基于用户行为的数据分析与挖掘

房如华 2011.09.10

SACC2011

关于我

- 房如华，酷讯旅游网 BI部门

- 联系方式
  - @房如华bluetent
  - bluetent@gmail.com

酷讯旅游网的BI团队

- 老公司的新部门
- 两个使命：
  - 产品运营工作的“推进器”
  - 让网站变得更“聪明”

小调查

- 有多少公司在使用自行开发的统计系统？
- 有多少公司已经开展了数据挖掘算法方面的实践？

用户行为分析是道哲学题：

- “你是谁？”
- “你从哪里来？”
- “你要到哪里去？”

“你是谁？”

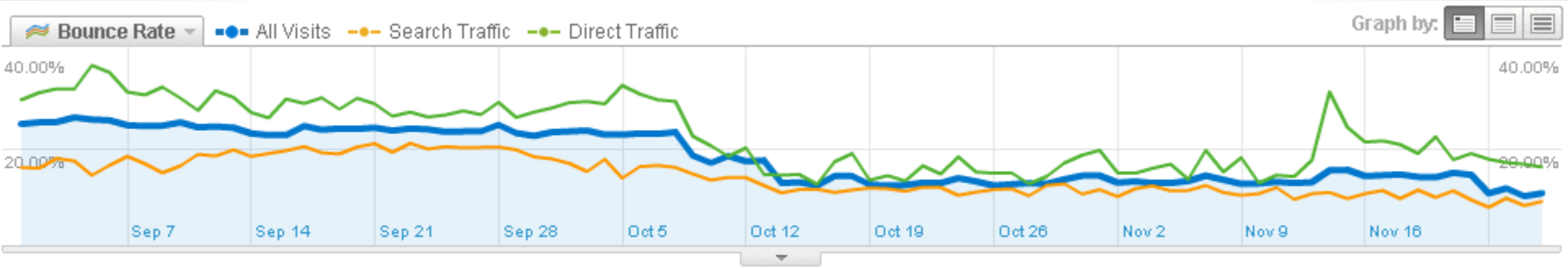
- 如何识别一个用户？
  - 按惯例，我们使用浏览器的cookie区分不同的用户
  - 推荐使用Guid算法进行生成用户的唯一ID
- 如何识别一次访问？
  - 生成访问的唯一ID，并使用cookie记录
  - 在cookie中记录会话的最后更新时间，超过N（如30）分钟则认为会话结束

“你从哪里来？”(1)

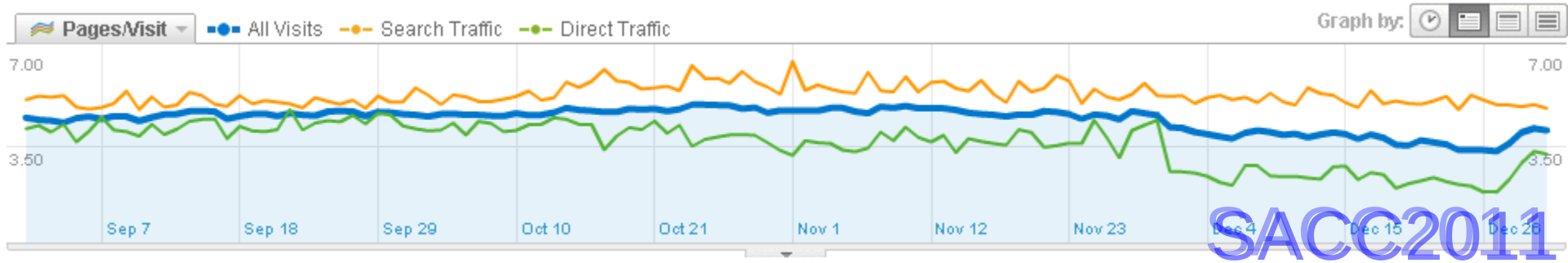
- 用户的流量来源有多种划分
  - 免费流量，付费流量
  - 不同的来路网站
    - 直接打开网址
    - SEO/SEM
    - 社会化网站
    - 付费广告

“你从哪里来？”(2)

- 为什么要关注流量来源？
  - 流量质量差异（以搜索引擎和自有流量为例进行对比）
    - 跳出率



    - 平均访问深度



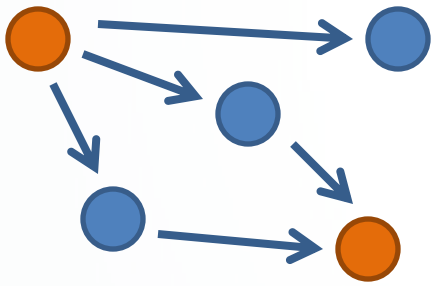
“你要到哪里去？”(1)

