



移动互联网场景中客户特征分类技术研究

张 青

(中国电信股份有限公司广东研究院 广州 510630)

摘 要:在分析客户特征分类理论的基础上,总结了目前主流的网页分类实现技术,并进一步分析了客户特征分类常用的挖掘算法及其适用场景,最后探讨了客户特征分类过程,为移动互联网中客户特征分类工作的开展提供了贴合实际的参考依据。

关键词:客户特征;网页分类;数据挖掘

doi: 10.3969/j.issn.1000-0801.2014.01.021

Research on Classification Technology for Customer Characteristic in Mobile Internet Scene

Zhang Qing

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

Abstract: Based on the analysis of the characteristic of customer classification theory, the current mainstream webpage classification technology was summarized, and common mining algorithms and applicable scene of customer classification were deeply analyzed. Finally, the classification process of customer characteristic was discussed to provide practical reference for developing customer feature classification work under mobile internet.

Key words: customer characteristic, webpage classification, data mining

1 引言

1.1 背景

当前移动互联网正处于爆发式增长阶段,OTT(over the top)服务提供商正在逐渐积累自身的客户资源,以提供基础文字、语音服务、互动性的分享与交互的应用,OTT服务的盛行占用了大量的网络带宽资源,却未给运营商带来相应的收益,同时还冲击到运营商传统的语音、短信、彩信等业务。OTT服务的出现使得数据流量所带来的价值不断地向OTT服务提供商转移,电信运营商“被管道化”的趋势越发明显,转型迫在眉睫。

在大数据时代,电信运营商必须加强数据流量经营理念,探索电信精细化运营方式,实现用户级的网络资源及

业务的控制,细分业务流量,将业务流量与用户、网络资源相匹配,实现对业务流量的甄别和控制。通过数据挖掘识别用户行为习惯和偏好,为用户提供个性化、差异化的信息服务,从而实现单位流量价值的最大化。

1.2 需求

长久以来对语音、短信等业务营销的支撑以及对客户的认知,主要基于客户基础资料 and 客户的语音、短信等通信行为,通过统计和数据挖掘等手段分析和认识客户,进而提供客户营销和服务等支撑。该方法在以往的业务发展中,对移动用户发展、ARPU值提高等起到了非常重要的作用。但面对支撑以移动互联网为基础的新业务发展目标时,该方法表现得力不从心,因为客户通信行为和基础资料等数据本身并不包含业务内容和客户偏好信息,从而无

法更好地实现移动互联网新业务的推广工作。

为了更好地支撑移动互联网业务的发展,需要对客户访问移动互联网的行为进行深入分析,获取访问行为里的丰富信息,如客户在手机上浏览了什么新闻和小说,通过搜索引擎搜索了什么关键字,在购物网站上购买了什么产品等。通过及时、准确地捕捉客户的偏好、业务热点等移动互联网客户特征,为后续的精细化运营奠定基础。

2 客户特征分类理论基础

客户特征是指通过计算、归纳、关联、数据挖掘等方式,利用生产系统中所有与客户相关的数据,包括客户资料、客户消费和接触等,为每个客户从不同角度标注个性化的数据标识。通常情况下,分析客户特征的维度包括:人口统计学特征、社会群体特征、性格心理特征以及电信业务特征,由此形成精细化营销的客户特征体系^[1]。

随着移动互联网逐渐渗透到人们生活、工作的各个领域,已有的特征体系缺乏对用户上网行为习惯和用户偏好等方面信息的收集,从而难以对用户进行完整“画像”,极大地影响了业务的精准营销和良性发展。

以下基于本体理论、马斯洛需求层次理论,研究移动互联网客户特征库的构建。

2.1 本体论

本体最权威的定义由 Gruber 提出,本体是对世界上客观存在物体的系统描述,是将现实世界中的某个领域抽象为一组概念和概念之间的关系。其中,概念是领域内具有公认语义的词语,概念间关系是实现领域内各种语义的描述。根据描述的对象可将本体分为常识本体和领域本体^[2]。其中,常识本体描述整个语言范围内的概念,领域本体描述特定领域中的概念。

