



专题:大数据技术与应用

移动互联网数据可视化技术及应用研究

张 青,陶彩霞,陈 肿

(中国电信股份有限公司广东研究院 广州 510630)

摘 要:大数据时代的到来给电信运营商带来了更多的机会和挑战,为从日趋庞大的数据中快速提炼出有用的信息,亟需研究大数据可视化技术,以有效支撑精准分析和决策。总结了移动互联网数据可视化新技术的特征,并结合多个移动互联网客户画像的典型场景,深入探讨数据可视化技术的实际应用。

关键词:电信运营商;移动互联网;数据可视化;大数据

doi: 10.3969/j.issn.1000-0801.2014.10.002

Research on Mobile Internet Data Visualization Technology and Application

Zhang Qing, Tao Caixia, Chen Chong

(Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: Big data era brings more opportunities and challenges to the telecom operators. In order to effectively support the accurate analysis and decision making, visualization technology is in urgent need of large data to extracting useful information from the increasingly large data quickly. The features of new technology of mobile internet data visualization were summarized, and combined with typical scenes of more mobile internet customer portrait, practical application of data visualization technology was deeply discussed.

Key words: telecom operator, mobile internet, data visualization, big data

1 引言

进入 Web 3.0,用户可以通过手机随时随地参与内容的创造和传播,形成移动与互联网融合的移动互联网时代。移动互联网具备基于个人身份/个人位置服务的特点,移动互联网时代的到来使得运营商面临的“被管道化”威胁更为严峻,终端、平台和应用已经成为新的赢利点。

对于电信运营商来说,流量是今后收入的重要利益增长点。为了更好地支撑移动互联网业务的发展,需要对移动互联网的数据进行合理采集、解读与表达,从中准确地捕捉用户的偏好、业务热点等移动互联网用户特征,数据可视化无疑是让数据更加易用和便于理解的最有效途径。

数据可视化顺应大数据时代的到来而兴起,本文在阐述数据可视化概念后,总结了移动互联网数据可视化新技术的特征,并进一步探讨了移动互联网用户数据可视化的应用场景。

2 数据可视化的定义和分类

数据可视化^[1]是指综合运用计算机图形、图像、人机交互等技术,将采集或模拟的数据映射为图形、图像、视频或动画,并允许用户对数据进行交互分析的理论、方法和技术。现代主流观点将数据可视化分为传统的科学可视化和信息可视化两类。

科学可视化面向科学和工程学领域的的数据,如含空间

坐标和几何信息的三维空间测量数据、计算模拟数据和医学影像数据等,重点探索如何以几何、拓扑和形状特征来呈现数据中蕴含的规律;信息可视化的处理对象是非结构化、非几何的抽象数据,如金融交易、社交网络和文本数据,其核心挑战是针对大尺度、高维复杂数据,如何减少视觉混淆对有用信息的干扰。由于数据分析的重要性,将可视化与分析相结合,形成一个新的学科——可视分析学。

可视分析学是一个多学科领域,涉及以下 4 个方面:一是分析推理技术,能使用户获得深刻的见解,这种见解直接支持评价、计划和决策的行为;二是可视化表示和交互技术,充分利用人眼带宽通道的视觉能力来观察、浏览和理解大量的信息;三是数据表示和变换技术,以支持可视化和分析的方式转化所有类型的异构和动态数据;四是支持分析结果的产生、演示和传播的技术,能与各种观众交流有适当背景资料的信息。可视分析的学科交叉组成及涉及学科如图 1 所示。

3 移动互联网数据可视化技术

移动互联网时代,综合多种媒体获取和理解信息已经成为信息传播的发展潮流。文本作为人类信息交流的主要载体之一,对其进行可视化能够有效帮助人们快速理解和获取其中蕴含的信息;近年来,社交网络发展迅速,其用户数量呈爆炸式增长,对社交网络进行可视化,将社交网络信息以生动、易理解的方式呈现,可以直观地解释隐藏在社交网络背后的结构模式;日志数据记录了用户随着时间变化的行为特征信息,用可视化的方式呈现日志数据中隐含的信息,理解被记录对象的行为特征。本节主要介绍移动互联网数据中文本数据、社交网络数据以及日志数据的可视化方法和技术。

