# 基于 MapReduce 的校园网用户网购偏好分析①

杨军超, 雒江涛, 申 健, 邓生雄

(重庆邮电大学 电子信息与网络工程研究院, 重庆 400065)

**摘 要**: 用户网购偏好发现是用户挖掘、电商营销以及用户个性化推荐的关键, 该文基于校园网流量, 提出了一种基于 MapReduce 的校园网用户网购偏好分析方法, 结合深度包检测(Deep Packet Inspection, DPI)与网络爬虫等技术, 对校园网用户网购行为进行了特征提取和识别. 以淘宝、天猫、京东三家电商网站为例, 对电商网站用户转化率进行了统计分析, 并分别对三个节假日校园网用户网购偏好进行了细致的分析.

**关键词**: MapReduce; 深度包检测; 校园网; 网购偏好分析.

## Online Shopping Preference Analysis of Campus Network Users Based on MapReduce

YANG Jun-Chao, LUO Jiang-Tao, SHEN Jian, DENG Sheng-Xiong

(Electronic Information and Networking Research Institute, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

**Abstract**: A method for online shopping preference analysis based on MapReduce is proposed in this paper. The campus network traffic is analysed using MapReduce model, in which the features of users online shopping behavior is extracted by four MapReduce jobs combined with deep packet inspection (DPI). Making use of those features occuring in different E-commercial websites and with the help of the product information database established by a web crawler, user online shopping conversion rates of E-commercial websites and category of purchased product are analysed and preference analysis results are presented.

**Key words**: MapReduce; deep packet inspection; campus network; online shopping preference analysis

## 1 引言

近年电子商务繁荣发展, 亚马逊(Amazon)个性推荐取得巨大成功, 引起了电商和企业对用户挖掘的重视, 电子商务数据挖掘成为研究的热点. 淘宝、京东等电商纷纷建立数据挖掘平台, 它们拥有丰富的用户交易数据, 可以做到精准营销. 但显然, 对跨电商的用户挖掘以及区域性用户群体的用户购物偏好分析等方面存在不足, 而对于中小电商而言, 建立数据挖掘平台不仅花费巨大, 而且存在用户数据不足等问题.

文献[1] 分析了大数据环境下的电商用户数据特征, 提出电商用户数据挖掘框架, 并探讨数据挖掘流程和主要的数据挖掘方法, 分别从电商平台, 用户和商家三者角度探讨了电商用户数据挖掘的应用问题. 文献[2]基于 Web 用户在自己本地计算机上的行为日志, 其包含用户在本地机器上浏览的所有网页的记录, 提出面向 Web 用户日志的电子商务竞争对手分析模型, 并结合实际数据对 11 家电子商务网站进行了分析, 但此方法实施性差, 用户隐私存在隐患. 文献[3]提出了深度包检测(DPI)技术在移动数据业务流量识别中的应用前景, 并提出了基于云计算的未来数据服务平台架构. 文献[4]提出一种基于 Hadoop 的网络流量分析系统, 并提出一种直接读取 libpcap 文件的 API, 实现了基于 IP 层和 HTTP 层的流量分析. 文献[5]基于 MapReduce 的网络流量分析方法, 相比传统的流分析工具计算时间提升了 72%. 文献[6]将 DPI 技术应用到移动互联网监测中, 实现对移动互联网流量进行了监控、分析和管理. 文献[7–8]分别对 DPI 技术在流量分析中的应用进行了相关研究. 基于以上研究的方法和

① 基金项目:重庆市应用开发计划(cstc2013yykfA40006); 2013 重庆高校创新团队建设计划(KJTD201312)
收稿时间:2015-02-06;收到修改稿时间:2015-04-02



© 中国科学院软件研究所 http://www.c-s-a.org.cn

思路, 本文提出了一种基于 MapReduce 的校园网用户网购偏好分析方法, 利用深度包检测技术解析校园网数据, 提取用户网购行为特征, 结合网络爬虫, 对校园网用户网购偏好进行细致的分析, 并对淘宝、天猫、京东三家电商网站的用户转化率进行了统计.

