软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1174−1187 [doi: 10.13328/j.cnki.jos.004958]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于移动终端日志数据的人群特征可视化*

张宏鑫[1], 盛风帆[2], 徐沛原[1], 汤 颖[2]

[1](CAD & CG 国家重点实验室(浙江大学),浙江 杭州 310058)

[2](浙江工业大学 计算机科学与技术学院,浙江 杭州 310023)

通讯作者: 汤颖, E-mail: ytang@zjut.edu.cn

摘 要: 随着我国移动互联网的迅猛发展,如何从海量移动终端日志数据中提取出有效信息,并进行合理、清晰的可视化分析,为工业界等提供有价值的统计分析功能显得尤为重要.目前,对于移动终端日志数据的研究和分析多是基于对单一属性的统计结果分析,如应用下载排行、用户留存率等.为了进一步挖掘移动终端日志数据背后深层次的隐含信息,更加准确地概括出移动终端用户的特征,提出了一种基于移动应用程序日志数据的人群特征分析与画像计算方法,构造了基于移动应用程序数据的主题模型,并将移动设备用户按照与不同应用主题的相关度进行聚类,得到了具有不同特征的人群,从而提出了基于层次气泡图和 Voronoi Treemap 的可视化展现与分析方案.进一步将人群特征与时间信息、地理位置信息相结合,从多角度可视化展现人群特征.最后,根据该研究内容,实现了 B/S 架构的日志数据可视化分析原型系统,并通过案例分析验证了该方法的有效性.

关键词: 数据可视化;主题模型;移动设备用户特征

中图法分类号: TP391

中文引用格式: 张宏鑫,盛风帆,徐沛原,汤颖.基于移动终端日志数据的人群特征可视化.软件学报,2016,27(5):1174−1187. http://www.jos.org.cn/1000-9825/4958.htm

英文引用格式: Zhang HX, Sheng FF, Xu PY, Tang Y. Visualizing user characteristics based on mobile device log data. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1174−1187 (in Chinese). http://www.jos.org.cn/1000-9825/4958.htm

Visualizing User Characteristics Based on Mobile Device Log Data

ZHANG Hong-Xin[1], SHENG Feng-Fan[2], XU Pei-Yuan[1], TANG Ying[2]

[1](State Key Laboratory of CAD & CG (Zhejiang University), Hangzhou 310058, China)

[2](School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: With the dramatic countrywide development of mobile internet, it becomes very important to extract valuable information from mobile device log data and report the analysis result through visualization method to help application developers and distributors maximize monetization opportunity. Currently, most of mobile log data analysis work is based on single dimension statistics, e.g., app download rank, and user retention rates. In order to mine deep information hiding behind mobile device log data and summarizes user characteristics. A method is proposed for analyzing users' characteristics and computing users' profile. An app topic model is constructed based on mobile log data, user clusters are build according to app topics, and two visualization methods are designed to show user characteristics clusters. Furthermore, user clusters are combined with time information and geographical information to show user characteristics from additional dimensions. Finally, a mobile log data visualization analysis B/S system is implemented to demonstrate the validity of the method by a case study.

Key words: data visualization; topic model; mobile device user characteristics

* 基金项目: 国家自然科学基金(61232011); 浙江省自然科学基金(LZ12F02002, LY14F020021); 国家科技支撑计划(2014BAH23F03)

Foundation item: National Natural Science Foundation of China (61232011); Natural Science Foundation of Zhejiang Province of China (LZ12F02002, LY14F020021); National Key Technology R&D Program of China (2014BAH23F03)

收稿时间: 2015-07-31; 修改时间: 2015-09-19; 采用时间: 2015-11-10

随着信息技术的不断发展,进入 21 世纪以来,智能手机产业飞速发展.智能手机可以让用户根据自己的需求和喜好安装各种功能的应用软件、各种类型的游戏,这是它吸引用户的主要特色之一.如何从含有众多应用程序的应用市场中为移动用户推荐他们确实需要的应用程序?如何将一款手机游戏推荐给喜欢该类型游戏的用户;另外,值得注意的是,用户在不断下载他们感兴趣的应用的同时也会卸载不再需要的应用,如何成功留住用户是一个重要问题.以上这些问题成为应用商店平台提供商、移动应用程序开发者都希望得到解决的问题.

通过移动终端日志信息得到的海量用户信息和应用程序数据,为商业研究和分析提供了宝贵的数据资源.目前,很多厂商都已经开始利用日志数据进行研究和分析.但是,目前的数据研究和分析大多基于对单一属性的统计结果分析,如应用下载排行、用户留存率等.如果希望得到人群的行为习惯、更为精确的用户特征,往往需要综合多维度的数据分析.但将多维度的数据同时进行清晰的呈现,对于可视化展示而言是很困难的.

主题模型(topic model)[1]属于概率产生式模型(generative model),是一种层次贝叶斯模型,可以以无监督的方式自动组织和理解文档,发掘一系列文档中抽象的主题,在自然语言处理、机器学习等领域都有广泛的应用.为了更加直观地展示主题模型得出的结果,帮助人们理解,主题模型越来越多地与可视化方法相结合,主题模型为数据可视化提供模型基础,可视化将主题模型结果直观地、可交互地进行展现.

本文将 LDA(latent Dirichlet allocation)主题模型引入到手机日志数据的分析中,提取出手机应用分类主题.并将手机用户按照与不同应用主题的相关度进行聚类,形成具有代表性的人群.将人群聚类结果与时间维度相结合,用于观察变化趋势.此外,还将人群聚类信息与地理位置信息相结合,从而进一步了解手机用户更为详细的信息,如分布情况.为了能够将结果以直观、易懂的方式展现给相关厂商和研究人员,帮助他们更加有效地对手机用户进行研究和分析,本文分别采用了层次气泡图、像素地域分布图等可视化展现方法.本文第 1 节讨论相关工作.第 2 节详细介绍人群特征的可视化研究的流程和方法以及具体的实现步骤.第 3 节介绍人群聚类信息结合空间维度信息的可视化.第 4 节进行总结,并对未来的研究方向提出设想.

1 相关工作

1.1 主题模型

最早的文本数据挖掘方法是基于向量空间模型(vector space model,简称 VSM)[2].随后,Landauer 等人提出了潜在语义分析模型(latent semantic analysis,简称 LSA)[3].LSA 通过线性代数中的奇异值分解(singular value decomposition,简称 SVD[4])方法来对单词-文档矩阵进行维数约减,从而将单词-文档映射到一个低维的潜在语义空间中[5].Hofmann 等人于 1999 年提出了概率潜在语义分析模型(probabilistic latent semantic analysis,简称 PLSA)[6],Blei 等人于 2003 年提出了潜在狄利克雷分配模型(即 LDA 模型)[7].

