LaD算法的F值明显高于传统K均值聚类，且聚类折线较为平缓，稳定性明显优于传统K均值聚类。通过划分区域，在数据紧密区选取初始聚类中心点，不仅降低一般短文本聚类过程中的波动，还考虑到移动图书馆用户评论内容的独有特征，降低需求识别价值低的评论文本干扰。综合考虑文档紧密性和平均距离，使其适用于移动图书馆评论语言表达的特殊性，同时保证初始聚类中心的均匀分布；对算法中欧氏距离进行标准化，更适合移动图书馆评论高维数据的相似性度量。通过本文改进后的算法，其聚类效果明显比原算法更有优势。

6 结语

基于传统K均值算法的文本聚类，操作简便，但存在稳定性较差、不适用于高维数据及聚类准确性较低等问题，本文尝试在研究该算法的具体原理后对其进行改进。结合移动图书馆用户评论高维数据的特点，将欧氏距离标准化，继而提出基于平均距离的K均值算法。考虑到移动图书馆评论的独有特征，降低需求识别价值小的评论文本干扰，加入文档紧密性衡量，确保初始聚类中心分布均匀，保证文本聚类的质量。本文对聚类算法作出改进，今后可利用所构建的移动图书馆领域专属词典辅助中文分词，结合其他研究方法（如结合条件随机场方法）深入挖掘用户需求，搭建自动用户需求获取系统，以便及时、迅速、客观地获取用户需求，并将用户需求反馈给移动图书馆服务平台加以改进，从而为用户提供更好、更优质的移动图书馆信息服务。

参考文献

[1] KARIM N S A,DARUS S H,HUSSIN R.Mobile phone application in academic library services: a students' feedback survey[J].Campus-Wide

Information System,2006,23(1):35-51.

[2] CHANDHOK S,BABBAR P.M-learning in distance education libraries: a case scenario of Indira Gandhi National Open University[J].Electronic Library,2011,29(5):637-650.

[3] 郑德俊,沈军威,张正慧.移动图书馆服务的用户需求调查及发展建议[J].图书情报工作,2014,58(7):46-52.

[4] 叶莎莎,杜杏叶.移动图书馆用户需求理论研究[J].图书情报工作,2014,58(16):50-56.

[5] RYOKAI K,AGOGINO A M,OEHLBERG L.Mobile learning with the engineering pathway digital library[J].International Journal of Engineering Education,2012,28(5):1119-1126.

[6] 侯桂楠.基于用户体验的移动图书馆服务模型研究[D].重庆:重庆大学,2013.

[7] 倪峰,李永明,郑德俊,等.移动图书馆服务平台的改进需求识别[J].图书情报工作,2016,60(2):17-23.

[8] 邹腊梅,肖基毅,龚向坚. Web文本挖掘技术研究[J].情报杂志,2007(2):53-55.

[9] 倪瑜泽,彭蓉,孙栋,等.基于用户评论的潜在演化需求发现方法[J].武汉大学学报(理学版),2015,61(4):347-355.

[10] 崔建苓,杨达,李娟.RERM:一种基于评论挖掘的需求获取方法[J].计算机应用与软件,2015,32(8):28-33.

[11] 龚才春.短文本语言计算的关键技术研究[D].北京:中国科学院研究生院(计算技术研究所),2008.

[12] 王仲远,程健鹏,王海勋,等.短文本理解研究[J].计算机研究与发展,2016,53(2):262-269.

[13] 李亚松.基于文本挖掘的用户评论分类解析系统的设计与实现[D].北京:北京邮电大学,2015.

[14] FLURY B.Algorithms for clustering data:Anil K. Jain and Richard C.Dubes Prentice Hall Advanced Reference Series in Computer Science Prentice Hall,Englewood Cliffs,NJ,1988[J].Journal of Statistical Planning & Inference,1989,21(1):137-138.

[15] 杨翔.针对短文本数据的聚类分析的算法及应用设计和实现[D].北京:北京邮电大学,2014.

[16] 李伟,黄颖.文本聚类算法的比较[J].科技情报开发与经济,2006,16(22):234-236.

[17] MACQUEEN J.Some methods for classification and analysis of multivariate observe[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press,1967:281-297.

[18] 朱建宇.K均值算法研究及其应用[D].大连:大连理工大学,2013.

[19] 左进,陈泽茂.基于改进K均值聚类的异常检测算法[J].计算机科学,2016,43(8):258-261.

[20] TZORTZIS G,LIKAS A.The minmax K-means clustering algorithm[J]. Pattern Recognition,2011,44(4):866-876.

[21] 张志祥.基于最大最小距离法的多中心聚类算法[J].计算机应用,2006,26(6):1425-1428.

[22] 傅德胜,周辰.基于密度的改进K均值算法及实现[J].计算机应用,2011,31(2):432-434.

[23] BRADLEY P S,FAYYAD U M.Refining initial points for K-means clustering[C]//Proceedings of the 15th International Conference on Machine Learning.[S.l.]:[s.n.],1998:91-99.

[24] 朱晓峰,陈楚楚,尹婵娟.基于微博舆情监测的K-Means算法改进研究[J].情报理论与实践,2014,37(1):136-140.

[25] CHOWDHURY G G.Natural language processing[J].Annual Review of Information Science & Technology,2003,37(37):51-89.

[26] 郑德俊,轩双霞,沈军威.用户感知的移动图书馆服务质量测评模型构建[J].大学图书馆学报,2015,33(5):83-92.

[27] 刘远超,王晓龙,徐志明,等.文档聚类综述[J].中文信息学报,2006,20(3):55-62.

[28] 裴超,肖诗斌,江敏.基于改进的LDA主题模型的微博用户聚类研究[J].情报理论与实践,2016,39(3):135-139.

作者简介

郑德俊，男，1968年生，教授，研究方向：信息服务与评价，E-mail：zdejun@njau.edu.cn。
朱婷婷，女，1993年生，硕士研究生，研究方向：移动图书馆服务，E-mail：2015114009@njau.edu.cn。
沈军威，男，1989年生，博士，研究方向：移动图书馆服务，E-mail：t2017013@njau.edu.cn。

Research on Demand Clustering of Mobile Library from User Reviews Based on the Improved K-means Algorithm

ZHENG DeJun, ZHU TingTing, SHEN JunWei
(Department of Information Management, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: The automatic clustering of mobile library user reviews helps to obtain user needs more accurately and efficiently. Based on the traditional K-means algorithm, this paper uses HT-LaD algorithm to improve the initial clustering center and uses the user's evaluation data of mobile library to prove it. The results show that it is feasible to use the improved K-means algorithm to complete the demand clustering of mobile library user comment text, and the clustering accuracy and stability are improved.

Keywords: Mobile Library; Improved K-means Algorithm; User Reviews; User Demands

（收稿日期：2017-09-01）