基于网络流量的用户网购行为分析方法, 可以为电商营销提供数据支持, 对于流量分析以及大数据挖掘提供一种新的视角和参考.

## 2 基于校园网流量的用户网购偏好发现

### 2.1 基本思路

基于校园网流量的用户购物偏好分析旨在通过解析用户网络流量, 对用户的购物偏好和电商网站用户转化率进行统计分析. 区别于基于 Web 服务器日志的用户行为分析方法[1], 本文的方法基于网络流量, 对用户购物偏好进行了分析, 能全面真实的反应用户偏好.

本文主要关注用户网购行为, 用户通过搜索或者一系列的导航, 找到有意向购买的商品后, 用户点击进入商品页面, 之后用户点击“立即购买”进行结算或者点击“加入购物车”稍后结算, 而用户的一系列行为均在 HTTP 请求内容中体现出来.

以下以在淘宝购物为例进行网络流量抓包分析, 通过对数据包分析, 可以发现, 用户 HTTP 请求内容, 有明显的网购行为标识特征, 如图 1 所示, 商品浏览其请求 URL 中有显著的“item.taobao.com”字段和“id=商品 ID”字段; “立即购买”其请求 URL 中有显著“buy.taobao.com/auction/buy_now.jhtml?” 字段和“item_id_num=商品 ID”字段; “加入购物车”其请求 URL 中有显著的“cart.taobao.com/cart.htm?”字段, 而其对应的 Referer 字段中则有“item.taobao.com/.....&id=商品 ID&……”特征.

图 1 淘宝网商品浏览流量特征

对天猫和京东网购行为抓取数据包分析发现: HTTP 报头的 GET 字段和 Referer 字段可以作为识别用户网购行为的标识, 如表 1 所示, 商品浏览、立即购买、加入购物车这三种用户行为, 都有唯一的特征字符来确定, 并且有对应的商品 ID 来确定用户网购的商品.

表 1 电商网站网购行为特征

| | 商品浏览 | 立即购买 | 加入购物车 |
|---|---|---|---|
| 淘宝网 | http://item.taobao.com/.....&id=商品 ID&..... | http://buy.taobao.com/auction/buy_now.jhtml?.....&item_id_num=商品 ID…… | http://cart.taobao.com/cart.htm?..... Referer:http://item.taobao.com/.....&id=商品 ID&..... |
| 天猫商城 | http://detail.tmall.com/……&id=商品 ID&…… | http://buy.tmall.com/..... Referer:http://detail.tmall.com/……&id=商品 ID&…… | http://fbuy.tmall.com/cart/..... Referer:http://detail.tmall.com/……&id=商品 ID&…… |
| 京东商城 | http://item.jd.com/商品.html | | http://cart.jd.com/cart/addToCart.html?.....&pid=商品 ID&..... |

如何确定商品 ID 与商品的对应关系是用户网购偏好分析的关键, 我们不仅需要知道用户网购行为的动作, 而且需要知道用户网购行为动作的对象, 即商品. 唯一的商品 ID 是与商品对应的关键, 本文中我们采取的方案是利用网络爬虫构建信息库, 信息库主要包括商品 ID、商品类别以及商品描述等. 用户网购行为特征中提取的商品 ID 与信息库建立映射, 从而得出用户网购的商品具体信息. 结合网购行为特征字符和网购对象信息即商品信息, 对用户的网购偏好进行统计分析.

### 2.2 基于校园网流量的用户网购偏好发现过程模型

本文主要基于网络流量进行用户网购偏好发现, 但网络流量数据量大, 对其的分析效率是需要解决的问题, Hadoop 是 Apcache 下的一个开源的分布式架构, 目前已经在各大企业得到广泛应用. 本文利用 Hadoop 的分布式计算模型(MapReduce)对网络流量数据进行解析和处理. 图 2 给出了基于校园网流量的用户网购



© 中国科学院软件研究所 http://www.c-s-a.org.cn

偏好发现过程模型, 其包括 6 个主要过程, 分别概述如下:

(1)用户数据采集: 主要利用数据采集卡采集校园网用户的网络数据, 为后续的用网购偏好发现提供原始数据依据.