传统判断两篇文档相似性的方法是,比较两篇文档共同包含的单词的多少,如 TF-IDF(term frequency-inverse document frequency)方法等[8,9].LDA 模型则假设一篇文档是由主题集合中的各个主题按照一定的比例构成的,而每一个主题又是由单词表中的单词按照一定的比例混合而成的.通过机器学习的方法可以得到文档的主题,从而判断两个文档是否相似.LDA 模型层次清晰,依次分为文档层、主题层和单词层.其中,文档和主题相关联,主题和单词相关联.可以通过学习文档集中的单词挖掘出所有潜在的主题信息,并通过这些信息来挖掘该文档集以外的其他文档的主题分布.

1.2 数据可视化

现代的可视化旨在研究大规模信息资源的视觉呈现[10],以及利用图形和图像的相关技术和方法将数据直观显示,为用户提供可交互操作等,帮助人们理解和分析数据[11].如今,可视化技术已成为一个基本的工具,用来揭示数据集中数据之间的关系和背后隐匿的信息[12].基于不同的显示需求、交互需求等,可视化的方法也是多种多样的.Treemap(矩形式树状结构绘图法)是一种在受限空间内展示树状数据结构的可视化方法[13],通过将矩形不断进行细分(slice and dice),可以在固定大小区域内展示多层级的数据信息,也可以比较直观地展示同层级数据之间的比较,但很容易在结果中出现细长的矩形,不利于辨别.为了解决这一问题,提出了 Voronoi Treemap

(泰森多边形树状结构图)[14]的方法,可以避免出现细长矩形的情况,达到更好的可视化效果.而且最外层的区域也不再限制为矩形,可以在任意形状内进行多层级数据的展示.马赛克图(Mosaic display)是一种用来展示关联表(contingency table)的图解法[15].马赛克图与 Treemap 的区别是:每一次将一个矩形切分成几个矩形,都等价于增加一个维度的信息.一般用于二维、三维、四维的低维数据的可视化展示.本文采用了嵌套圆圈的形式来展示结果,并将这种展示形式命名为气泡图[16].用一个圆圈将构成信息包含起来,符合集合的表示形式,也能够展示数据结果的层次关系,更为用户提供了方便的交互.与传统的、一旦生成就固定不变的二维表格表现方式相比更加灵活多变,通过缩放操作可以为用户清晰地展示用户关心的数据细节,也可以进行整体上的宏观比较.

2 基于 LDA 主题模型的人群聚类

本文的研究课题是基于手机日志数据的人群特征可视化,尝试挖掘出手机日志数据背后隐含的数据信息,更加深入地了解手机用户的人群特征.如图 1 所示,根据这一方向和目标,我们确定了研究的基本流程,主要分为 4 个步骤:(1) 获取研究所需的相关日志数据;(2) 对这些数据文件进行筛选,选取并整理出有效的数据;(3) 构造可视化系统的原始模型,探索对日志数据的可视化方法和工作流程,并最终对这些数据进行可视化转换;(4) 根据可视化展现结果,结合实际情况进行观察和对比分析,挖掘出更有效的信息,并不断调整改善模型.

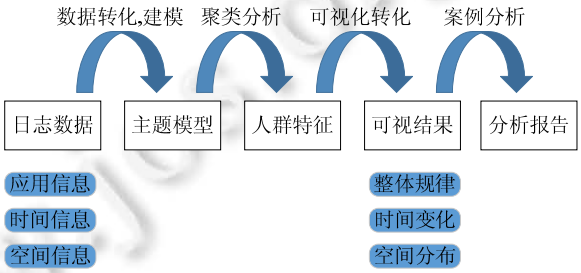


Fig.1 Overview of the method

图 1 本文方法概览

2.1 问题分析

为了挖掘出手机用户的用户特征及行为习惯,本节研究内容所需要的数据主要有两类:一类是通过日志采集到的每天的手机用户安装的应用程序列表信息,包括用户 ID、安装的应用程序 ID 列表、卸载的应用程序 ID 列表以及年、月、日等时间信息;另一类是从各大应用市场得到的应用程序所属的分类标签信息,包括应用程序 ID 和分类标签名称列表.

传统的基于手机日志数据的可视化大多是以二维表格的形式展现统计结果,而我们希望可以挖掘出一些潜在的更为有价值的信息,因为经过对真实用户数据的观察以及用户调研,我们发现,手机用户安装的应用程序往往是可以进行归类的,如微博、微信、QQ、人人网、开心网等应用都是属于社交类应用程序.我们把这样的一个集合定义为一个主题“社交类主题”,我们可以发现,有些用户对某一主题类别的应用程序安装得特别多,这是很有实际价值的信息.另一方面,很多人都不会只安装某一主题类别的应用程序,如果能够挖掘两类不同主题类别应用程序安装数量的比例关系,也是很有价值的信息.例如,“安装游戏主题类别应用程序多的人往往也会安装手机安全手机清理主题类别的应用程序”,结合实际使用场景,玩游戏多的人往往对手机性能要求比较高,所以会通过手机清理类应用程序及时清理手机上的“垃圾内存”.

因为每天都会产生大量的手机日志数据,所以,如何从庞大的日志数据中筛选出有用的数据信息进行建模分析显得十分重要.主题模型需要做的数据预处理工作主要有以下几点:

(1) 系统会每天采集很多次日志数据,但同一天同一部手机上安装的应用程序相关信息往往是一样的,所以需要做去除掉重复的记录.

(2) 对于用于收集数据的系统而言,每天都会有新激活的设备,也会存在不再激活、销毁的设备.对于主题模型而言,由于通过 LDA 学习出来的主题模型不受总的手机设备的影响,所以我们对于每个月的数据,可以固定地只观察一部分确定的手机设备的日志数据.

(3) 每天系统获取到的手机日志数据包含该手机今日新安装的应用程序信息及卸载的应用程序信息,但不包括该手机当前状态下的所有应用程序信息.所以需要额外维护一张数据表,用于存储我们观察的手机设备的历史累积的安装的应用程序列表,每天根据日志得到的应用程序变化数据以及前一天的应用程序状态数据,计算得到当天的应用程序数据信息.

2.2 主题模型的建立

在传统的手机日志数据可视化中,关于手机应用程序信息的展示往往是对单一信息进行简单的统计结果后的展示,如应用程序装机量的排名、某一应用的用户留存率统计等.另外一种对手机应用程序信息的可视化展示是统计整个手机应用市场中按类别划分后的应用程序的安装数量,最后得到的是手机市场中不同类别应用程序的安装数量排名.这样的统计数据具有一定的价值,可以知道,目前应用市场中哪类应用程序安装量最大,但也存在如下不足之处:

(1) 每个手机用户都不太可能只安装一个类型的应用程序,都会安装多个类型的应用程序,所以只统计得到的某一分类的用户数量,却无法得知该用户安装其他分类应用程序的多少.所以,这种方法是不能代表具有某类特征的人群的,一类人群的手机应用程序往往是由不同分类的应用程序按照一定比例混合组成的.