基于本体论对电信客户特征进行分类,可以让商家对客户的行为偏好进行有针对性的营销。以下从渠道偏好、消费偏好、访问偏好、搜索偏好、专题偏好和应用偏好6个方面构建其特征本体分类,如图1所示。渠道偏好是客户在办理电信业务时所使用的渠道,了解客户的渠道偏好能够帮助电信企业选择最合适的产品和服务渠道;消费偏好是客户在电信业务消费和套餐选择上的偏好;访问偏好是客户访问的互联网方式和时段等的行为特征与偏好;搜索偏好是客户搜索互联网信息所使用的搜索引擎偏好以及表现出的搜索行为特征;专题偏好是客户浏览互联网信息时所感兴趣的专题信息,对于捕捉客户的消费倾向具有重

要意义;应用偏好是客户在使用各种应用软件方面表现出的偏好和行为特征。

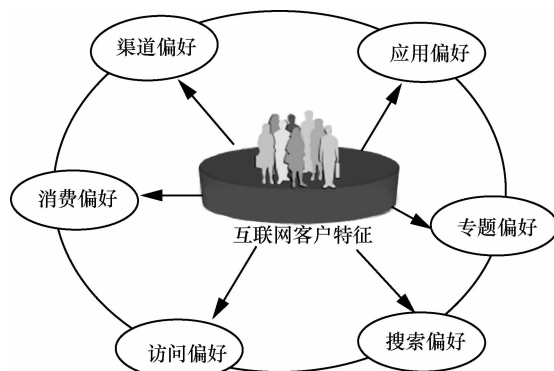


图1 电信客户在互联网业务中特征本体分类

2.2 马斯洛需求层次理论

社会心理学家马斯洛研究发现,所有人的需求分为5个层次,即生理生存需求、自身个体安全需求、归属感和社交需求、认可和尊重需求以及自我实现需求,依次由较低层次到较高层次。移动互联网作为满足人类需求的手段,至少已经拓展到4个方面:信息渠道、沟通、娱乐和生活助手功能。业界普遍认为,网民的上网行为可归结为“类马斯洛理论”,如图2所示。

1级生存需求:对食物、水、空气和住房等的需求,级别最低,人们在转向较高层次的需求之前,总是尽力满足这类需求。在互联网,信息渠道功能作为互联网的基本功能已被中国网民广泛使用,门户最初提供的生活、房产等网站内容满足网民最基本的生理需求。

2级安全需求:包括对人身安全、生活稳定的需求。安全需求对网民来说简单得多,即网页是否含有病毒、流氓软件、弹窗广告以及相应的自己不喜欢的如各种含有色情、暴力等垃圾内容,当然还包括网站速度快、内容不错、能让网民有一个稳定的上网环境等。

3级社交需求:包括对友谊、情感以及隶属关系的需求。移动互联网的社交需求通过即时通信、电子邮件、网上阅读、网上音乐、网络游戏、网上银行和网上炒股等方式实现。目前用户的这些需求还没有被完全激发,使用率不高。移动互联网还没有深入大多数网民生活的方方面面。

4级尊重需求:既包括对成就或自我价值的个人感觉,也包括他人对自己的认可与尊重,目前在一些论坛、QQ群、资源共享等领域体现得较为明显。一个网民在论坛发表的一个好帖或者分享的一段精彩视频,如果能够得到