(2)用户数据解析: 主要基于 MapReduce 对原始数据进行预处理, 主要包括数据包重组与合成.

(3)网购行为特征提取: 根据电商网站网购行为特征, 利用深度包检测技术对用户数据进行网购行为特征提取.

(4)网购行为识别: 主要依据用户行为特征以及用户行为对象(商品), 结合网络爬虫建立的信息库, 完成对网购行为的识别.

(5)电商网站用户转化率统计: 主要对用户在电商网站网购行为进行统计

(6)用户网购偏好统计与分析: 根据用户网购行为识别的结果, 对用户偏好(主要包括用户网购网站选择分布、商品类别分布等).

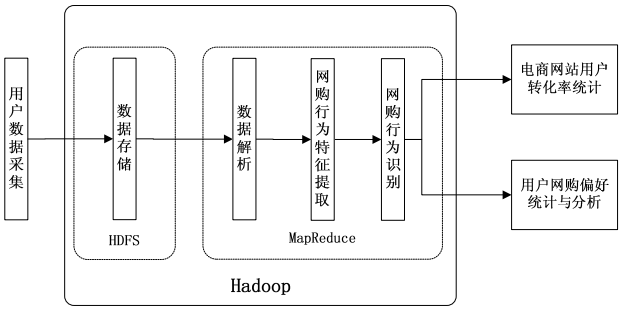


图 2　基于校园网流量的用户网购偏好发现过程模型

## 3　基于MapReduce的深度包检测的实现

### 3.1 MapReduce 与深度包检测

MapReduce[9]的计算过程可以表示为两个功能: map 和 reduce, 其中 Map 功能生成中间键/值对, 并将中间 key 相同的 value 分为一组传递给 reduce. Reduce 功能接受中间 key 和与之关联的所有 value, 融合(merge)这些值, 生成一个更小的值的集合. 图 1 给出了一个包含 2 个 map 实例和 3 个 reduce 实例的数据处理过程.

深度包检测技术[6–7](Deep Packet Inspection)是一种基于应用层的流量检测和控制技术, 深度包检测技术广泛用于数据包应用类型分析、用户行为分析, 以及入侵检测、病毒/蠕虫检测等方面, 是数据挖掘的重要手段. 其中基于流量分析的方法主要包括①传输层端口分析. 许多应用使用默认的传输层端口号, 例如 HTTP 协议使用 80 端口. ②特征字匹配分析. 一些应用在应用层协议头, 或者应用层负荷中的特定位置中包含特征字段, 通过特征字段的识别实现数据包检查、监控和分析③通信交互过程分析. 对多个会话的事务交互过程进行监控分析, 包括包长度、发送的包数目等, 实现对网络业务的检查、监控和分析.

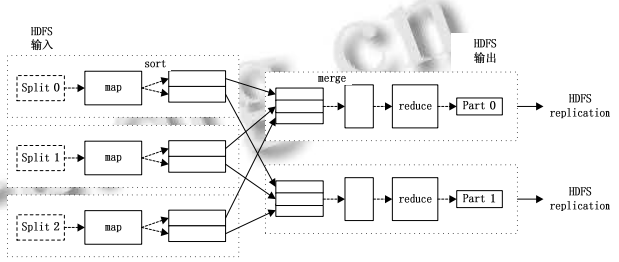


图 3　MapReduce 数据流程图

本文主要采用特征字匹配的深度包检测方法, 利用 1.1 小节的网购行为特征, 基于 MapReduce 对用户数据进行数据处理. 将深度包检测技术融合到 MapReduce 计算模型中.