(2) 按照类别进行划分得也不够细化、精确.因为应用市场给一个应用程序划分的大类比较概括,不够细化,如,游戏类下面又可以具体分为塔防类游戏、跑酷类游戏、射击类游戏、解谜类游戏等,工具类下面又可以具体分为输入法、浏览器、词典等.每一个具体的小分类下面,会包括各种名称的应用程序.所以,按照这些更为细化的标签进行分类,可以得到更为准确的结果.

(3) 一个应用程序可能会有多个标签,所以只归纳为一个大类进行统计不够准确.不同的应用程序也可以组合,形成新的主题分类.因为应用市场中会不断有新的应用程序产生,也可能会产生新的应用程序类别,所以主题分类需要可以据此动态改变.

我们则将 LDA 主题模型巧妙地引入到手机日志数据可视化与分析中来.定义手机应用程序所包含的多个分类标签对应单词,一部手机即对应一篇文档,采集到的所有手机用户即为语料库,通过提取应用主题来分析手机用户.

通过对用户的手机上所安装的应用程序标签数据进行分析,我们可以得到该手机用户潜在的主题信息.基于此,我们可以判断两名手机用户是否属于同一类型的用户,我们可以将类似的用户聚集在一起形成人群.在对所有手机用户进行聚类之前,LDA 主题模型很好地帮助我们对应用数据特征进行了降维操作,将原本每个手机上数十个千差万别的应用名称提取为该手机用户与几个(本文为 5 个)主题的相关度.使用这 5 个主题,按照不同比例组合,就可以代表一个手机用户的特征,从而可以很容易地对所有用户进行聚类操作,解决了直接基于手机安装的应用程序名称进行聚类,由于不同用户之间安装的应用程序差异很大、特征点太多,无法进行有效聚类的问题.

2.2.1 建立应用标签的词袋模型

首先,我们需要将手机上的应用程序数据与 LDA 模型的输入数据文档中的单词建立对应关系.将应用名称直接作为单词,会导致整个语料库单词太多、词频太小,无法有效学习出整个语料库中的文档主题信息.而将应用程序名称对应为预先维护好的几个分类标签后,可以达到名称的规范统一,保证语料库的单词量适中,词频适中,还可以增加代表该手机用户的标签数量.所以,我们使用应用程序的分类标签信息作为单词,建立了每台手机的词袋模型,见表 1.

Table 1 Bag of words model for application tags

表 1 应用标签词袋模型

应用名称	分类标签	词袋
APP Name1	Label 1 Label 2	Label 1×1 Label 2×2 Label 3×1 Label 4×1
APP Name2	Label 2 Label 3	
APP Name3	Label 4	

为了得到更好的分析结果,使主题模型更加准确地代表手机用户,我们尝试对采集到的每个手机上安装的应用程序数据增加对应的打分机制.因为手机用户安装一个应用程序后,可能从未使用过该应用程序,可能使用过很多次该应用程序,可能过了一两天后将该应用程序卸载,可能该应用程序自安装后很多天都没有卸载,所以我们设计了如下的打分机制:对于手机上安装的每个应用程序,如果该应用安装当天就卸载,则认为是不得分的;该应用每在该手机上留存 1 天,对应的分类标签便增加 1 分;留存大于等于 10 天以上的应用,我们认为该应用一直在该手机上,对应的分类标签得 10 分.手机用户每实际启动该应用程序一次,便为该应用程序对应的分类标签增加 1 分.

增加打分机制后,得分越高的应用标签,我们认为该标签类别的应用在该手机上使用得越多,与该手机用户特征关系越大.与 LDA 模型中的词频相对应,一篇文档中出现次数越多的单词越能代表该篇文档.

当然,和 LDA 处理文档时会剔除像“and”这种在每篇文档中都会出现多次的无意义单词一样,我们也会剔除“免费软件”这种会属于很多应用程序、在每个手机上都会出现多次的、没有实际意义的标签.而且我们采集的日志数据,统计的是手机用户自己安装的应用程序信息,不包括原生 ROM 自带的如短信、电话等,这些每部手机买来就已经安装好的应用程序.

通过对分类标签增加打分机制,我们得到了能够更加准确代表手机用户应用数据的词袋模型.

2.2.2 手机主题特征模型

将一部手机看作一篇文档,该手机的所有应用程序对应的分类标签作为单词,根据 LDA 主题模型,我们可以得到公式(1):

$$P(\text{标签}|\text{手机})=P(\text{标签}|\text{主题})\times P(\text{主题}|\text{手机}) \tag{1}$$

更具体地,每一部手机与 T 个(通过反复实验等方法事先确定)主题的一个多项分布相对应,将这个多项分布记为 θ.每个主题又与分类标签库中的 V 个标签的一个多项分布相对应,将这个多项分布记为 Φ.θ和 Φ分别有一个带有超参数 α和 β的狄利克雷先验分布[17].对于一部手机 d 中的每一个标签,我们先从该手机所对应的多项分布 θ中选择一个主题 z,然后,我们再从主题 z 所对应的多项分布 Φ中选择一个标签 w.将这个过程重复 N_d 次就产生了手机 d,其中,N_d 是手机 d 中的总标签数.这个生成过程可以用如图 2 所示的盘子表示法(plate notation)表示.

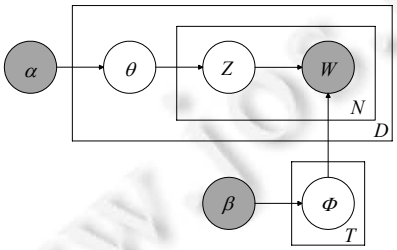


Fig.2 Plate notation of the topic characteristics model[7]

图 2 主题特征模型的盘子表示法[7]

盘子表示法图中的阴影圆圈代表可观测变量(observed variable),非阴影圆圈代表潜在变量(latent variable),箭头表示变量之间的条件依赖性(conditional dependency),方框表示重复抽样.如果给定了 α和 β,那么文档的主

题分布θ、主题向量$z=(z_1,\ldots,z_n)$以及单词向量$w=(w_1,\ldots,w_n)$的联合分布如公式(2):

$$P(\theta,z,w\mid\alpha,\beta)=P(\theta\mid\alpha)\prod_{n=1}^{N}P(z_n\mid\theta)P(w_n\mid z_n,\beta) \tag{2}$$

其中,$p(z_n|\theta)$实际上就是对应z_n=i 的θ_i分量.上式对θ和z在全部取值区间内积分(或累加),以消去θ和z,便得到了一篇文档中单词的边缘分布如公式(3):

$$P(w\mid\alpha,\beta)=\int P(\theta\mid\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}P(z_n\mid\theta)P(w_n\mid z_n,\beta)\right)\mathrm{d}\theta \tag{3}$$