3.1 文本与文档可视化

文本信息在人们日常生活中几乎无处不在,如新闻、邮件、微博、小说和书籍等。面对海量涌现的电子文档和类文本信息,利用传统的阅读方式解读电子文本已经变得越来越低效。因此,利用可视化和交互的方式生动地展现大量文本信息中隐含的内容和关系,是提升理解速度、挖掘潜在语义的必要途径之一。

文本可视化的研究主要包括基于文本内容的可视化、基于文本关系的可视化、基于多层面信息的可视化^[2]。

(1) 基于文本内容的可视化

面对海量文本,人们需要对每个文本或者整个文本集合的主要内容进行快速浏览。最典型的可视化形式是“标签云”(tag cloud)。标签云一般使用字体的大小和颜色对关键词的重要性进行编码。越重要(权重越大)的关键词的字体越大,颜色越显著。

(2) 基于文本关系的可视化

基于文本关系的可视化研究文本的内外关系,帮助人们理解文本内容和发现规律。常用的可视化形式有树状图和节点连接的网络图,主要关注文本的内部结构和语义关系。

(3) 基于多层面信息的可视化

基于多层面信息的文本可视化主要研究如何结合信息的多个方面帮助用户从更深层次理解文本数据,发现其内在规律。其中,包含时间信息的文本可视化近年来受到越来越多的关注,时间信息提供了关于文本内容变化、数据规律等方面的重要信息。

3.2 社交网络可视化

社交网络服务是基于互联网中人与人之间的相互联系、信息沟通和互动娱乐的运作平台。Facebook、微信、微博等都是当前普及的社交网站。社交网络是一个网络型结构,

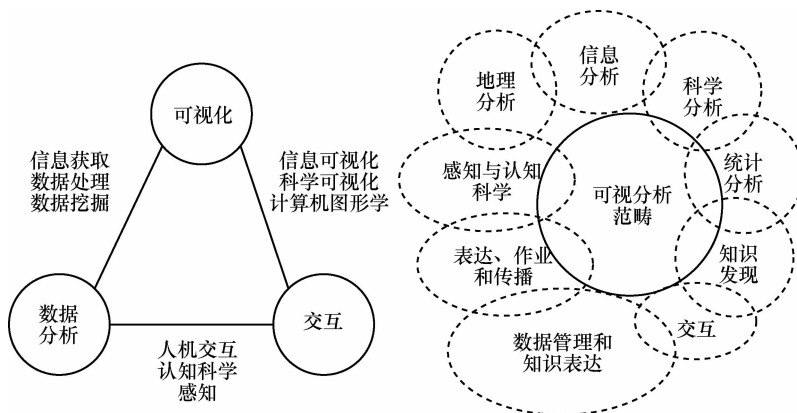


图 1 可视分析的学科交叉组成及涉及学科



由节点和节点之间的链接组成。这些节点通常指个人或者组织,节点之间的链接关系有朋友关系、亲属关系、关注或转发关系、支持或反对关系或者拥有共同的兴趣爱好等。

根据可视化所需展示的内容,社交网络的可视化方法可以分为4类:结构型、统计型、语义型和时序型,具体介绍如下。

(1)结构型

结构型可视化着重于展示社交网络的结构,即体现社交网络中的参与者和他们之间关系的拓扑结构。常用的可视化方法是节点链接图,其中的节点表示社交网络的参与者,节点之间的链接表示两个参与者之间的某种联系,如亲属关系、拥有共同兴趣爱好等,反映一个社交网络中的聚类、社区、潜在模式等。