### 3.2 具体实现

本文通过四个 MapReduce job(作业)实现了基于 MapReduce 的深度包检测, 结合网络爬虫建立的信息库[10–12], 实现了对校园网用户网购偏好的发现, 其具体实现流程如图 4 所示, 具体描述如下:

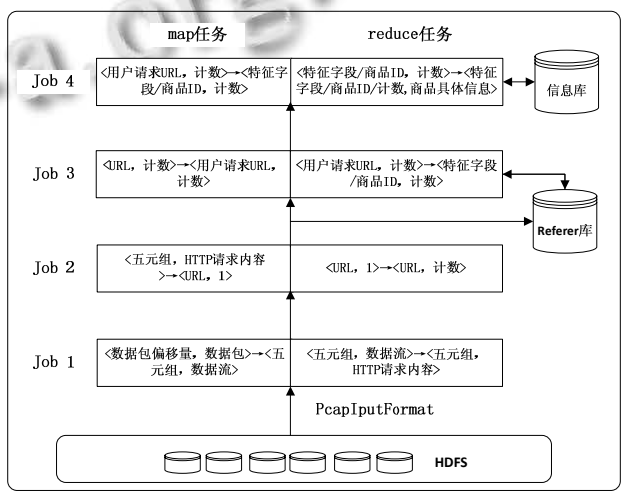


图 4　基于 MapReduce 的深度包检测

(1)job 1　数据包解析和 HTTP 请求重组

通过 PcapIputFormat[4]从 HDFS 中读取数据包, 以



© 中国科学院软件研究所 http://www.c-s-a.org.cn

数据包在文件的偏移量为 key、数据包为 value 作为 map 阶段的输入, map 任务对数据包解码, 以五元组(源 IP、目的 IP、源端口、目的端口、传输协议)进行数据包分流, 并将时间戳、TCP 序列号添加到数据包净荷的前面, 用以 reduce 阶段数据重组, 最终形成以五元组为 key、数据流为 value 作为 map 阶段的输出. reduce 阶段以 map 任务的输出作为输入, reduce 任务根据数据包的时间戳、TCP 序列号对同一五元组的数据流进行 HTTP 请求内容重组, reduce 阶段结果输出以五元组为 key、HTTP 请求内容为 value 的形式.

(2)job 2 URL 提取和次数统计

以 job 1 reduce 输出结果作为 job 2 map 阶段的输入, map 任务提取 HTTP 请求的 Host 字段和 GET 字段, 通过拼接 Host 字段和 GET 字段还原完整的 URL[8]. 同时 map 任务提取 Referer 字段, 将其存入数据库的 Referer 库中. map 阶段以 URL 为 key、数字 1 为 value 作为输出, reduce 阶段以 map 任务的输出作为输入, reduce 任务对提取到的 URL 进行次数统计, reduce 结果输出 URL 为 key、计数为 value 的形式.

(3)job3 用户请求 URL 确认和网购行为特征提取

map 阶段输入即为 job 2 reduce 结果, map 任务将 key 值与 Referer 库中的 Referer 字段匹配, 匹配上则说明此 URL 为用户请求 URL, 否则抛弃, 完成对用户请求 URL 的确认. map 阶段输出结果以用户请求 URL 为 key、计数为 value 的形式. reduce 阶段以 map 任务的输出作为输入, reduce 任务根据用户网购行为特征对用户请求 URL 进行过滤提取特征, reduce 结果输出以网购行为特征字段/商品 ID 为 key、计数为 value 的形式.

(4)job4 网购行为识别与统计

Job4 仅需 map 任务即可完成网购行为识别与统计, 以 job 3 reduce 阶段的结果为输入, map 任务通过 key 值中的商品 ID 与信息库的商品信息进行映射, 得到商品具体信息, 结果输出以网购行为特征字段/商品 ID/计数为 key、商品具体信息为 value 的形式, 完成对用户网购行为的识别与统计.

## 4 实验与分析

(1)电商网站用户转化率分析

以用户在电商网站访问的所有商品页面次数总计即为电商网站用户商品浏览量 PPV(Product Page View), 以用户在电商网站“立即购买”和“加入购物车”的购买行为次数总计即电商网站用户购买量. 用户转化率=用户购买量/用户商品页面浏览量, 用户转化率是衡量用户进入电商网站后一系列浏览后真正购买商品的情况, 它一定程度上体现了用户网购对于电商网站选择的偏好, 本文以校园网五个学生公寓互联网流量进行了分析.