对于含有M篇文档的文档集:

$$P(D\mid\alpha,\beta)=\sum_{d=1}^{M}P(w_d\mid\alpha,\beta) \tag{4}$$

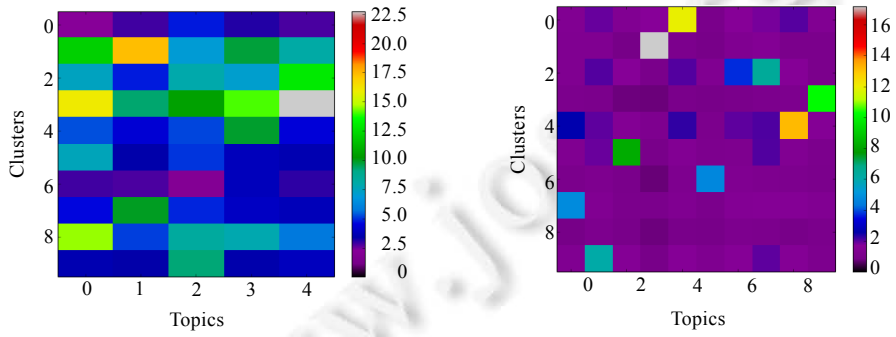
$$P(D\mid\alpha,\beta)=\prod_{d=1}^{M}\int P(\theta_d\mid\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_n}p(z_{dn}\mid\theta_d)P(w_{dn}\mid z_{dn},\beta)\right)\mathrm{d}\theta_d \tag{5}$$

LDA 的训练过程,就是估算使公式中$P(D|\alpha,\beta)$取得最大值的参数α和β;LDA 的预测过程,则是通过已知的α和β预测文档的主题分布θ以及主题和单词的分布Φ.该模型有两个参数要推断:一个是“文档-主题”分布θ,另一个是“主题-单词”分布Φ.推断方法主要有 LDA 模型的作者 Blei 博士等人提出的变分推断算法(variational inference)[18]、最大期望算法(expectation maximization),还有现在常用的 Gibbs 抽样法[19].成功解出θ和Φ后,得到了表示手机在主题上的分布和主题在标签上的分布.

我们经过多次实验,认为提取出 5 个主题更具有代表性和实际意义.这 5 个主题分别为工具类、娱乐类、生活类、游戏类与社交类,每个主题由若干个分类标签组成.《互联网周刊》对外正式发布了《2014 年中国 APP 排行榜 TOP 500》榜单,抛弃了过往唯“下载量”论的评选方式,评选中不仅衡量应用在用户中的受欢迎程度,更综合考量了应用本身的创新性、实用性以及对未来应用发展的引领作用.数据显示,App 类型基本上被社交、游戏、生活、娱乐、工具类应用所占据,说明我们所提取的主题与实际调查是一致的.

2.2.3 手机特征聚类

通过上述 LDA 模型学习,我们同时得到了每个手机与 5 个主题的相关度.此时,用 5 个主题的相关度就可以代表该手机用户的特征.如某手机用户和“游戏”、“社交通信”、“手机工具”、“多媒体”、“生活服务”5 个主题的相关度依次为 0.8,0.3,0.3,0.2,0.1,我们可看出,该手机用户属于比较典型的游戏玩家人群.将所有手机用户都以主题 5 维向量来表示后,我们就可以用 *K*-Means 聚类算法[20]将具有相同主题特征的用户聚集,形成具有特征的代表性人群.基于 MDL 标准[21]以及多次实验,我们选择聚类成 5 类人群更具有代表性且易于观察.我们用该人群分别和 5 类主题的相关度大小来代表该人群的特征.如图 3(a)所示,横坐标代表 5 类主题,纵坐标代表聚成的 5 类,每个色块代表该类人群和对应主题之间相关度,色调越暖,相关度就越大.



(a) 使用分类标签 (b) 使用应用名称

Fig.3 Clustering results

图 3 聚类结果

图 3 是我们采用不同的信息作为“单词”的实验结果.图 3(a)是我们采用应用程序的分类标签信息作为单词,建立词袋模型,参与 LDA 计算;图 3(b)所示为我们直接使用应用程序的名称作为单词以建立词袋模型,用于

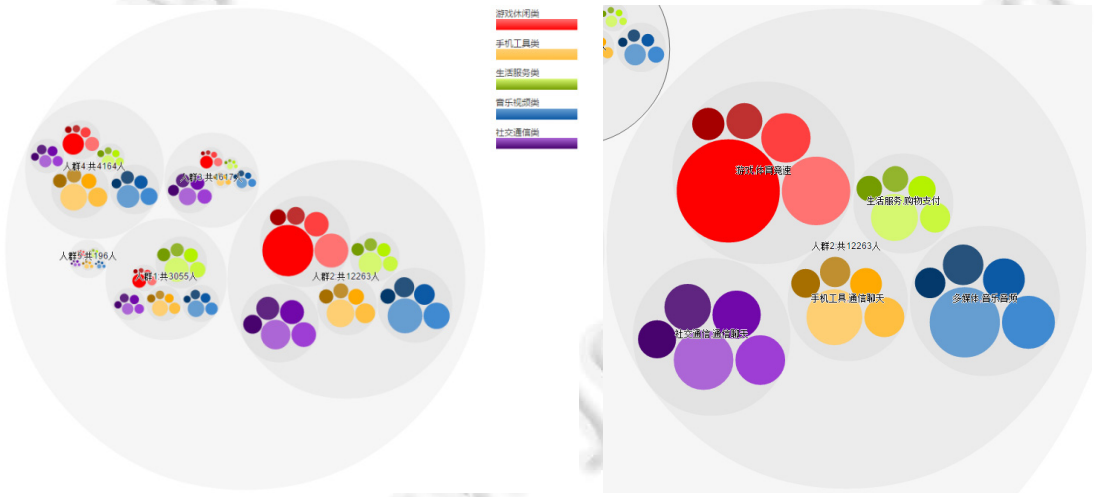
LDA 计算.可以看到:使用应用程序名称作为单词,由于词库太大、词频太小,无法有效提取出主题模型,导致每个聚类后的簇(cluster)和每个主题的相关度都差不多,没有区分度.而使用应用标签信息作为单词,最后的效果比较理想,每个类簇都有各自的特征,之间存在较明显的差异.

2.3 可视化转换

对海量的手机日志数据进行有针对性的筛选和处理后,通过网页的形式将结果直接展示,提供给用户自己去挖掘数据之中的价值,不应该仅仅是一副由程序计算后生成的图片或表格,而应该是一个可以进行交互的应用,使得用户可以方便地进行操作.用户根据自己的视角来获取感兴趣的内容,并可以通过交互的方式逐步缩小兴趣点的范围.当用户通过筛选确定自己感兴趣的人群或主题后,可视化系统可以将这部分人群的应用程序数据进行详细的展示,并提供关键字段的数据导出功能,供用户更近一步地进行深度分析和使用.基于上述对本次可视化研究的意义和目的的探讨以及对用户数据的分析,逐渐探索出了一个基于手机日志数据对手机用户人群特征的可视化流程和方法,这也是本文可视化研究的核心.