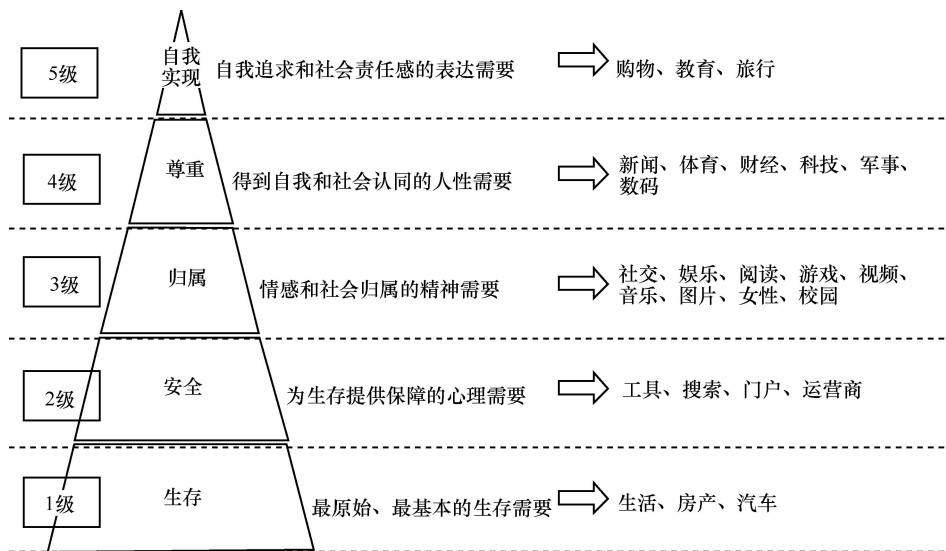


图2 采用马斯洛需求层次理论搭建客户行为框架

其他网友的认同,引起网友的关注,都将是网民尊重需求的实现。

5级自我实现需求:是最高层次的需求,指个人价值得到社会实现的追求,如努力获得某种成就,尽量发挥自我潜能,追求崇高理想的实现等。一方面,利用网络资源获得个人价值、成就感、满足成功欲望的产品或服务,如网上教育、网络购物、网上旅行预订等;另一方面,通过网上求职方式为人们提供有用的就业信息,为消费者提供发展或表现自身价值的场所或机会。

3 网页分类技术

目前移动互联网活动大多通过浏览器或应用完成,通过对上网记录进行数据挖掘和行为分析可获得用户上网行为偏好,从而获得用户所关注的信息。如何从浏览网页后产生的类似URL数据中分析得到用户行为意向?目前对网页URL进行自动分析、归纳网页类别已经成为快速且有效解决这一问题的关键,也是构建客户行为特征库的关键。URL分类库的构建如图3所示。

3.1 URL 特征分类

(1)规则 URL

使用多级域名或目录对网站内容URL进行编码:建立热门网站频道和URL分类的映射关系;通过爬虫技术收集热门网站的具体URL,进行URL分类,如http://sports.sina.com.cn/nba/: 体育健身—体育竞技—球类运动—篮球。

(2)无规则 URL

无规则或所有栏目使用同一格式编码:从上网日志中获取用户访问的无规则URL;通过网络爬虫爬取URL对应的网页内容;利用分词技术,把内容拆分成单个词;利用文本分类算法,通过关键词匹配,识别URL分类。URL地址单词解析示例:http://movie.youku.com/: 休闲娱乐—影视—影视类型—电影。

对于热门网站,重点抓取、分析,同时为提升匹配的准确率和覆盖率,采用智能聚类分类模型。该模型对用户偏好分析采用多级分类方式(如新闻一级分类,可进一步细分为二级的国内新闻、国际新闻等),见表1。

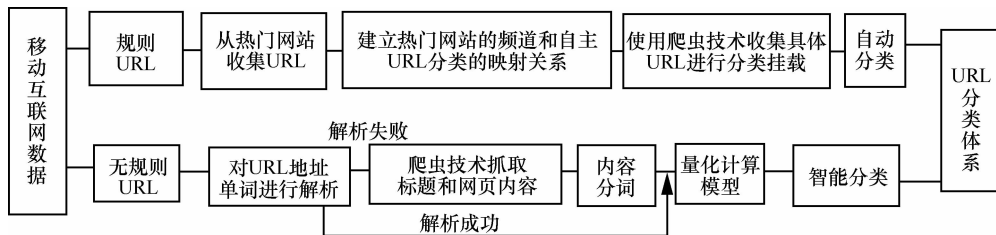


图3 URL分类库构建

表 1 URL 特征分类示例

URL	一级分类	二级分类	三级分类	四级分类
zhj8.sinaapp.com/*	生活服务	资讯	财经	黄金
static.alimama.com/wangwang/wwupdate	公司企业	商业经济	广告营销	广告营销
static.apc.360.cn/cms	电脑网络	软件下载	休闲娱乐	游戏下载
static.paipaiimg.com/js/version	生活服务	购物	商家购物	购物综合
traffic.tuyulbs.com/LocationServer/le	生活服务	旅游交通	地图导航	地图导航
tsh.pengyou.com/cgi-bin/ticker	休闲娱乐	社交	交友	交友综合
wafp.pptv.com/webdelivery/webafp	休闲娱乐	视频	在线视频	综艺
qidian.cn/wap2/chapter/chapter.do	文化教育	文学	小说	原创
qp.qq.com/cgi-bin/cgi_imgproxy	休闲娱乐	社交	交友	交友综合
qqgameplatedn.qq.com/social_hall/js	休闲娱乐	游戏	游戏平台	游戏平台
res.neng.com/res/apk_sp	电脑网络	软件下载	办公软件	办公应用
res.xjol.g.1360.com	电脑网络	软件下载	系统软件	系统软件
quote.zhijinwang.com/xml/ag.txt	生活服务	资讯	财经	黄金