如表 2 所示, 淘宝网在“双十一”、“双十二”、“元旦”用户转化率分别达到了 14.3%、9.1%、6%, 并且其拥有比较高的用户购买量. 天猫商城在这三个时间段的用户转化率只有 4%、3.5%、3.2%, 相对淘宝网来讲较低, 而且仅在“双十一”有较高的购买量. 京东商城在这三个时间段的用户转化率分别达到了 16.9%、17.4、16.7%, 是三家中最高的, 说明到京东购买商品的用户目的性较明确, 但是这三个时间段内的用户商品浏览量和用户购买量比较低, 京东需要在吸引用户进入其网站方面加强营销策略.

表 2 电商网站用户转化率

| 电商网站 | 双十一 | | 双十二 | | 元旦 | |
|---|---|---|---|---|---|---|
| 淘宝网 | 17183 | 14.3% | 11234 | 9.1% | 10327 | 6% |
| | 2554 | | 1037 | | 616 | |
| 天猫商城 | 21714 | 4% | 5193 | 3.5% | 4310 | 3.2% |
| | 809 | | 183 | | 139 | |
| 京东商城 | 540 | 16.9% | 384 | 17.4% | 471 | 16.7% |
| | 91 | | 67 | | 79 | |

(2)用户网购偏好分析

结合网络爬虫建立的三家电商网站的商品信息库, 本文对三个时间段的用户网购行为进行了细致的分析, 由于文章篇幅限制, 仅给出基于频道分类下的统计结果, 分别如图 5、图 6、图 7 所示, 可以总结如下:

淘宝网和天猫商城在“双十一”相对于其他两个时间段的用户购买量要高出一倍以上, 说明淘宝网和天猫商城对于“双十一”的营销策略比较成功. 而京东商城在“双十一”并没有较高的用户购买量, 淘宝网和天猫商城在用户群体中依旧占据较大的优势.

三家电商网站的用户购买量多集中于服饰和运动类, 其中淘宝网和天猫商城服饰类的用户购买量占到了 50%以上. 京东商城除服饰类和运动类外, 数码产品类也较为集中, 说明京东商城在学生用户群里中有较高的信誉度.



© 中国科学院软件研究所 http://www.c-s-a.org.cn

三家电商网站在三个时间段内的用户购买量多集在女性用户上, 可见电商网站的优惠信息对女性更吸引力, 电商网站可以针对女性用户做更深入的营销.