2.3.1 层次气泡图

通过 LDA 主题模型计算得到的应用程序主题信息以及手机用户和主题的相关度、主题和分类标签的相关度后,我们选择使用嵌套的圆圈,如图 4(a)所示,即气泡图来进行展现计算结果.层次气泡图最外层圆圈代表人群,中间一层 5 个圆圈代表该人群与 5 个应用主题的相关度比例关系,最里面一层的 5 个小圆圈代表最能表示该主题的 5 个分类标签,同样用面积去编码其相关度.用一个圆圈来将其构成信息包含起来,符合认为概念中集合的表示形式.通过 LDA 主题模型计算得到的结果,本身具有嵌套关系,用户人群由主题组成,主题又由分类标签组成,所以我们将圆圈也进行嵌套,这样更符合数据结果的层次关系.气泡图更为用户提供了方便的交互.用户选择了感兴趣的人群后,可以点击代表该人群的圆圈,人群圆圈将放大;用户可以看到该人群的具体特征,主题信息的构成比例,如图 4(b)所示;当用户选择了进一步想了解的主题的圆圈后,该主题圆圈将放大,用户可以观察到该主题里包含的具体应用程序分类标签及其对应的比例.这种层层递进的表现形式,可以使用户先对所有人群有总体的概览,进行不同人群的比较;选择了想进一步了解的人群后,可以通过放大的交互操作,得到更为详细的信息展示.



(a) 气泡图全貌 (b) 人群 2 详细特征

Fig.4 Hierarchical bubble chart

图 4 层次气泡图

我们通过 *pack*()函数将后台 LDA 主题模型计算得到的 json 格式的结果数据,在指定的网页空间范围(长和宽)内,先后计算代表人群、主题、标签的大圆、中圆、小圆的半径,圆的位置都尽可能地相切.圆的面积大小分

别与人群大小、每个人群和 5 个主题的相关度大小、每个主题及其内的 5 个分类标签的相关度大小相关.5 个主题选取了 5 个具有明显差异的色系来代表,每个主题内的分类标签采用同一个色系的颜色,并按照固定的差值进行深浅变化.

2.3.2 基于时间维度的动态比较

气泡图将 LDA 主题模型计算得到的人群聚类结果以嵌套圆圈的形式为用户进行展现,用户可以选取日期来查看该日的人群聚类信息.如果增加时间维度,就可以帮助用户察看一段日期范围内的人群变化趋势.

如果只是将不同天的数据进行切换,会显得很突兀,不自然,所以我们采用动画将每天的可视化结果串联起来.先使用 pack 函数计算得到后一天日期的气泡图中圆圈的半径及位置坐标,利用 *transition* 函数,可以对网页上的每个圆圈元素按照指定的方向及速度进行变换,实现串联动画.

然而,如果只是通过平移动画将前后两天的气泡图串联起来,那么,由于在通过 *pack* 函数计算每天的圆圈半径和位置坐标时,为了更有效地使用空间,会优先考虑尽可能地将圆圈之间形成相切关系,所以前后两天的代表人群的大圆圈的位置可能会出现较大变动,并且动画变换过程中会出现交叉,如图 5 中矩形框框出部分所示.这样产生的动画会变化较大,在观察连续的多天数据时变化剧烈,不方便观察人群的变化趋势.

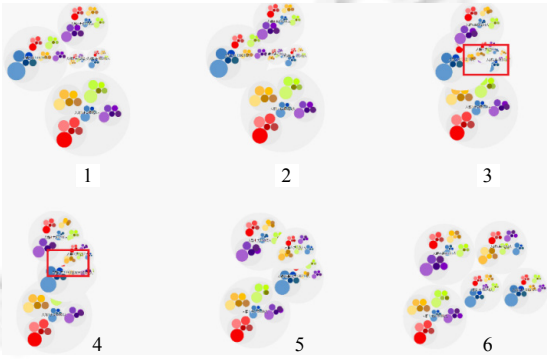


Fig.5 Bubble chart's timing transformation result (with overlap)

图 5 气泡图时序变换结果(交叉重叠)

因此,我们需要将代表 5 类人群的大圆圈的相对位置尽可能地固定下来,于是,我们对 pack 函数进行了改进,在计算大圆圈的位置坐标时优先依据人群编号进行排序操作,将人群 1~人群 5 从 9 点钟位置依次按照顺时针顺序进行排列.如图 6 所示,当代表人群的大圆圈位置相对固定后,一段日期范围内的动画也达到了平滑过渡的效果,使得用户观察变化趋势更为直观了.

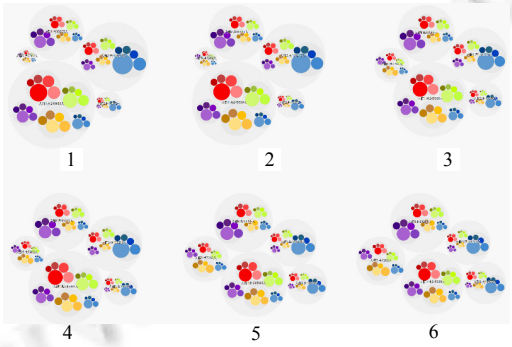


Fig.6 Bubble chart's timing transformation result (translate smoothly)

图 6 气泡图时序变换结果(平滑过渡)

2.3.3 Voronoi Treemap

气泡图可以层次分明、清晰地将人群特征和主题相关度信息进行可视化展示,但占用的面积较大.如果想看到深层次的细节信息,则需要通过点击放大后察看,适合在空间充足的网页中进行可视化展示.但在手机移动端,如手机浏览器、微信中查看,或者在网页中显示缩略图时,就会显出不足之处.

为了满足在较小的空间内也可以让用户查看到整体的人群聚类结果,我们采用泰森多边形树图(voronoi treemap)[22]来展现数据.Voronoi Treemap 中的基本图形单位是任意形状的小多边形,用面积代表这个小多边形数据的大小,由于每个小多边形可以是任意形状的,所以可以充分利用空间,不留间隙.每块区域都是不规则的多边形,再配以辅助颜色后,可以比较清楚地区分层级关系.最终得到的 Voronoi Treemap 形式展示的可视化结果如图 7 所示.