(2)统计型

社交网络某些特性统计变量的分布(如节点的度数、中心性、聚类系数)可用柱状图、折线图、饼图等基本统计图表进行可视化。

(3)语义型

社交网络是现实世界的反映,蕴含着丰富的语义信息。对复杂社交网络中的语义信息进行可视化,可以有效地发现社交网络中的舆情和突发事件等。

(4)时序型

社交网络中的用户行为具有时间戳,将时间信息作为属性融入社交网络的可视化,可反映社交网络的动态变化情况。

3.3 日志数据可视化

日志数据可以理解作为一种记录所观察用户行为信息的数据。日志数据记录了对象随着时序变化的行为特征信息,用可视化的方法呈现日志数据中隐藏的大量不规则数据信息,可有效帮助用户挖掘日志数据中所含的信息,理解被记录用户的行为特征^[9]。

针对不同领域、不同类型的日志数据,有不同的可视化需求和方法,主要包括商业交易、移动轨迹和系统日志数据3类,具体介绍如下。

(1)商业交易数据可视化

电子商务交易平台每时每刻都在记录用户的交易信息,包括个人信息和每一笔交易记录,用户个人信息与交易记录具有巨大的分析价值,如分析买家的购买记录和个人特征,可挖掘出特定类型商业的潜在购买用户。商业交易数据的可视化,可直观形象地展示数据,提高分析效率。

(2)移动轨迹数据可视化

GPS等空间定位技术以及无线通信和移动计算的快速发展,使得实时跟踪和记录移动对象的轨迹或其他相关信息变为现实,移动数据轨迹信息的可视化通常结合地理信息,直接在地图上展示,代表性方法有热力图(heatmap)。

(3)系统日志数据可视化

系统日志数据记录了机器或集群的运行性能等信息,被广泛用于实时监控。基于可视化的系统日志数据分析是未来趋势,如在线日志可视分析软件 Loggly、Splunk 等。

4 移动互联网用户数据可视化应用

移动互联网时代,以用户的位置信息、终端机型、时间信息为基础,结合用户个人习惯信息、在线用户状态、用户需求,为运营商提供了一幅轮廓清晰的客户画像。根据这样的画像,精确的广告投放正在变成现实。在此基础上引入可视化技术,使数据的说服力、可读性和逻辑性得到了增强,以便于了解庞大的多维数据间的相互关系及发展趋势。

可视化不是单独的一个算法,而是一个流程。除了视觉映射外,也需要设计并实现其他关键环节(如前端的数据采集、处理和后端的用户交互)。这些环节是解决问题必不可少的步骤,并且会直接影响可视化效果。选取客户画像中的3个典型场景:用户行为特征、社交关系及兴趣关系特征、用户位置特征,进一步探讨数据可视化技术的实际应用。

4.1 用户行为特征数据可视化

在流量经营精细化方面,借助 DPI (deep packet inspection,深度分组检测)技术^[4]等,建立客户超级细分模型,为各细分群组客户打上互联网行为标签,从而帮助运营商完善用户的360°画像,深入了解用户行为偏好和需求特征;其次,根据用户行为偏好,推送合适的业务,并根据对用户特征的深入理解,建立用户与业务、资费套餐、终端类型、在用网络的精准匹配,实现全程精准营销。

以下根据用户的行为数据进行特征偏好计算和展示。

(1)数据采集

从 DPI 数据采集设备获取已生成的数据文件,将文件中的用户信息入库,得到3G用户互联网访问记录清单,并从业务系统获取用户基本信息(包括性别、年龄、在网时长、套餐、终端等)。

(2) 数据处理和转换

DPI 分析日志经抽取后,按两个关键库(URL 规则库、应用库)完成用户客户端和访问内容的准实时归类和统计,基于分析数据归类统计结果,输出以下关键内容:TopN 应用、访问偏好分类、访问次数、用户兴趣度。针对用户搜索行为,首先对搜索关键词进行预处理,以排除原始数据中一些无用或冗余的信息;接着进行分词和词干提取处理,抽取其中的特征信息,形成搜索关键词、搜索分类、搜索次数、搜索偏好度等分析挖掘数据。DPI 日志处理规则如图 2 所示。