3.2 应用分类

电信运营商根据目前市场及各省的手机应用下载量,选择建立标准应用库^[3],建设步骤介绍如下。

(1)与主流互联网厂商合作,分享/爬取互联网应用的下载地址、介绍、标签等和使用过程中产生的 URL、IP 地址等信息。

(2)对于无统一地址的应用,根据服务器分布进行地域验证。

(3)建立标准应用库与兴趣点分类标准映射关系。

(4)建立自有基地业务与标准应用库映射关系,如图 4 所示。

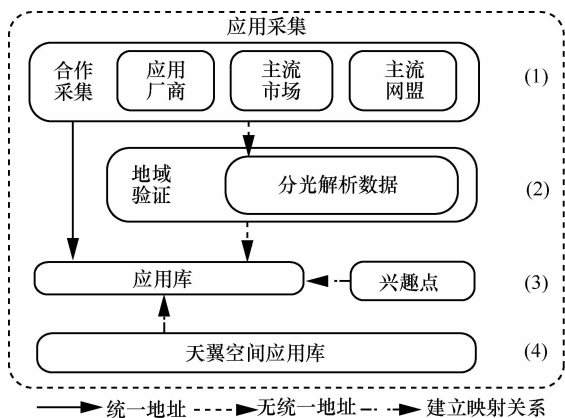


图 4 天翼空间应用与标准应用库映射关系

应用解析过程如图 5 所示,具体介绍如下。

(1)对用户的下载解析,通过域名和 URL 即可识别为家庭—购物类。

(2)对用户使用的解析,可分为公有和私有协议两种情况,分别解析为社交—交友类和交互方式—视频类。

3.3 Web 内容分词

从大量的 Web 页面描述数据中发现信息,进而抽取知识的过程,称为 Web 内容挖掘,该方法主要用于信息过滤、知识抽取、网络文档分类等场景。由于 Web 中文内容中的词与词之间无间隔,需要进行分词处理。国内汉语自动分词系统采用的分词算法主要有以下 3 种类型^[4]。

(1)基于字符串匹配的分词方法

该方法是按照一定的策略将待分析的汉字串与一个“充分大”的机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功。主要使用最大匹配法、逆向最大匹配法、逐词匹配法、设立标志法等进行分词。

(2)基于理解的分词方法

这种分词方法通过让计算机模拟人对句子的理解,达到识别词的效果。主要在分词的同时进行句法、语义分析,利用句法信息和语义信息处理歧义现象。

(3)基于统计的分词方法

该方法对语料中相邻共现的各个字的组合频度进行统计,计算它们的互现信息。这种方法只需对语料中的字组频度进行统计,不需要切分词典,因而又叫无词典分词法或统计取词方法。

4 客户特征分类常用挖掘算法

Web 数据挖掘是从移动互联网及其相关资料和行为中提取有用的模式和隐含信息,在进行基于数据挖掘的用户行为特征分析中,常用的算法包括聚类算法、决策树算法、关联规则算法及时间序列算法等,具体介绍如下。

