EMC 杯数据创新大赛数据集说明(网络数据集)

1　数据集概况

网络数据集由上海交大 WiFi 网络用户的上网流量统计产生，该 WiFi 网络覆盖交大主要校区，WiFi 热点涵盖了教室、宿舍、公共活动建筑、以及部分室外开阔场地，从一个别样的角度记录了大学校园的生动生活。数据集由交大网络信息中心和 OMNILab 联合提供，在隐私处理的基础上，将用户的上网统计信息以比赛数据集的形式进行公开，参赛队伍在通过授权后方可使用该数据集参加本次技术创新大赛。

该网络数据集由原始的用户 HTTP 上网记录统计产生，记录了 WiFi 用户在 2014-09 至 2015-01 五个月间的上网习惯及时空轨迹信息。这里对首先对数据集涉及到的基本概念进行说明：

- **无线网络**：这里特指交大的 WiFi 无线网络，由数千个独立的 WiFi 热点（也称 Access Points, AP）和中心网络控制器组成。用户手持设备通过连接 AP、登陆后使用互联网服务。该网络提供了用户的识别信息和空间位置数据。
- **网络流量**：网络流量是用户使用网络上网时产生的数据流，该数据流包含了用户的上网时间和应用类型。通过对用户的通信过程进行会话（session）分割，得到了用户会话粒度的上网时序数据。
- **网络会话**：指一段时间内用户连续的上网行为。在该数据集中，网络会话以超时的方式进行定义，即如果用户的在一定时间段内没有上网，接下来的网络通信被认为是一个新会话的开始。这里我们采用 **5 分钟**为会话分割的阈值。
- **统计特征**：这里指针对单个用户的单个会话，提取的关于该会话的网络流量和服务使用的统计数据。

2　数据集字段说明

该数据集通过严格的匿名化处理，去除用户的身份识别信息和上网轨迹，仅保留时空行为的统计信息，为比赛提供了丰富信息的同时很好地保护了用户的隐私。

网络流量比赛数据集包含两个基本的数据表: net_traffic.dat 和 net_users.dat，即网络数据表和用户特征表。

- **网络数据表**是数据集的主体，以单次网络会话为最小时间粒度，记录了网络会话的统计特征（以英文逗号 ’,’ 分隔），具体包括：

用户 ID：long，如 1000

上网地点：string，如东上院

会话开始时间(UNIX 时间)：long，单位毫秒，如 1412229603742

会话持续时间：long，单位毫秒，如 360000

服务提供商：string，如腾讯微信

服务类型：string，如即时通信

服务一级域名：string，如 qq.com

通信字节数：long，单位 Byte，如 11656

发送的 HTTP 请求数：long，如 4

本条记录再数据集中表示成：

```
1000,东上院,1412229603742,360000,腾讯微信,即时通讯,qq.com,11656,4
```

- **用户特征表**记录了用户的身份特征信息，包括：

  用户 ID：long，如 1000

  性别：boolean，如 0 或 1

  生日：int，如 1993

  年级：int，即入学年，如 2014

两份数据表中的用户 ID 是关联的，即同一 ID 表示同一个用户。详细数据格式说明请参考 net_traffic.schema 和 net_users.schema 文件。

*UPDATE：为了在用户隐私和数据质量之间达到平衡，比赛发布的最终数据集对服务提供商、服务类型、通信字节数、以及 HTTP 请求数进行了扩展：如果在一次会话中间用户使用了多个应用程序，对应的特征按**流量大小**顺序显示在各个字段，每个字段用**英文分号**（;）分隔。*

3 数据集基本统计特征

本次比赛发布的网络数据集总共包含大约 20000 个有效匿名用户的时空、网络行为信息，数据条目 1200W 条，采用 CSV 格式存储，字段内的不同特征值使用分号分隔。