主流视频网站评论分析系统

OLAP 系统搭建

沈祥壮

中山大学

2019 年 10 月 29 日

项目概览

- 项目的核心是对主流的一些视频网站的评论系统进行深入的分析，力图增强对用户反馈的理解，以便更好的完善社区的服务。
- 本次项目以国内最大的视频网站 bilibili 为研究对象，手动抓取网站的评论数据，并将数据导入数据仓库，建立 OLAP 系统。
- 此外，整个项目前期主要是以探索性数据分析为主，后面会陆续将新的数据融合到数据仓库，并对数据进行深入分析。

OLAP 概览

本次 OLAP 系统的搭建过程中，使用的数据源来自之前自己写的 B 站视频评论的爬虫，即 Bilibili-Comments-Spider 。OLAP 系统的搭建采用了 Python 的 Cubes 框架，可视化部分采用 Cubesviewer 。在使用过程中，对框架本身的执行效率进行了优化，主要是数据库插入数据的优化，极大提升了程序的运行效率。

数据集描述

本次数据集 (*BilibiliGaomu.csv*) 来自 B 站其中一部番剧的评论数据，共包含 *mid*, *username*, *rpid*, *gender*,*content*, *ctime*, *likes*, *rcount* 八个字段，共 77474 条评论数据。因为框架的具有十分良好的可扩充性，经过简单的预处理 (*Preprocessing.py*)，可以十分方便地将之前爬虫采集的数据集成进去。处理后的数据集 (*data.csv*) 包含如下字段。

字段	含义
sex	评论人性别
$username$	评论人昵称
$likes$	评论获赞数
$wordnum$	评论字数
$month$	评论月份

OLAP 系统搭建流程

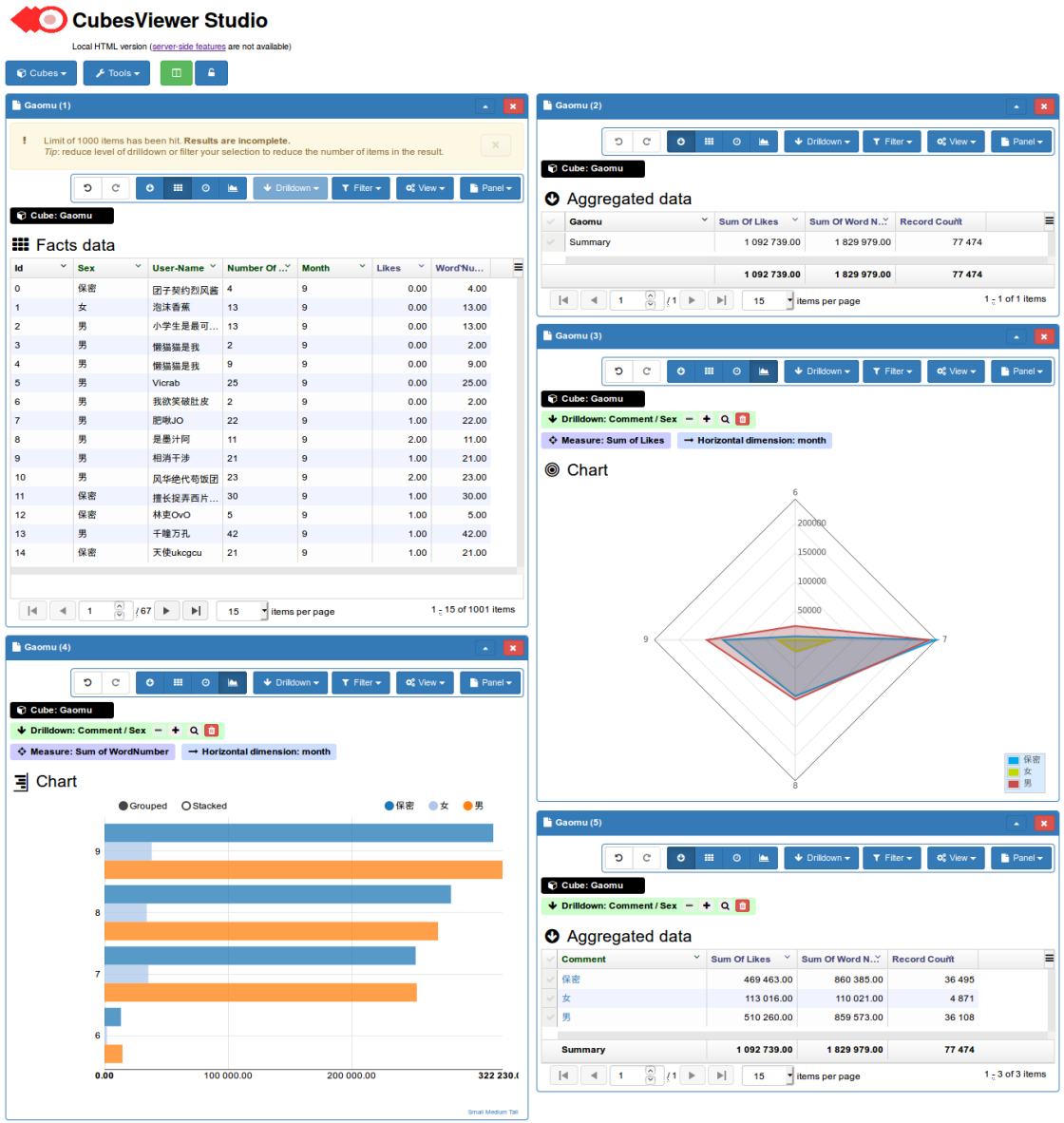
1. 将抓取的数据导入数据仓库
2. 构建数据模型
3. 配置本地 OLAP 服务器
4. 配置可视化服务
5. 将服务部署到远程服务器
6. 登录到 OLAP 系统
7. 开始进行联机分析