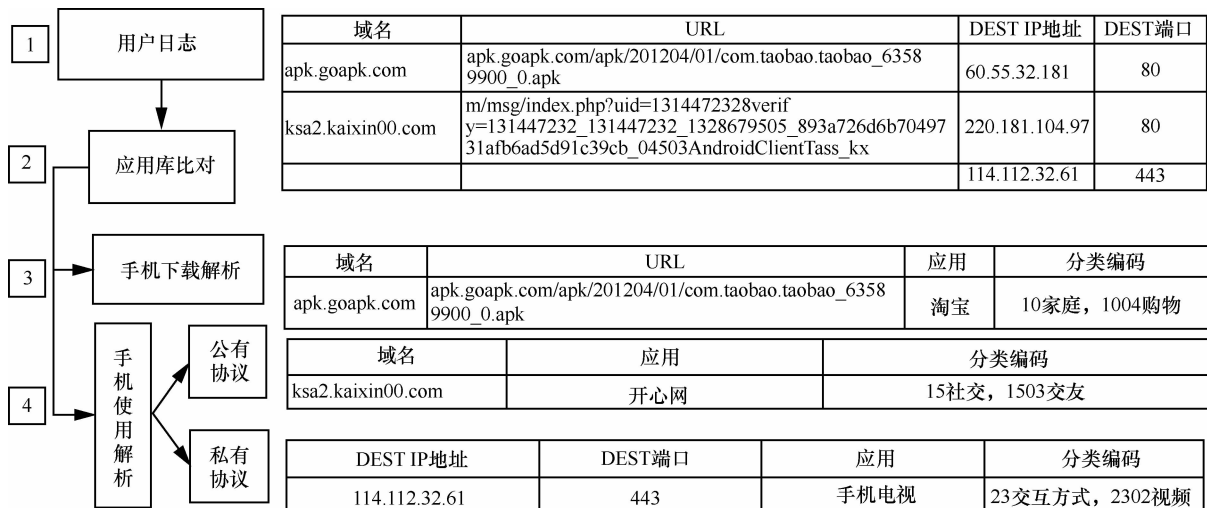


图5 应用解析过程

(1) 聚类算法

聚类算法属于探索性的数据分析方法。通常利用聚类算法将看似无序的对象进行分组、归类,以达到理解研究对象的目的。聚类结果要求组内对象的相似性较高,组间对象的相似性较低。

从统计学的观点看,聚类算法是一种通过数据建模简化数据的方法。传统的统计聚类分析方法包括系统聚类法、分解法、动态聚类法等。

在用户行为分析中,很多问题可以借助聚类算法解决,如网站的信息分类、网页的点击行为关联性以及用户社交圈问题等。聚类算法在电信领域中主要用于客户特征识别、客户群体细分等。

(2) 决策树算法

在数据挖掘分类技术中,决策树分类算法^[9]是应用最广的归纳推理算法之一,对噪声数据有很好的稳健性。决策树算法有如下优点:形成了易理解规则;计算量相对较小;可处理连续和离散数据;可显性表达属性的重要程度。利用决策树可以快速创建挖掘模型,且模型易于解释。决策树最常见的数据挖掘任务是分类,但和聚类算法不同,决策树是一种有监督的数据挖掘算法。在电信领域可以用决策树算法进行客户流失分析、风险管理和广告定位等方面的用户行为分析。

(3) 关联规则算法

关联规则是数据挖掘的一个重要课题,用于从大量数据中挖掘出有价值的数据项之间的相关关系。关联规则的任务在于减少潜在的大量杂乱无章的数据,使之成为少量的易于观察理解的静态信息。Agrawal 于 1993 年首先提出

了挖掘顾客交易数据库中的关联规则问题,后续研究人员对关联规则的挖掘问题进行了大量研究^[6],主要包括对原有算法的优化,如引入随机采样、并行思想等,以提高算法挖掘规则的效率,并对关联规则的应用进行推广。通过关联分析可以找出用户上网行为之间的关联规则,这些规则可以帮助企业针对用户的个性化需要提供更多的服务。对于电信运营商,可以分析客户购买的电信产品之间的关联,从而通过交叉销售或捆绑销售方式增加销售额。

(4) 时间序列算法

时间序列通过对数据进行分类汇总获得,可以进行长期趋势、循环分量、季节分量和不规则分量等预测分析,从而实现预测、制模、特征提取。在电信领域,可以用时间序列算法对用户的上网行为信息做出预测,进行趋势分析和周期性分析,从而推荐合适的产品套餐,以便于拉动上网流量。

5 客户特征分类过程

(1) 信息识别

首先根据客户的移动互联网行为数据找出客户访问的网站、使用的客户端、上网时段以及访问内容的类别信息,从而获取客户的上网内容和时段等的偏好,主要包括匹配规则梳理、匹配规则优化,如图 6 所示。

为了提升匹配的准确率和覆盖率,建议采用智能聚类分类模型。该模型对用户偏好分析采用多级分类方式,一级分类基于访问网址的分类属性建立;二级分类是对一级分类的细分,每个一级分类下延伸出多个二级分类以及更细层次的分类,所有网页首先采用地址方式匹配分类,如无法匹配,再通过网页内容分词技术、网络标签及关键字

