基于 Hadoop 的 Web 用户识别与新闻智能推荐算法研究

林中明[1],李文敬[2,3]

(1.广西师范学院 计算机与信息工程学院;2.科学计算与智能信息处理广西高校重点实验室,广西 南宁 530023;
3.广西师范学院 物流管理与工程学院,广西 南宁 530001)

摘 要:为了解决大数据时代用户阅读时遇到的"信息过载"与"信息迷失"问题,提出了基于 Hadoop 平台的用户准确识别与新闻推荐算法。首先基于 MAC 地址识别用户,通过对用户浏览轨迹的离线和在线挖掘,建立用户兴趣模型。然后对新闻关键词进行聚类,结合协同过滤和启发式方法,基于关键词对用户进行新闻的智能推荐。实验结果表明,基于 MAC 地址的算法比基于 IP 地址的算法用户识别率提高了 30%。

关键词:云计算;新闻推荐;Web 日志挖掘;Hadoop;MAC 地址

DOI:10.11907/rjdk.161378

中图分类号:TP312 **文献标识码**:A **文章编号**:1672-7800(2016)005-0027-03

0 引言

根据 ZDNET《数据中心 2013:硬件重构与软件定义》[1]年度技术报告显示,2013 年中国产生的数据总量超过 0.8ZB,预计到 2020 年,产生的数据总量将是 2013 年的 10 倍。海量的 Web 信息让人们感觉到信息过载和信息迷失,如何快速精准地识别用户并为其推荐感兴趣的内容成为了当今的研究热点[2]。根据新闻阅读与设备使用情况的调查问卷[3]数据显示,95%的人都是在电脑、手机、平板等电子设备上获取新闻资讯,而且 80%的人在阅读新闻时并未处于登录状态,即无法通过用户的登录信息给用户推荐相应内容。面对海量的新闻资讯,文献[4]针对用户识别存在的问题提出了 IASR(IP,Agent,Session and Referrer)算法,通过引入会话(Session)来识别用户;文献[5]提出了基于用户浏览行为的建模,提高了同一个 IP 下用户的识别率;文献[6-8]提出了基于 URL 相似度的会话识别方法。但这些方法并不能改变 IP 对于识别用户的限制,所以不能从本质上提高用户识别率。因此,利用 Hadoop 大数据平台,对无登录信息的用户进行快速身份识别和新闻信息的个性化推荐,相关研究具有重要的现实意义和潜在的经济价值。

1 海量 Web 日志与用户识别

MAC 地址是网卡物理地址,由网络设备制造商生产时写在硬件内部,因此世界上任意一个拥有 48 位 MAC 地址的网卡都有唯一标识[9],且 MAC 地址与网络无关。通过在 Web 日志中加入 MAC 地址,可以实现用户的唯一性识别,增加用户识别的准确性。

用户识别是个性化新闻推荐的基础和关键,详细有用的用户数据将决定新闻推荐的效果。由于 Web 日志中包含了访问主机 IP、访问时间、访问页面、请求方式等信息,详细记录了用户的访问轨迹,生成巨大的数据量及数据类型,因此将通过 Web 日志作为用户识别的数据源。本文将记录分为长期记录和短期记录,一般将 10 天以前的访问日志作为长期记录,最近 10 天的访问日志作为短期记录。针对长期记录,通过 Hadoop 平台进行离线处理。短期记录则在用户使用过程当中,以信息增量的形式补充到推荐算法中来。

2 基于 MAC 地址的用户识别算法

2.1 算法基本思想

Hadoop 的核心是 Map/Reduce。Map/Reduce 是一

基金项目:国家自然科学基金项目(61163012,61363074);广西科学研究与技术开发计划项目(桂科攻 1598010-3);广西高校科学技术研究项目(2013YB147);广西研究生教育创新计划项目(YCSZ2014187)

作者简介:林中明(1988-),男,广西富川人,广西师范学院计算机与信息工程学院硕士研究生,研究方向为云计算;李文敬(1964-),男,广西邕宁人,广西师范学院物流管理与工程学院教授、硕士生导师,研究方向为并行计算、云计算。