登录 OLAP 系统

目前已经将项目部署到服务器之上，登录搭建好的 OLAP 系统流程只需要如下三个步骤：

1. 使用浏览器打开文件
   bilibili/cubesviewer/html/studio.html
2. 在弹出的对话框中输入服务器地址
   http://119.23.209.92:5009/
3. 开始进行联机分析

可视化结果展示


CubesViewer Studio
Local HTML version (server-side features are not available)
Cubes
Tools
Gaomu (1)
Limit of 1000 items has been hit. Results are incomplete.
Tip: reduce level of drilldown or filter your selection to reduce the number of items in the result.
Drilldown
Filter
View
Panel
Cube: Gaomu
Facts data
Id
Sex
User-Name
Number Of ..
Month
Likes
Word Nu...
0 保密 团子契约烈风酱 4 9 0.00 4.00
1 女 泡沫香蕉 13 9 0.00 13.00
2 男 小学生是最可... 13 9 0.00 13.00
3 男 懒猫猫是我 2 9 0.00 2.00
4 男 懒猫猫是我 9 9 0.00 9.00
5 男 Vicrab 25 9 0.00 25.00
6 男 我欲笑破肚皮 2 9 0.00 2.00
7 男 肥喇JO 22 9 1.00 22.00
8 男 是墨汁呵 11 9 2.00 11.00
9 男 相消干涉 21 9 1.00 21.00
10 男 风华绝代苟饭团 23 9 2.00 23.00
11 保密 擅长捉弄西片... 30 9 1.00 30.00
12 保密 林吏OvO 5 9 1.00 5.00
13 男 千瞳万孔 42 9 1.00 42.00
14 保密 天使ukcgcu 21 9 1.00 21.00
items per page
1 - 15 of 1001 items
Gaomu (2)
Aggregated data
Gaomu
Sum Of Likes
Sum Of Word N..
Record Count
Summary 1 092 739.00 1 829 979.00 77 474
1 092 739.00 1 829 979.00 77 474
1 - 1 of 1 items
Gaomu (3)
Drilldown: Comment / Sex
Measure: Sum of Likes
Horizontal dimension: month
Chart
Gaomu (4)
Measure: Sum of WordNumber
Grouped
Stacked
保密
女
男
0.00
100 000.00
200 000.00
Gaomu (5)
Comment
保密 469 463.00 860 385.00 36 495
女 113 016.00 110 021.00 4 871
男 510 260.00 859 573.00 36 108
Summary 1 092 739.00 1 829 979.00 77 474
1 - 3 of 3 items