基于互联网日志的用户行为分析算法和用户模型的研究

孙卓豪

南开大学 & 电子信息科学与技术

2016 年 3 月 24 日

题解

各模块分别介绍

- 基础数据采集
- 数据预处理
- 主成分分析
- 基于 Hadoop 的协同过滤算法

系统整体介绍及实例分析

已完成工作、未来计划及目前未解决问题

- 已完成工作与未来计划
- 目前未解决问题

题解

- 基于互联网日志，即数据的来源是网络运营商路由器或者交换机流量记录，以及大型互联网公司服务器访问记录
- 用户行为和用户模型，表现为协同过滤算法中的 User-Item 矩阵

综上，本文的主要工作是从路由器、交换机或者服务器采集基础数据，经过预处理后，用协同过滤算法获得 User-Item 矩阵，再根据 User-Item 矩阵进行预测用户行为或者推荐物品。最后将本文工作整合起来设计一个推荐系统。

各模块分别介绍

基础数据采集

本模块主要工作是从网络运营商路由器或者交换机流量记录，以及大型互联网公司服务器访问记录中截取有用的用户行为信息：用户的 ID，请求的 URL，请求时间，请求地点，使用设备类型，服务器返回内容等。

各模块分别介绍

数据预处理

本模块主要工作是对上一模块收集到数据中“服务器返回内容”进行语义分析。语义分析得出的关键词与已有的关键词列表对比，更新关键词列表，同时生成用户兴趣信息：用户的 ID，用户感兴趣的内容，行为时间，行为地点，使用设备类型。再将用户兴趣信息输入分布式系统，已初步构建 User-Item 矩阵。

各模块分别介绍

主成分分析

由于用户数目巨大及用户感兴趣方面较多，上一模块 User-Item 矩阵维数高而且稀疏，本模块主要工作是对矩阵进行降维，减小后一模块计算复杂度。降维主要使用主成分分析算法，通过合并类似的维度减小矩阵的维度。

各模块分别介绍

基于 Hadoop 的协同过滤算法

对上一模块生成的 User-Item 矩阵使用基于 Hadoop 的协同过滤算法。此时 User-Item 矩阵中每个用户对每种物品的喜好程度都已知，只需对每个用户对每种物品的喜好程度排序，则可以获得对每个用户推荐的物品列表。至此用户兴趣分析已完成。

系统整体介绍及实例分析

这一部分主要是在实际环境中采集数据，切分训练集与验证集。用训练集数据输入各模块运行算法，获得 User-Item 矩阵。将结果与验证集比较，判断推荐系统准确程度。同时可以调节推荐系统各参数及各模块间协作方式以获得推荐系统最佳运行配置。

已完成工作、未来计划及目前未解决问题

已完成工作与未来计划

目前已完成资料查找工作，已完成 25% 的写作工作，即已完成各模块分别介绍中的数据预处理部分。

计划在中期检查前完成各模块分别介绍，中期检查后完成系统整体介绍及实例分析。

已完成工作、未来计划及目前未解决问题

目前未解决问题

- 原本计划申请雅虎某 10TB 的数据，结果申请后雅虎客服未回复。即目前未找到大量数据进行模拟
- 如果使用亚马逊云服务，集群每个机器安运行小时收费，小时费率从 0.011USD/小时到 0.27USD/小时不等。数据存储服务单独计费，如在 S3 中存储 10TB 数据，每月需要支付 668USD。价格十分昂贵。其他云服务公司价格类似
- 若购买硬件，自行配置 2 ～ 3 台服务器的集群，价格也不便宜，同时安装配置过程比较麻烦

所以现在不知道应该如何进行实例分析。