(3) 可视化映射

用户行为特征属于高维数据,难点在于如何呈现单个数据点的各属性值分布以及比较多个数据点之间的属性关系。区域可视化将全部数据点的全部属性以区域填充的方式在二维平面布局,并采用颜色、距离等视觉通道呈现数据属性的具体值,所有数据点在空间中布局排列,方便用户进行对比。

用户移动互联网行为特征的数据可视化映射包括以下 4 个方面:

- 用户的基本属性,包括性别、年龄、在网时长、套餐、终端等;
- 用户的移动互联网访问偏好,即访问分类,不同颜色、距离反映不同热度;
- 用户的应用使用偏好,直接显示用户最常使用的 TopN 个应用,不同颜色、距离反映不同使用频率;
- 用户的搜索偏好,类似总体搜索偏好的分析方法,

采用词云+分类统计方式,不同颜色、距离反映不同搜索频次。

由于涉及海量日志信息的处理以及大量节点图形的表现,为了在可视化显示上尽可能地实现高效,采用 Highcharts 实现图形展示。如图 3 所示,圆圈中心的头像代表性别,外一圈是用户的基本信息,最外圈是用户的偏好(颜色、距离的不同代表不同的热度)。



图 3 用户行为特征数据可视化

4.2 用户社交关系及兴趣关系可视化

通信关系是电信用户之间常规的通信往来信息,如通话、短信、彩信等,由于电信企业对用户计费的需要,这类信息一般在企业各类业务计费系统中长期保存,是最常用来进行电信用户社交网络分析的通信关系数据,一般采用计费系统中原始的通话详单记录(call detail record, CDR)。

以下根据用户的通话详单进行社交关系分析,并在用

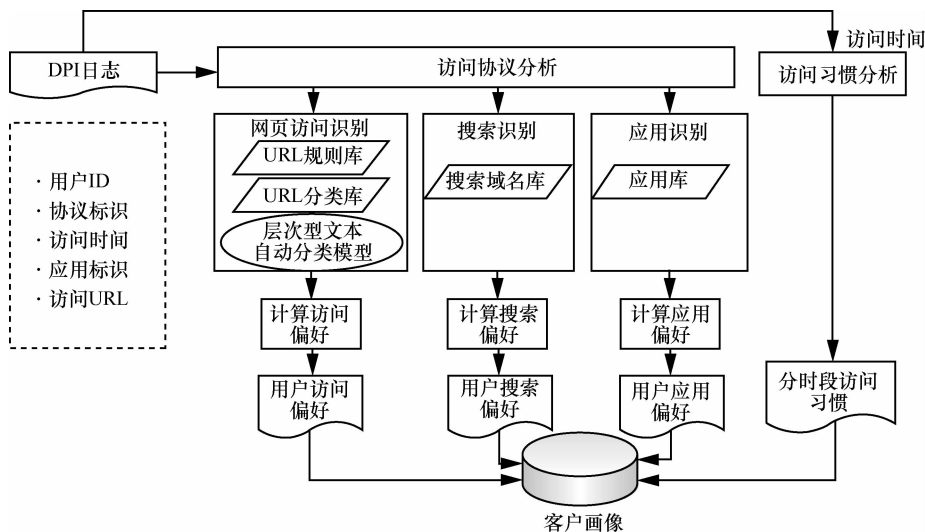


图 2 DPI 日志处理规则



户的社交关系圈内进行用户兴趣图谱的计算和展示。

(1)数据采集

对于社交网络分析^[9],最重要的数据是支撑建立节点间边的数据。在电信企业,这些数据包括通话详单记录数据,代表用户与用户间连接的方式。由于CDR数据量庞大,有必要以月为周期进行数据抽取,并根据业务问题分析,按月进行月均指标计算。