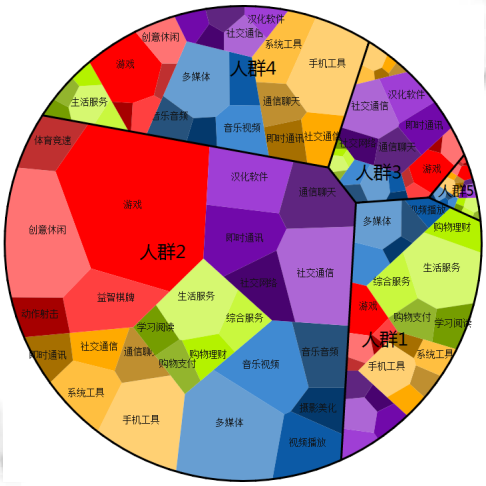


Fig.7 Visualization of user clustering results using voronoi Treemap

图 7 人群聚类结果的 voronoi 树图可视化

Voronoi Treemap 可以更加充分地利用空间,同时在一层画布中展示出所有层级的分布信息.但是由于每次 Voronoi Treemap 的生成过程都需要多次迭代后才可以求出“稳定”的分布结果,所以不适合串联起来展示一段日期范围内的多天数据.可以想象:图像会处于一直的运动状态,会让用户无法分辨每日的数据,也无法观察到变化趋势.而且在空间允许的条件下,气泡图可以更加清晰地展示数据,尤其是层级信息,圆圈也更符合人们对于集合的一般表现形式.Voronoi Treemap 更适合在手机等小屏幕设备上总体地展现人群聚类可视化效果.

3 可视化实例分析

3.1 总体规律

通过 LDA 学习后得到了 5 个主题(topic),与每个主题相关度最大的 5 个分类标签见表 2.

Table 2 Topic constitution

表 2 主题构成

主题 1	主题 2	主题 3	主题 4	主题 5
手机工具	多媒体	生活服务	游戏	社交通信
系统工具	音乐视频	综合服务	创意休闲	汉化软件
社交通信	视频播放	购物理财	益智棋牌	即时通信
通信聊天	音乐音频	购物支付	体育竞速	通信聊天
即时通信	摄影美化	学习阅读	动作射击	社交网络

我们的实验共采集了 2 万多名手机用户的数据.每个用户都和 5 个主题有相关度的值,也就是说,每个用户都有一个 5 维向量,我们使用 *K*-Means,按照用户相互之间的相似性(也就是向量之间的距离),把用户聚类成 5 类人群,对于每一类里面的用户,求这一类里所有用户的向量的平均值,得到这一类的中心点.5 类用户人群的中心点见表 3.

Table 3 Characteristics analysis of different category users

表 3 不同人群的特征分析

	人数	主题 1	主题 2	主题 3	主题 4	主题 5	特征分析
人群 1	3055	8.259	7.223	23.154	3.623	5.231	尤其喜欢生活服务类
人群 2	12263	3.282	4.168	2.799	6.404	4.509	各类主题安装数量相差不大,游戏类稍多
人群 3	1617	5.342	7.904	3.524	8.106	31.958	尤其喜欢社交聊天类
人群 4	4164	16.511	9.526	4.497	6.018	4.289	手机工具系统工具类应用相对多于其他类
人群 5	196	60.137	42.081	79.376	33.62	34.397	对各类应用都很感兴趣,生活服务和系统工具类应用安装最多

从表 3 的人群聚类结果中我们可以发现,人群用户数量最多的是人群 2.该人群的特征为:各类主题的应用程序都安装一些,各类之间差别不大,游戏类应用稍微多一些.这和我们实际调查得到的数据相一致,人群 2 确实可以代表大多数手机用户的特征.各个人群可视化结果如图 7 所示.

3.2 同一人群的用户数据分析

我们从 2 万多名用户数据中随机选取了 5 名用户,他们的人群聚类结果以及 5 个主题类别的相关程度见表 4.

Table 4 Relevancy No.1 between sampling users and topics

表 4 抽样用户和主题的相关度 1

	主题 1	主题 2	主题 3	主题 4	主题 5	聚类结果
用户 1	6.978	0.527	0.557	4.577	0.911	人群 2
用户 2	3.318	0.532	0.559	14.663	1.518	人群 2
用户 3	0.521	0.518	16.506	0.518	1.526	人群 1
用户 4	5.470	0.528	2.668	4.527	20.394	人群 3
用户 5	0.762	10.070	0.833	4.489	8.434	人群 2

表 4 中,用户 1、用户 2、用户 5 实际手机上安装的应用信息见表 5.

Table 5 Installed application information No.1 of the sampling users

表 5 抽样用户安装应用程序信息 1

	应用标签(数量)
用户 1	输入法(1)浏览器(1)即时通信(1) 视频播放(1) 新闻阅读(1) 游戏(1)竞技飞行(1)动作射击(1) 社交网络(1)
用户 2	系统工具(1)安全杀毒(1)即时通信(1)社交通信(1) 视频播放(2) 教育学习(1)图书阅读(1) 游戏(4)飞行射击(1)动作射击(1)
用户 5	音乐视频(3)图书动漫(2) 办公学习(1)学习阅读(1) 体育竞速(1)策略经营(1) 汉化工具(1)社交通信(1)社交网络(2)

用户 1、用户 2、用户 5 实际手机上安装的应用程序信息对应的分布直方图如图 8 所示.

从图 8 中可以看出:用户 1 安装的应用数量不多,基本上每个主题类别应用都安装了,主题 1(工具类)和主题 4(游戏类)稍微感兴趣一些;用户 2 和用户 1 类似,对主题 4(游戏类)相对更感兴趣;用户 5 安装应用数量较多,但并没有偏爱某一主题,每类主题应用安装的数量差不多.

图 9 所示为人群 2 的特征,人群 2 各类主题的安装数量比较均衡,游戏类稍多,其次是多媒体类与生活服务类.从图 7 所示的不同人群特征可视化结果可见,人群 1、人群 3~人群 5 都对某一主题应用有一定的偏向性.用户 1、用户 2 都是不同主题类别应用安装数量较近,对游戏类较感兴趣的用户;用户 5 则对应用类别没有明显倾向性.所以,将这 3 个用户划分到人群 2 这类人群是合理的.