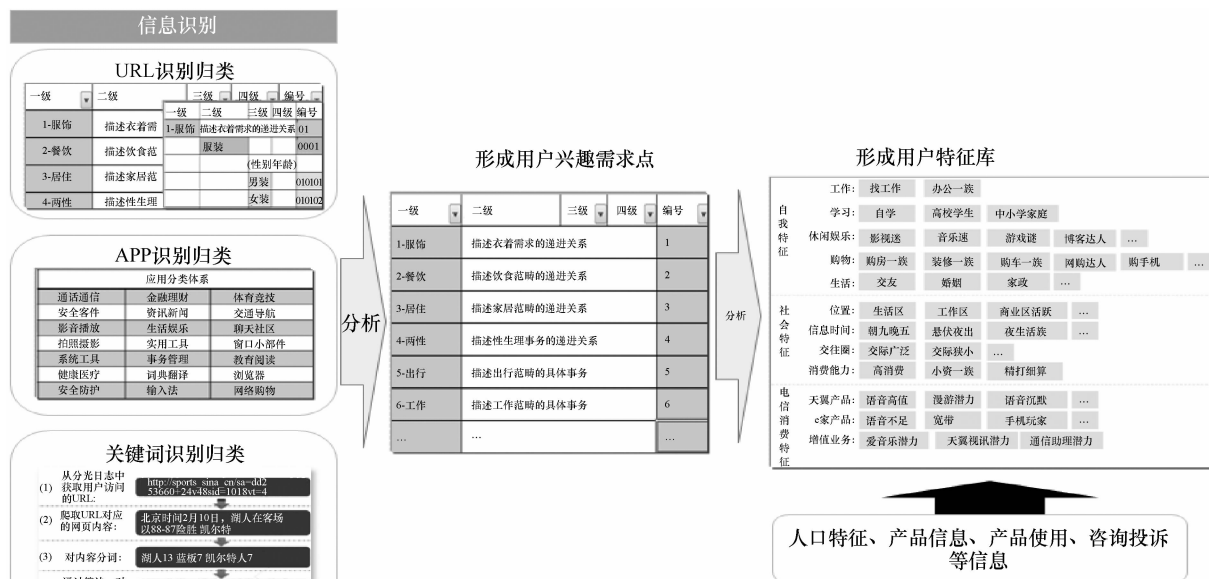


图6 客户特征分类过程

定位,根据用户访问网页确定网页属于哪个二级分类及下一级分类。

(2)形成用户兴趣需求点

定时将用户上网行为规律(如用户上网流量、用户在线时长、用户上线次数、用户上线时段等)信息以及包括用户文化、娱乐、医疗、购物等方面偏好的访问量数据按日进行汇聚,得到用户各类兴趣需求点按日计算的访问量统计值。

(3)形成客户特征库

客户特征的知识规则由数据挖掘模块获取,其知识规则由访问序列模式、聚类规则、关联规则等组成,这些知识规则反映了互联网用户的个性行为特征和之间的关联关系。客户特征库标签主要根据客户的生命周期、消费行为以及网络行为,分析出客户的数据业务偏好、终端偏好、产品偏好等内容并进行挖掘分析,从而获取客户特征数据。

6 结束语

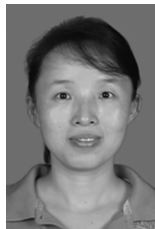
大数据时代对电信运营商营销的精准性和实时性提出了更高的要求。本文在分析客户特征分类理论的基础上,总结了主流的网页分类实现技术,进一步分析了客户特征分类常用的挖掘算法及其适用场景,并对客户特征分类过程进行深入研究,为移动互联网下客户特征分类的开展提供了贴合实际的参考依据。未来,基于移动互联网客

户特征的针对性营销还可以增加基于信令数据的客户行为轨迹的研究,结合实时营销和后服务理念进行综合的针对性的营销研究,在低成本、高效益运作的大背景下,为营销的精准性提供更有力的支撑。

参考文献

- 邹农基, 冯俊文. 客户知识的关键维度及其管理过程. 技术经济, 2006(11)
- 顾金睿, 王芳. 关于本体论的研究综述. 情报科学, 2007(6)
- 赵慧玲. 移动互联网的现状与发展方向探索. 移动通信, 2009(1)
- 邱均平, 文庭孝, 周黎明. 汉语自动分词与内容分析法研究. 情报学, 2005, 24(3):309~317
- 陈霖, 宋燕辉. 数据挖掘技术在电信集团客户管理系统中的应用技术交流. 电信快报, 2008(3)
- 刘永, 陈治平. 数据挖掘在电信套餐预演中的应用研究. 计算机工程与设计, 2008, 29(15)

[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。

(收稿日期:2013-09-20)