个可用于大数据处理的离线计算模型，它将一个任务分成多个细粒度的子任务，并将这些子任务分配到计算节点上进行并行处理，以缩短任务完成时间。将 Web 日志等份划分后，利用 Map/Reduce 对 Web 日志作长期记录处理。

利用 Hadoop 平台得到用户长期记录下的每个 MAC 地址对应用户的集合文件，这是一个庞杂的文件，将通过基于 URL 相似性的用户识别算法对集合文件进行处理，得到此 MAC 对应用户的 100 条最感兴趣页面的排序文件。

定义长期记录的日志文件为集合 $L=\{l_1,l_2,\cdots\cdots,l_m\}$，通过 map 过程得到每个 MAC 对应的集合文件 $K=\{k_1,k_2,\cdots\cdots,k_n\}$，再通过 reduce 过程，得到对应生成的用户长期访问文件为 $MAC=\{MAC_1,MAC_2,\cdots\cdots,MAC_r\}$，每个文件里包含了此 MAC 地址对应用户的所有长期访问记录。在 K 的每个文件中包含有访问时间、IP、URL、访问时长、访问次数字段。针对短期日志文件，根据最近 10 天该 MAC 地址用户的所有访问记录，同样生成一个短期的访问记录文件。在用户进入站点后，根据用户的长期和短期记录生成一个综合的用户访问记录文件，与用户未读新闻对比后进行推荐。

2.2 特征标签选择

由于一篇文章中经常存在多个分页形式，且每个分页的访问次数和浏览时间基本相同，所以要将同属一篇文章多个分页的 URL 记录合并。对 k_i 中 URL 具有相似性的记录进行合并，$cos(URL_i,URL_j)$ 为两条 URL 的余弦相似性，$Smax_i$ 为合并的记录中访问次数最多的，$\bar{t}_i$ 为合并的记录中访问时间的平均值，num_i 为合并的记录条数。

$$\cos(URL_i,URL_j)=\frac{URL_1\cdot URL_2}{\|URL_i\|\cdot\|URL_j\|} \tag{1}$$

如果 $\cos(URL_i,URL_j)$ 等于 1，说明这两条记录同属一篇文章，在此基础上，遍历 k_i 得到：

$$A=\begin{bmatrix}1 & \cdots & \cdots & \cdots & \cos(k_n,k_1)\\ \cos(k_1,k_2) & 1 & \cdots & \cdots & \vdots\\ \cos(k_1,k_3) & \cos(k_2,k_3) & \cdots & \cdots & \vdots\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \cos(k_1,k_n) & \cos(k_2,k_n) & \cdots & \cdots & 1\end{bmatrix} \tag{2}$$

A 是一个对称矩阵，继续降维，得到：

$$A_{num_1}=num_1\begin{bmatrix}\cos(k_2,k_1) & \cdots & \cdots & \cos(k_n,k_1)\\ 1 & \cdots & \cdots & \cdots\\ \cos(k_2,k_3) & 1 & \cdots & \cdots\\ \vdots & \vdots & \vdots & \vdots\\ \cos(k_2,k_{n-numi}) & \cdots & \cdots & 1\end{bmatrix} \tag{3}$$

A_{num_1} 是合并了第一篇文章之后的矩阵，num_1 为第一篇文章合并的记录条数。继续计算矩阵中第一列合并的所有项 $\{k_1,k_2\cdots\cdots k_{num_1}\}$ 中对应的时间进行求和，平均值 $\bar{t}_i$ 为：

$$\bar{t}_i=\frac{\sum_{i=1}^{num_1}(k_1(t_1),k_2(t_2),\cdots\cdots,k_{num_1}(t_{num_1}))}{num_1} \tag{4}$$