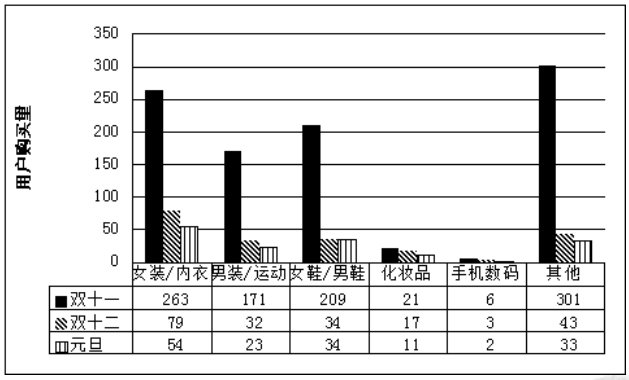


图 5　淘宝网用户网购商品分布

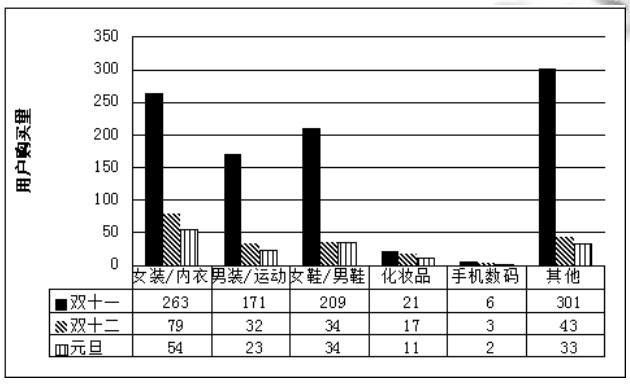


图 6　天猫商城用户网购商品分布

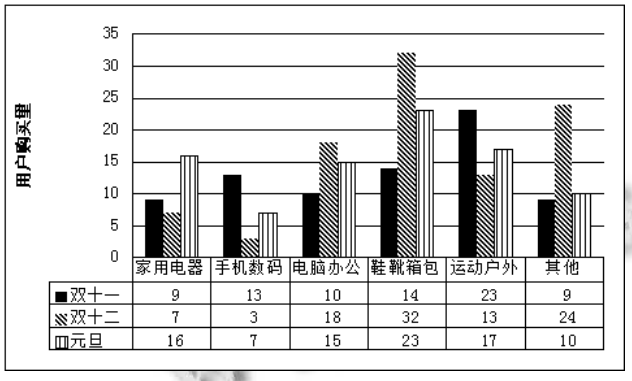


图 7　京东商城用户网购商品分布

## 5　结论

本文提出了一种基于 MapReduce 的校园网用户网购偏好分析的方法, 基于校园网流量对用户网购偏好进行了分析, 本文的主要贡献在于将深度包检测技术应用到流量分析中, 基于 Hadoop 的 MapReduce 计算模型将校园网流量的用户网购行为特征提取出来, 利用网络爬虫建立商品信息库的方案, 对校园网用户的网购行为进行了深入识别. 结果显示该方法也较好的扩展性和可行性. 对于用户群体的网购偏好分析以及潜在用户挖掘都具有巨大的商业价值, 同时也为网络流量的分析提供了新的视角和参考.

## 参考文献

1 徐国虎,孙凌,许芳.基于大数据的线上线下电商用户数据挖掘研究.中南民族大学学报(自然科学版),2013,32(2):100–105.

2 赵洁,温润,周峰,金培权.基于 Web 用户日志的电子商务领域对手分析—以 11 家电子商务网站为例.信息资源管理学报,2013,4:53–62.

3 雒江涛,舒忠玲,梁燕.移动数据业务透视.北京:人民邮电出版社,2014:285–298.

4 Lee Y, Lee Y. Toward scalable internet traffic measurement and analysis with hadood. ACM SIGCOMM Computer Communication Review, 2013, 43(1): 6–13.

5 Lee Y, Kang W, Son H. An internet traffic analysis method with MapReduce. 2010 IEEE/IFIP Network Operations and Management Symposium Workshops. 2010. 357–361.

6 Lu XM, Cao WH, Huang XS, Huang FY, He LW, Yang WH, Wang SB, Zhang XT, Chen HS. A real implementation of DPI in 3G network. 2010 IEEE, Global Telecommunications Conference (GLOBECOM 2010). 2010. 1–5.

7 Trivedi U. A self-learning stateful application identification method for deep packet inspection. 2012 8th International Conference on Computing Technology and Information Management (ICCM). 2012. 416–421.

8 Rezvani M, Ignjatovic A, Bertino E, Sanjay J. Provenance-aware security risk analysis for hosts and network flows. 2014 IEEE Network Operations and Management Symposium (NOMS). 2014. 1–8.

9 Hadoop.http://hadoop.apache.org/.

10 郑力明,易平.基于 HTMLParser 信息提取的网络爬虫设计.微计算机信息,2009,15:123–124.

11 孙立伟,何国辉,吴礼发.网络爬虫技术的研究.电脑知识与技术,2010,15:4112–4115.

12 Arif B, Nisa A.U, Shafi Q, Qureshi HN, Siddiqui UH, Tariq T. Web crawlers to detect security holes. 2013 International Conference on Open Source Systems and Technologies (ICOSST). 2013. 133–140.



© 中国科学院软件研究所 http://www.c-s-a.org.cn