百视通 Bestv 数据格式说明

Bestv 数据的日志 log

Bestv 数据的日志 log 包括 4 个部分，原始数据存放在服务器（166.111.138.102）：

a. 2013 年 1 月 1 日——2013 年 5 月 13 日，百视通服务器日志 log
   原始 log 存放路径：目录/mnt/2t/Bestv_cdn/10.49.8.11
   原始压缩文件为：/media/Seagate Expansion Drive/BesTV_log/VoD_2013/10.49.8.11.tgz
   后期数据处理目录$Dir：/mnt/2t/Bestv_cdn
b. 2013 年 9 月，包括来自三个 cdn 的日志 log：
   a) 2013 年 9 月 10 日——2013 年 9 月 30 日，百视通服务器日志 log
   原始 log 存放路径：目录/mnt/2t/Bestv_cdn2/third/bestv
   b) 2013 年 9 月 12 日——2013 年 9 月 30 日，蓝汛服务器日志 log
   原始 log 存放路径：目录/mnt/2t/Bestv_cdn2/third/lx
   c) 2013 年 9 月 3 日——2013 年 9 月 30 日，网速服务器日志 log
   原始 log 存放路径：目录/mnt/2t/Bestv_cdn2/third/ws
   三个 cdn 日志的原始压缩文件存放于：/media/Seagate Expansion Drive/BesTV_log/VoD_2013/third
   后期数据处理目录$Dir：/mnt/2t/Bestv_cdn2
c. 2014 年 1 月——4 月百视通播放记录，包括直播和点播日志 log：
   a) 2014 年 1 月 1 日——2014 年 4 月 19 日，百视通点播日志 log
   原始 log 存放于外置硬盘中，存放路径：/media/Seagate Expansion Drive/BesTV_log/VoD
   后期数据处理目录$Dir：/mnt/2t/2014
   b) 2014 年 1 月 1 日——2014 年 4 月 21 日，百视通直播日志 log
   原始 log 存放于外置硬盘中，存放路径：/media/Seagate Expansion Drive/BesTV_log/Live
   后期数据处理目录$Dir：/mnt/2t/2014_live

Bestv 数据视频信息：

原始记录：
目录 /media/Seagate Expansion Drive/BesTV_log/VoD_2013/document
原始视频信息记录存在一定的错误率，后期在使用时进行了一定的处理，处理格式话的视频信息存在在各部分数据对应处理目录中：一般包括$Dir/info/session_video 和$dir/info/session_video_fix，其中 session_video_fix 用于更早 session_video 错误的视频信息。
处理后的信息包括：video_code、video_name、video_length、video_type、other。

Bestv 数据用户设备情况：

使用 client 参数标识，如下：
PC TV Pad Phone
```
Common.client = [
    [0, "N/A"],
    [1, "iPad"],
    [2, "iPhone"],
    [3, "lePad"],
    [4, "Android Phone"],
    [5, "Android Pad 7\""],
    [6, "Android Pad 10\""],
    [7, "PC"],
    [8, "Full Stream"],
    [9, "Android TV"],
    [10, "IPTV Pocket"],
    [11, "TV Cloud"],
    [101, "Ctv"],
    [102, "Ott Pad"],
    [103, "Ott Stb"],
    [127, "Tbd"],
    [-1, "Test"]
];
```

Bestv 视频数据格式化处理：

百视通的视频服务分为非 dash 和 dash 两种，dash 通过苹果的 HLS 来实现，主要给苹果设备提供视频服务，在 dash 的 log 数据中，大部分有用户 id 标识符。Dash 部分的 log 根据日记的记录情况可以分为：带 user_id 的记录和不带 user_id 的记录。带 user_id 的记录在开始会请求 index.m3u8 文件，之后请求 ts 文件。不带 user_id 的记录在最开始的请求的的 m3u8 文件分为 index.m3u8 和非 index.m3u8（例如 HLSVodService.m3u8），在数据格式上有所区别。
目前对 bestv 各部分数据的处理如下：

1. 统计所有 log 对应的用户访问设备（user agent）统一编号，使用增量更新的方式进行更新。第 a 部分数据设备的统计列表为：文件 /mnt/2t/Bestv_cdn/device/total_server，第 b 部分的设备统计列表为：文件 /mnt/2t/Bestv_cdn2/device/total_device，第 c 部分的设备统计列表为：点播 VoD，/mnt/2t/2014/device/all_device；直播 Live，/mnt/2t/2014_live/device/all_device。
   此信息不一定准确，后使用 client_code 来标识用户设备，目前只在判断是否同一个 session 时使用。
2. 格式化日志 log 数据。

提取其中分析需要的数据，包括用户 ip、用户设备 id、视频 code（用于对应视频信息）、视频文件下载时间、视频码率、视频文件名称、http_code、下载数据量。同时对于 dashbu 的记录，根据用户 ip、视频 code 以及用户设备来判断是否是相同一次观看记录（session），把同一个 session 的记录放在一起。（user_id 为 000000000000000，表示没有 user_id；disk_id 为-1，表示没有 disk_id）。

对应代码为：$Dir/code/session/{Get_seesion.java, Get_session.sh}

a) 第 a 部分非 dash 数据的格式化：目录/mnt/2t/Bestv_cdn/no_dash/，格式为：user_ip、video_path、device_id、video_code、disk_id、download_time、timestamp、video_bitrate、file_name、http_code、download_size。