2.3 权重计算与排序

基于改进的归并排序，引入最大访问次数 $Smax_i$、合并记录中访问时间的平均值 $\bar{t}_i$ 与合并条数 num_i 作为参数，它们的权值分别为 x、y、z，则每篇文章的对应权值为：

$$Q_i=Smax_i\cdot x+\bar{t}_i\cdot y+num_i\cdot z \tag{5}$$

最后，根据权值从 MAC_i 中得到此 MAC 用户的 100 条最感兴趣的记录文件。

3 基于关键词的协同过滤智能推荐算法

当前有很多种智能推荐算法，主要有基于内容的推荐、协同过滤推荐和基于知识的推荐。基于内容的推荐是提取对象中的特征属性，通过用户信息与待推荐对象的匹配程度进行推荐，但这种算法对特征提取方法的依赖程度很高，无法准确地描述用户特征；协同过滤推荐是通过聚合待推荐用户的相似用户评价的所有对象，计算对象与用户之间的效用值进行推荐，对于新对象和新用户都存在冷启动和稀疏性问题；基于知识的推荐是在特定领域构建规则来进行基于规则和实例的推理，不存在冷启动和稀疏问题，但知识很难建模。

本文结合各推荐算法的优缺点，提出一种基于关键词的协同过滤智能推荐算法。一般地，在系统中的每一篇文章都包含有最能体现这篇文章主题的关键词。通过对关键词的聚类，避免了项目的冷启动问题，并去掉了项目特征提取的步骤。对从用户模型中得到的此 MAC 用户的 100 条最感兴趣的记录文件，对关键词进行聚类。得到关键词聚合文件 $W=\{(w_1,q_1),(w_2,q_2),\cdots\cdots,(w_n,q_n)\}$，其中 q 为 w 的出现次数。利用启发式方法，先计算文章关键词之间的相似度，再对所有待推荐文章对此 MAC 用户的效用值进行计算，得到推荐子集。同时假设待推荐文章的关键词为 $W'=\{w_1',w_2',\cdots\cdots,w_m'\}$。

该算法以用户的关键词聚合文件 W、待推荐文章的关键词文件 W'、已访问过的文章集合文件 MAC 作为输入。判断要推荐的文章用户是否访问过，如果没有，则计算两篇文章关键词的相似性：

$$\cos(w_i,w')=\sum_{i=1}^{n}\sum_{j=1}^{m}w_iw'_j/\sqrt{\sum_{i=1}^{n}w_i^2}\sqrt{\sum_{j=1}^{m}w_j'^2} \tag{6}$$

计算待推荐文章 P 对用户的效用值：

$$r_{p(i)}=\sum_{i=1}^{n}cos(W_i,W')\times q_i \tag{7}$$

对待推荐文章的效用值进行排序，并得到从大到小排序的推荐子集为：

$$R_p=\text{argmax}r_p \tag{8}$$

具体算法流程如下：

Input W，W'，MAC

if $URL\notin MAC$ //该 URL 不在用户已访问的记录中

DO

$\cos(w_i,w')$ //计算两文章的关键词相似性

$r_{p(i)} = \sum_{i=1}^{n} cos(W_i, W') \times q_i$　//计算文章 p 的效用值

R_p = argmaxr_p　//得到推荐子集

找到 R_p 满足

if △ MAC (mac) ≠ 0　//已访问过新的文件

重复①－⑦步骤

END

4　实验结果与分析

实验在由 5 台 HP DL380G5 服务器组成的集群上进行，其中，一台是主节点，一台是任务调度节点，5 台都可以作为计算节点及数据存储节点。同时，采取 Xen 的虚拟化技术，使同一节点上同时并发执行多个 MapReduce 操作。5 台服务器均安装 hadoop－0.20.0 和 JDK。实验程序是在 PHP 集成开发环境中开发的。测试数据集来自某地方综合新闻资讯网站的 Web 服务器日志。为了验证该 Web 日志分析平台的有效性及高效性，做了以下 2 个实验。

实验 1：在 Hadoop 平台上对 Web 日志中的 MAC 和 IP 地址数量分别进行统计。通过比较发现，基于 MAC 地址比基于 IP 地址辨别用户的算法识别率高出了 30％以上，且随着记录时间的变长，用户的识别率还会继续扩大。这表明基于 Web 日志分析的新闻推荐使用基于 MAC 地址的用户识别算法能够准确地识别用户，且不依靠用户前台的数据，减轻了前台数据的处理压力。