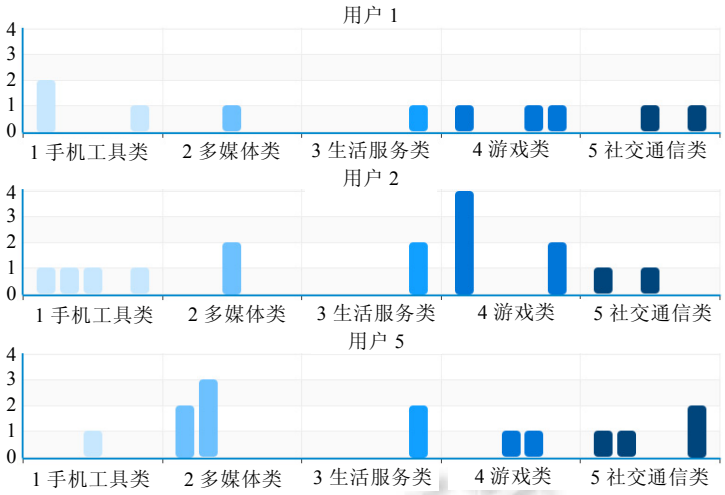


Fig.8 Distribution histogram No.1 of the sampling users' installed applications

图 8 抽样用户安装应用程序的分布直方图 1

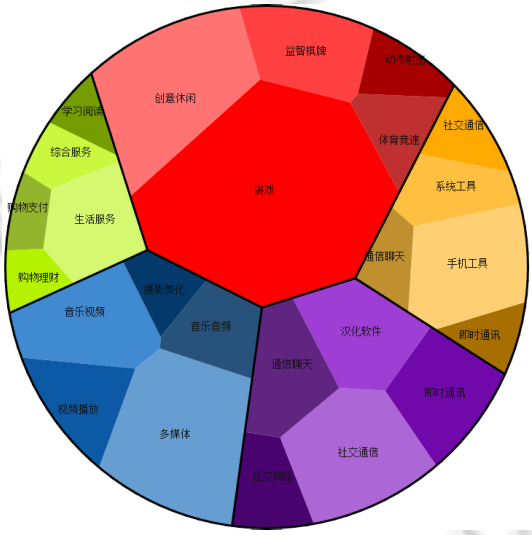


Fig.9 Visualization of the 2nd category users' application distribution

图 9 人群 2 应用程序分布特征可视化

3.3 不同人群的用户数据分析

我们继续使用上面随机选取的 5 名用户数据,他们的人群聚类结果以及 5 个主题类别的相关度见表 6.

Table 6 Relevancy No.2 between sampling users and topics

表 6 抽样用户和主题的相关度 2

	主题 1	主题 2	主题 3	主题 4	主题 5	聚类结果
用户 1	6.978	0.527	0.557	4.577	0.911	人群 2
用户 2	3.318	0.532	0.559	14.663	1.518	人群 2
用户 3	0.521	0.518	16.506	0.518	1.526	人群 1
用户 4	5.470	0.528	2.668	4.527	20.394	人群 3
用户 5	0.762	10.070	0.833	4.489	8.434	人群 2

其中,用户 1、用户 3、用户 4 实际手机上安装的应用信息建表 7.

Table 7 Installed application information No.2 of the sampling users

表 7 抽样用户安装应用程序信息 2

	应用标签(数量)
用户 1	输入法(1) 浏览器(1) 即时通信(1) 视频播放(1) 新闻阅读(1) 游戏(1) 竞技飞行(1) 动作射击(1) 社交网络(1)
用户 3	图书动漫(1) 金融理财(3) 综合服务(1) 新闻资讯(2) 学习阅读(3) 中文游戏(2) 即时通信(1) 通话增强(1) 网络浏览(2)
用户 4	通话增强(2) 购物支付(1) 综合服务(1) 新闻阅读(3) 游戏(1) 竞技飞行(4) 网络游戏(2) 社交网络(6) 社交微博(2) 即时通信(3) 通信聊天(1) 汉化工具(2)

用户 1、用户 3、用户 4 实际手机上安装的应用程序信息对应的分布直方图如图 10 所示.

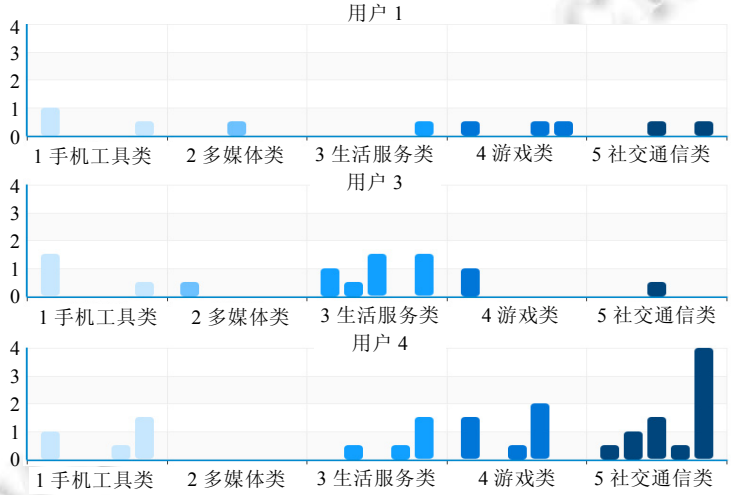


Fig.10 Distribution histogram No.2 of the sampling users' installed applications

图 10 抽样用户安装应用程序的分布直方图 2

根据表 7 和图 10,我们可以看到:用户 1 安装的应用数量不多,基本上每个主题类别应用安装了 1 个,主题 1(工具类)和主题 4(游戏类)稍微感兴趣一些;用户 3 可能是个比较注重生活的人,因为虽然安装应用数量不多,但很多应用都是主题 3(生活服务)类别的;用户 4 应该是个社交达人,因为该用户安装了较多应用,但社交类应用安装的数量具有明显优势.

从上面的分析可以得出:用户 1、用户 3、用户 4 各自具有不同的特征,所以将他们划分到不同的人群是合理的.

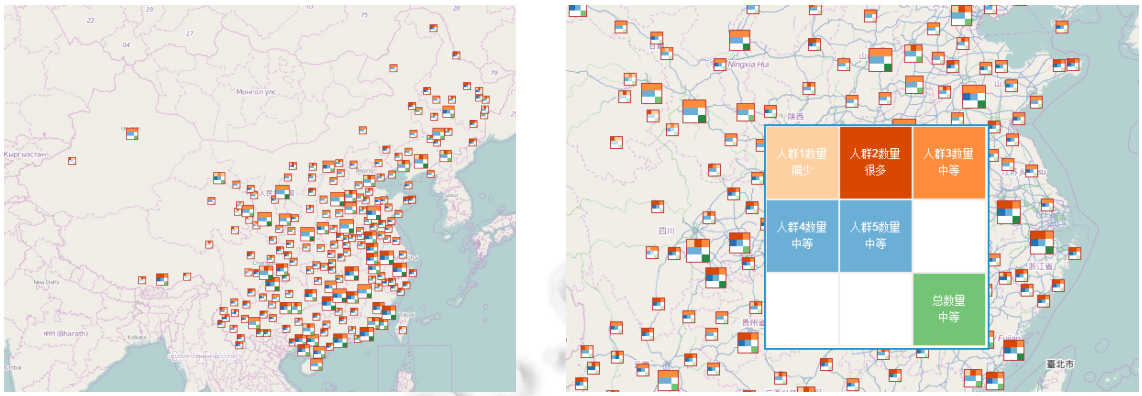
3.4 结合地理信息的人群聚类可视化

在气泡图可视化展示人群主题特征信息的同时,为了帮助用户进一步了解手机用户更为详细的信息,我们将人群聚类信息与其他维度信息相结合.在本节中,我们将着重研究人群聚类信息与地理位置信息相结合后的可视化方法:通过马赛克地域分布图,宏观地展示全国每个地方不同人群的分布情况.