   第 a 部分 dash 数据的格式化：目录/mnt/2t/Bestv_cdn/session_format/，格式为：user_id、user_ip、video_path、device_id、video_code、disk_id、download_time、timestamp、video_bitrate、file_name、http_code、download_size、client_code。

b) 第 b(1) 部分 dash 数据的格式化：目录/mnt/2t/Bestv_cdn2/session_format/，格式为：user_id、user_ip、video_path、device_id、video_code、disk_id、download_time、timestamp、video_bitrate、file_name、http_code、download_size、client_code。

   第 b(2) 部分 dash 数据的格式化：文件/mnt/2t/Bestv_cdn2/session_lx/format，格式为：user_id、user_ip、video_code、video_code、download_time、timestamp、video_bitrate、file_name、http_code、download_size。（蓝汛没有记录 download_time，均为 0.0）

   第 b(3) 部分 dash 数据格式化：文件/mnt/2t/Bestv_cdn2/session_ws/format，格式为：user_id、user_ip、video_code、video_code、download_time、timestamp、video_bitrate、file_name、http_code、download_size。

c) 第 c 部分数据格式化：目前已经完成处理了 2014 年 4 月 1 日—2014 年 4 月 15 日半个月的点播和直播数据。点播 VoD：/mnt/2t/2014/session_format，格式为：user_id、user_ip、video_path、device_id、video_code、disk_id、download_time、timestamp、video_bitrate、file_name、http_code、download_size、client_code。直播 Live：/mnt/2t/2014_live/session_format，格式为：user_id、user_ip、video_path、device_id、video_code、disk_id、download_time、timestamp、video_bitrate、file_name、http_code、download_size、client_code。

3. 整理了从 video_code 到视频信息的对应关系，使用增量更新原则，存放于各部分数据对应下： $Dir/info/session_video 和$dir/info/session_video_fix。
4. session 记录提取。

   对于 dash 的每一次观看记录，进一步进行了处理，统计了播放视频分片的数量，提取了其中码率切换的次数，计算了视频播放的百分比等。

   第 a 部分数据的信息存放于/mnt/2t/Bestv_cdn/session/sessions 文件。

   第 b(1)部分数据的信息存放于/mnt/2t/Bestv_cdn2/session/sessions 文件。

   第 b(2)部分数据的信息存放于/mnt/2t/Bestv_cdn2/session/sessions_lx 文件。

第 b(3)部分数据的信息存放于/mnt/2t/Bestv_cdn2/session/sessions_ws 文件。

第 c 部分数据的信息存放于点播/mnt/2t/2014/session/*和直播/mnt/2t/2014_live/session/*

其中数据对应的格式为：server_name、user_id、user_ip、video_path、disk_id、video_code、start_time(20130910 010513)、end_time、start_line、end_line、start_in_file、end_in_file、start_in_file_f、end_in_file_f、log_num、s_num、s1_num、s2_num、s3_num、s4_num、m3u8_num、total_time、total_size、average_speed、switch_num。

5. 码率切换分析。

对码率切换进行了分析，对有视频信息的观看记录，统计了其观看过程中发生码率切换的情况，包括切换类型、切换时间点、切换次数等，其中每一行记录表示播放过程中发生的一次码率切换行为。

对应代码：$Dir/code/session/Bitrate_types.java

第 a 部分数据信息：/mnt/2t/Bestv_cdn/session/bitrate_types/summary

第 b(1)部分数据信息：/mnt/2t/Bestv_cdn2/session/bitrate_types/summay

第 c 部分数据信息：点播/mnt/2t/2014/bitrate_types/summary

每一行对应信息表示：server, userid, ip, video_path, user_agent, video_code client_code, device, starttime, endtime, video_name, video_type, video_length, play_chunks, play_percent, if>80%, bs_num, bs_type, next_bs_type, bs_chunk_time, bs_chunk_percent, bs_timestamp, bs_chunk_file, continue_chunk_num, bs_change_value, from_bitrate_back, from_bitrate to_bitrate, next_bitrate, bs_status。

6. 用户偏好训练。

使用用户观看视频的记录对用户在视频类别上的偏好进行了训练。进行训练时针对观看记录大于 20 次的用户。

第 a 部分数据：/mnt/2t/Bestv_cdn/session/bitrate_types/interest2/

第 b(1)部分数据：/mnt/2t/Bestv_cdn2/session/bitrate_types/interest/

第 c 部分数据：/mnt/2t/2014/bitrate_types/interest/

对应代码：各自目录下 part.py，cal.py，count.py。

得到的结果为用户在各视频类别下的偏好情况，目录下 user_inter_format 文件。

7. 用户体验预测。

利用用户观看记录对用户体验进行建模预测，最后使用了随机森林的方式，先后使用 R 和 MATLAB 来训练预测。

第 a 部分数据：/mnt/2t/Bestv/session/bitrate_types/train/

第 c 部分数据：/mnt/2t/2014 /bitrate_types/train/

其中目录下 py 目录中存放了相关数据处理的代码，dat 目录下存放了处理得到的用于模型训练和预测的数据集。

使用 MATLAB 进行随机森林训练预测的代码存放于/mnt/2t/workspace/matlab 目录。