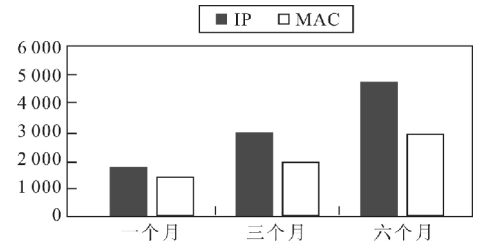

图 1　对比 IP 和 MAC 的用户识别率

实验 2：在实验 1 的两个平台上对大小不同的 4 个 Web 日志文件分别进行处理，计算执行时间，结果如图 2 所示。

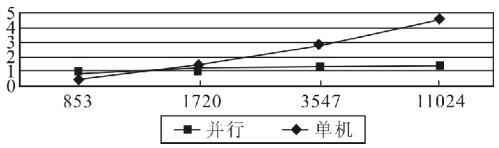

图 2　单机系统与并行系统执行时间对比

同时，本文分别在计算节点数为 2、3、5 时对一个 Web 日志进行了分析，计算执行时间，其结果如图 3 所示。

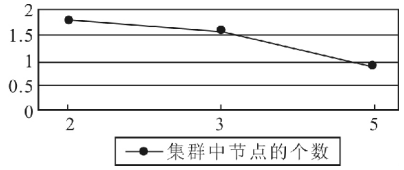

图 3　集群中不同节点个数对执行时间的影响

从以上结果可以看出，利用 MAC 地址的唯一性来识别用户是一个切实可行的方法。当处理的数据量较小时，基于 Hadoop 的 Web 日志分析平台由于需要生成及传输中间文件和最终文件，开启 Hadoop 也需要一定时间，因此并行运算的总时间反而大于单机执行时间。但随着数据量增大，基于 Hadoop 的并行处理平台将数据分割后分派给多个节点并行处理，使并行运算的总时间小于单机执行时间，且随着输入数据的增加，两者执行效率的差距也越来越大。从图 3 可以看出，集群中拥有的节点数目越多，基于 Hadoop 的并行处理平台效率越高。

5　结语

针对目前用户阅读新闻普遍遇到的信息过载问题及用户不登陆浏览的阅读习惯，基于 MAC 的用户识别算法提高了新闻推荐中的用户识别率。同时针对运行于单机集中平台上的 Web 日志分析系统不能满足海量数据处理的问题，本文在对云计算的 Hadoop 集群框架研究的基础上，给出了一种基于 Hadoop 集群框架的 Web 日志分析方法。实验结果表明，该平台能够获取隐含的、有实用价值的信息，执行效率高。

参考文献：

[1] 张广彬，盘骏，曾智强. 数据中心 2013：硬件重构与软件定义[R]. ZDNet 企业解决方案中心，2013.

[2] 张诚，郭毅. 数据挖掘与云计算——专访中国科学院计算技术研究所何清博士[J]. 数字通信，2011(3)：5-7.

[3] 新闻阅读与设备使用情况的调查问卷[EB/OL]. http://www.lzm07.com/index.php？file＝v.html.

[4] 吴永辉，王晓龙，丁宇新，等. 基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J]. 电子学报，2010(11)：2620-2624.

[5] 何希真. 基于用户反馈信息的新闻推荐系统设计与实现[D]. 济南：山东师范大学，2015.

[6] 谢润泉. 基于隐式专家的个性化新闻推荐[D]. 厦门：厦门大学，2014.

[7] 宋科. Hadoop 平台下基于 LDA 的新闻推荐算法研究[D]. 成都：西南石油大学，2015.

[8] 周松松，马建红. 基于 URL 相似度的会话识别方法[J]. 计算机系统应用，2014(12)：191-196.

[9] 谢俐，何勇，杨乐. 网卡 MAC 地址探究[J]. 今日科苑，2008(4)：190.

（责任编辑：黄　健）