通过 LDA 模型计算得到由不同主题按一定比例组成的用户人群特征后,用户自然而然很想知道他们关注的人群关于地理位置信息上的分布,这样可以更加有针对性地对特定地区进行分析调研.用户通过宏观的观察后,在人群分布多即热门地区继续推送该类人群感兴趣的主题类别的应用程序,或者对冷门地区加大推广力度.

我们采集到的手机日志数据中包含用户 ID、应用程序名称、标签信息、连接时所在的网络 IP 地址信息,可以通过每个人群所包含的用户 ID 信息确定该用户所在网络的 IP 地址.同时,我们可以通过 IP 地址转换得到对应的 GPS 地理位置坐标信息,这样,我们就得到了在每个地域上的每个人群用户数量的统计信息.

每个 GPS 经纬度坐标点,需要展示的数据总共有 6 个:人群 1~人群 5 分别的用户数量、总用户数量.我们采用 9 宫格形状的正方形来展示数据.每个经纬度坐标点对应一个正方形,该正方形由 3×2 个小正方形组成,每个小正方形对应 1 个计算后的值.每个小方格用颜色深浅代表该数据值的大小.采用上述这种马赛克图的方式来可视化展现数据,是因为在地图上,采样的 GPS 坐标点较为密集,而正方形面积利用率大,可以展示尽可能多的数据.同时,每个正方形划分为 3×2 个小正方形,结构规整,利于辨识,也方便不同的正方形之间、相同位置的小正方形上的数据比较,如图 11(a)所示.同时,为方便用户观察更为详细的信息,当用户将鼠标移动到某个正方形上后,我们将在旁边显示该正方形的放大后的详细信息,如图 11(b)所示.



(a) 人群地域分布全貌 (b) 浮动显示详情信息

Fig.11 Mosaic display of the the national population's geographical distribution

图 11 全国人群地域分布马赛克图

4 总结与展望

本文完整地给出了一个基于移动终端日志的在线可视分析系统.从海量的移动终端日志数据中,经过数据分析、主题建模,有效提取出了手机用户的特征.并通过可视化方法,将结果进行清晰展示.为移动应用开发商、应用市场平台开发商等进行分析和研究提供了帮助.

通过使用真实数据进行实验我们发现:使用从不同手机应用市场获取的应用名称-分类标签映射表,将对主题模型建立、人群聚类结果的好坏产生影响.所以,未来我们计划投入一定的人力,人工完善应用程序和分类标签的对应关系表,规范化应用程序的语义标签,从而进一步提高主题模型的准确性、人群聚类结果的代表性.移动终端日志数据中含有很多维度的信息,未来我们也将研究其他维度数据的意义,挖掘出更多有实际价值的信息.听取可视化专家、数据专家的反馈意见,并结合实际用户的使用感受,进一步修改可视化方案,尝试新的可视化展示方法.

致谢 感谢莲子数据(http://www.lotuseed.com)为本文提供了可靠的实验数据.

References:

[1] Blei D, Carin L, Dunson D. Probabilistic topic models. Signal Processing Magazine, 010,27(6):55−65. [doi: 10.1109/MSP.2010.938079]

[2] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975,18(11):613−620. [doi: 10.1145/361219.361220]

[3] Deerwester SC, Dumais ST, Furnas GW, LandauerTK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41(6): 391−407. [doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9]

[4] Golub GH, Reinsch C. Singular value decomposition and least squares solutions. Numerische Mathematik, 1970,14(5):403−420. [doi: 10.1007/BF02163027]

[5] Raghavan VV. Wong SKM. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, 1986,37(5):279−288. [doi: 10.1002/(SICI)1097-4571(198609)37:5<279::AID-ASI1>3.0.CO;2-Q]

[6] Hofmann T. Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 1999. 50−57. [doi: 10.1145/312624.312649]

[7] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[8] Salton G, Yang CS. On the specification of term values in automatic indexing. Journal of Documentation, 1973,29(4):351−372. [doi: 10.1108/eb026562]

[9] Salton G, Yang CS, Yu CT. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 1975,26(1):33−44. [doi: 10.1002/asi.4630260106]

[10] Eick SG. Graphically displaying text. Journal of Computational and Graphical Statistics, 1994,3(2):127−142.

[11] Stasko J. Information visualization. 2008. http://www.cc.gatech.edu/classes/AY2004/cs7450_spring/

[12] Feng YD, Wang GP, Dong SH. Information visualization. Journal of Engineering Graphics, 2001,(Suppl.):324−329 (in Chinese with English abstract).

[13] Johnson B, Shneiderman B. Tree-Maps: A space-filling approach to the visualization of hierarchical information structures. In: Proc. of the IEEE Conf. on Visualization'91. IEEE, 1991. 284−291. [doi: 10.1109/VISUAL.1991.175815]

[14] Balzer M, Deussen O, Lewerentz C. Voronoi treemaps for the visualization of software metrics. In: Naps T, ed. Proc. of the 2005 ACM Symp. on Software Visualization. New York: ACM Press, 2005. 165−172. [doi: 10.1145/1056018.1056041]

[15] Friendly M. A brief history of the mosaic display. Journal of Computational and Graphical Statistics, 2002,11(1):89−107. [doi: 10.1198/106186002317375631]

[16] Wang W, Wang H, Dai G,Wang H. Visualization of large hierarchical data by circle packing. In: Grinter R, ed. Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. New York: ACM Press, 2006. 517−520. [doi: 10.1145/1124772.1124851]

[17] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006, 101(476):1566−1581. [doi: 10.1198/016214506000000302]

[18] Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Analysis, 2006,1(1):121−144. [doi: 10.1214/06-BA104]

[19] Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Methodological), 1993,55(1):3−23.

[20] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley & Sons, 1990.

[21] Rissanen J. Modeling by shortest data description. Automatica,1978,14(5):465−471. [doi: 10.1016/0005-1098(78)90005-5]

[22] Okabe A, Boots B, Sugihara K, Chiu SN. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. 2nd ed., England: John Wiley & Sons, 2009.

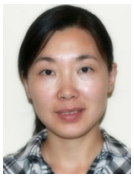
附中文参考文献:

[12] 冯艺东,汪国平,董士海.信息可视化.工程图学学报,2001(增刊):324−329.

张宏鑫(1975－),男,浙江杭州人,博士,副教授,CCF 专业会员,主要研究领域为计算机图形学,可视化,云计算.

徐沛原(1992－),男,硕士,主要研究领域为信息可视化.

盛风帆(1992－),女,硕士生,CCF 学生会员,主要研究领域为虚拟现实,信息可视化.



汤颖(1977－),女,博士,副教授,CCF 专业会员,主要研究领域为信息可视化,计算机图形图像,虚拟现实.