(2)数据处理和转换

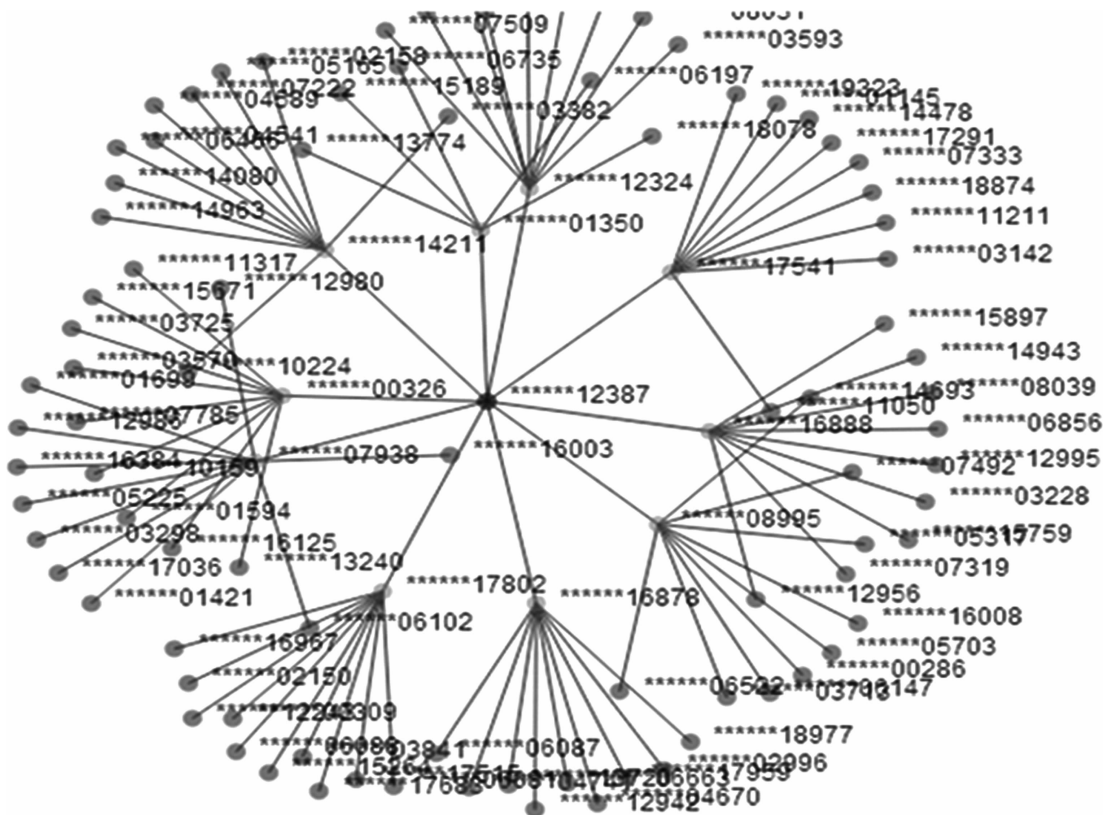
描述一个社交网络,最常用的指标就是一阶中心度和二阶中心度。一阶中心度指与一个特定节点存在直接联系的节点数量;二阶中心度则指与原始点直连的节点的连接点的个数。这两个基础指标的计算非常重要,是确定电信用户社交网络中关键节点(即有影响力用户)的核心指标,是用户影响力因素指标计算的主要构成。

另外,由于来话和去话的资费价值不同,用户连接的两个不同方向有不同意义,因此一阶和二阶中心度的来话、去话要分别进行计算,而用户间通话的次数、时长以及平均单次通话时长都将一并考虑,作为用户间关系强度基础指标分别进行计算。

(3)可视化映射

根据数据特征,采用节点链接图展现用户社交关系:节点表示社交网络的参与者,节点之间的链接表示两个参与者之间的某种联系,如通信关系、拥有共同兴趣爱好等。根据节点链接的方向进行聚类,方向一致的聚在一起,从而降低视觉复杂度,清楚地显示复杂的网络图,并且体现节点层次的连接关系。用户可以快速、准确地找到自己感兴趣的节点,如图4所示。

图4中,当用户对某个具体的节点感兴趣时,可以选中该节点,则该节点、与其相连接的节点以及相应边都突出显示出来,而其他的节点及边的不透明度降低。同时,有关选中节点的统计信息,如性别、语音通话次数、消息发送次数等将被显示出来。一个节点表示一个手机用户,节点的大小表示用户通信频度。边用来连接两个节点,表示用户间社交关系的紧密程度。此外,可以通过柱状图直观地看出该节点好友的亲密度统计信息,如图5(a)所示。基于用户社交圈展现用户访问偏好的兴趣相似度情况,如图5(b)、图5(c)所示,柱状高度、颜色体现用户间兴趣的相似情况。



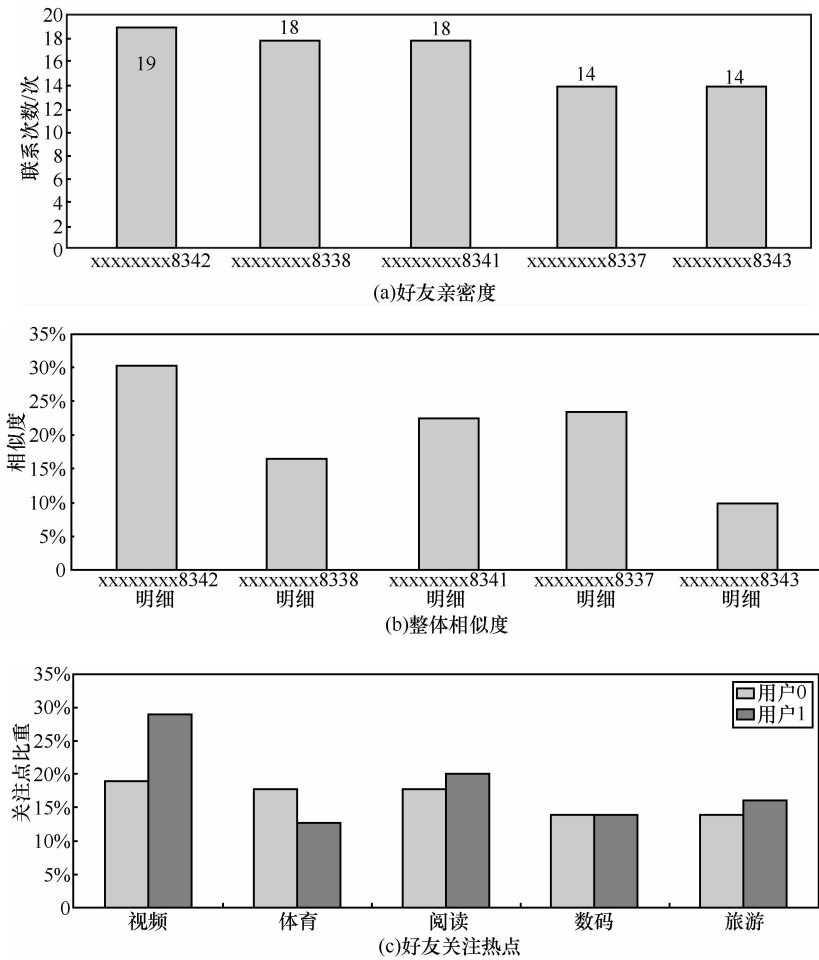


图 5 用户兴趣关系数据可视化

4.3 用户位置特征可视化

用户轨迹分析的目的在于从用户的历史行为轨迹中挖掘和解释用户的日常行为规律。在对这一问题的研究中,聚类分析是最常用的技术手段,然而由于缺乏必要的验证信息,对聚类结果的解读通常需要结合特定的时空上下文进行。如 Kirmse 等人对用户日常行为特点的研究^[6],使用 meanshift 聚类根据用户的历史行为轨迹数据生成了其日常驻留区域,并结合时间特点给出结果的语义解释(如用户在白天驻留时间最长的地点是工作场所,晚上驻留时间最长的地点是家里)。

以下基于用户注册的基站情况变化,展示用户一天的位置移动轨迹情况。

(1) 数据采集

原始的数据是由基站 ID 与对应的时间序列组成的。如果用户处在某个基站范围内的时间比较长,就会出现有几条记录的情况,将对后续的分析产生影响。因此,将相同

的基站记录合并为一条记录,同时,原始记录对应的时间也需要合并,以得到进入该基站和离开该基站的时间。

(2) 数据处理及转换

前面的数据预处理,已经将原始数据表中的噪声数据去除,重新构造了数据的记录格式,构造成 $[T_{start}, T_{end}, areaID, cellID]$ 的三元组形式。还需进一步对重构的基站数据(即地点信息)进行再处理,构造出用户移动路径的集合。在基站记录中,在该位置停留的时间占整个记录的比例若大于某个阈值,这里即路径终点。在用户移动的过程中,如果基站发生切换,离开前一个基站的时间就是该基站的结束时间,该数据被记录并生成一条新的记录,新记录存储进入前一个基站的时间,即开始时间以及基站的结束时间。同时,要存储当前基站的开始时间,直到下一次基站切换发生。

考虑现实生活中这样的情况,用户到达工作地点开始工作,或者到达目的地长时间没有移动,那么对应于数据中,就会出现一条记录的开始时间和结束时间的间隔很



图6 用户位置特征可视化

长,可以认为该路径应该终止。这里定义了地点持续时间,即与该地点对应的基站的通信时间,就是在一条记录中基站的结束时间和开始时间的差,该时间差表示处在某个基站信号范围内的时间。

(3)用户移动轨迹可视化映射

用户移动轨迹属于地理空间数据可视化,这里采用地理信息的线性数据可视化,表达多对象轨迹变化的地图,如图6所示,曲线的宽度代表流量的大小,如实地呈现轨迹的源头、合并、分散、路径改变和汇入的动态过程,是一种基于聚类 and 层次结构的地理信息简化方法。

5 结束语

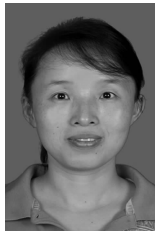
在大数据时代,数据的数量和复杂度的提高带来了数据探索、分析、理解和呈现的巨大挑战。除了直接统计或者数据挖掘的方式,可视化通过交互式视觉表现的方式帮助人们探索和解释复杂的数据。本文在分析数据可视化理论的基础上,总结了主流的移动互联网数据可视化技术,并结合客户画像的典型场景,对数据可视化过程进行深入的实践研究,为下一步数据可视化应用的开展提供贴合实际的参考依据。

参考文献

- 1 陈为,张嵩,鲁爱东.数据可视化的基本原理和方法.北京:科学出版社,2013

- 2 唐家渝,刘知远,孙茂松.文本可视化研究综述.计算机辅助设计与图形学学报,2013(3)
- 3 乔志伟.基于用户行为的3G业务分析与探讨.移动通信,2010(12)
- 4 张青,陈翀,向勇.深度分组检测技术研究及在流量经营中的应用.电信科学,2013,29(8)
- 5 漆晨曦.电信客户社交网络分析方法与营销应用探讨.电信科学,2012,28(7)
- 6 陈康,黄晓宇,王爱宝等.基于位置信息的用户行为轨迹分析与应用综述.电信科学,2013,29(4)

[作者简介]



张青,女,中国电信股份有限公司广东研究院高级工程师,主要从事大数据处理、数据仓库、数据挖掘及客户营销维系等方面的研究工作。

陶彩霞,女,中国电信股份有限公司广东研究院高级工程师,主要研究方向为大数据、移动互联网、云计算。

陈翀,女,中国电信股份有限公司广东研究院工程师,主要研究方向为大数据、云计算。

(收稿日期:2014-09-15)