- 网站的终极目标：促使用户形成转化效果

电子商务	支付订单
社会化媒体	一篇原创文章
搜索引擎	跳转至来源网站

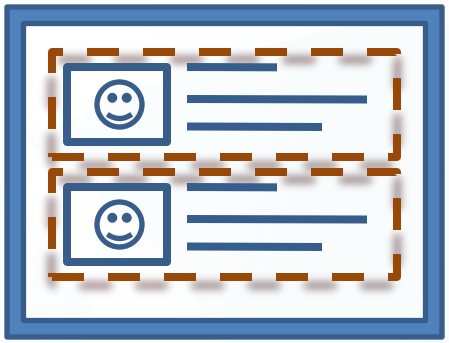
“你要到哪里去？” (2)

- 在转化的过程中，用户会留下各种痕迹



页面访问路径

页面点击行为



页面结构化数据

现在我们回到主题

- 基于用户行为的数据分析与挖掘的目标
  - 根据用户的访问路径、页面点击、访问内容等信息，发现共性，找促使网站产生更好转化效果的方法。

工作流程

数据采集 (1)

- 采集哪些数据？
  - 网页浏览行为（Pageview）
  - 转化效果
  - 用户在页面上的点击行为
  - 页面元数据

数据采集 (2)

- 如何采集？（以酷讯旅游网为例）

网页浏览行为	• javascript异步采集，get参数携带字段值。
页面元数据	• 将结构化数据树状存储。
点击行为	• 向DOM节点挂载onclick事件。
转化效果	• 通过统计中间页强制重定向。

数据采集 (3)

- 定义数据的格式
  - 以方便数据清洗和分析为第一要务
  - 根据数据规模、维护难度选择不同的方案
- 选择数据的存储方式
  - \t \n分割的文本
  - 关系型数据库
  - Hadoop
- 选择合适的数据流向
  - 拉
    - 从上游系统向数据分析引擎单向推送数据
    - 保证数据分析引擎与上游系统是互相独立的

数据采集 (4)

- 常见问题
  - 测量误差
    - 因统计代码异步加载导致某些请求未被统计到
  - 数据收集错误
    - 中文字段的乱码
  - 数据收集遗漏
    - 字符串太长，超过了字段限制而被截断
- 我们会在数据清洗环节进行解决！

与上游数据商的关系很重要

- 例：向数据表增加last update time字段

```
CREATE TABLE `logs` (
...,
`last_update_time` TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE
    CURRENT_TIMESTAMP,
...
)
```

field1	field2	...	fieldn	last update time
aaa	111		xxx	2011-01-01 12:34:56
bbb	222		yyy	2011-01-01 12:34:57
ccc	333		zzz	2011-01-01 12:34:58

上游数据商的数据库

field1	field2	...	fieldn	last update time
aaa	111		xxx	2011-01-01 12:34:56
bbb	222		yyy	2011-01-01 12:34:57
ccc	333		zzz	2011-01-01 12:34:58

商业智能数据库

(蓝色为更新的数据）

- 说服上游数据商调整数据结构，能够形成双赢。

数据清洗 (1)

- 什么是数据清洗？
  - ETL = Extract, Transform, Load（提取，转换，加载）
- 为什么要数据清洗？
  - 脏
    - 例：性别字段非男非女，IP字段包含字母
  - 复杂
    - “北京海淀酒店” = “北京市海淀区酒店”？
  - 不完整
    - 字段太长被截断，导致内容失去意义
- 高达75%的数据分析初始工作时间会花在这里。

数据清洗 (2)

- 常见的数据清洗工作示例

工作内容	示例场景	解决方案
过滤	网站记录用户一些行为数据，通常使用cookie进行记录，如果用户禁用了cookie或清除过cookie，就会造成统计到的数据不完整。	丢弃
消重	同一个用户，在一段很短的时间内，多次点击同一个按钮或者刷新同一个页面，如果不进行处理，则将会影响对数据分析阶段的数据准确性，给数据分析带来错误的结果。	设定阈值，超过阈值的记录进行丢弃
格式化	用户搜索关键词存在乱码或者过长	尝试判断编码格式，并进行转换
预处理	日志中会记录用户访问的IP地址，但是没有记录用户所在地，这样无法通过数据分析确定用户的所属信息，不利于城市以后的推广信息的推送。	通过内部的IP2City功能，将日志中的IP地址处理成城市，并对城市建立省>市>区的父子关系。便于从多个角度进行数据分析。

数据统计 (1)

- 基于用户行为的数据，要统计哪些？（以酷讯旅游网为例）

数据统计 (2)

- 常用第三方流量统计系统

- 为什么我们还要做自己的数据统计呢？
  - 各种个性化的需求
    - 例：无法支持任意维度的统计
    - 例：频道间的内部交叉流量无法识别

数据统计 (3)

- 酷讯旅游网内部统计系统（labrador）简介

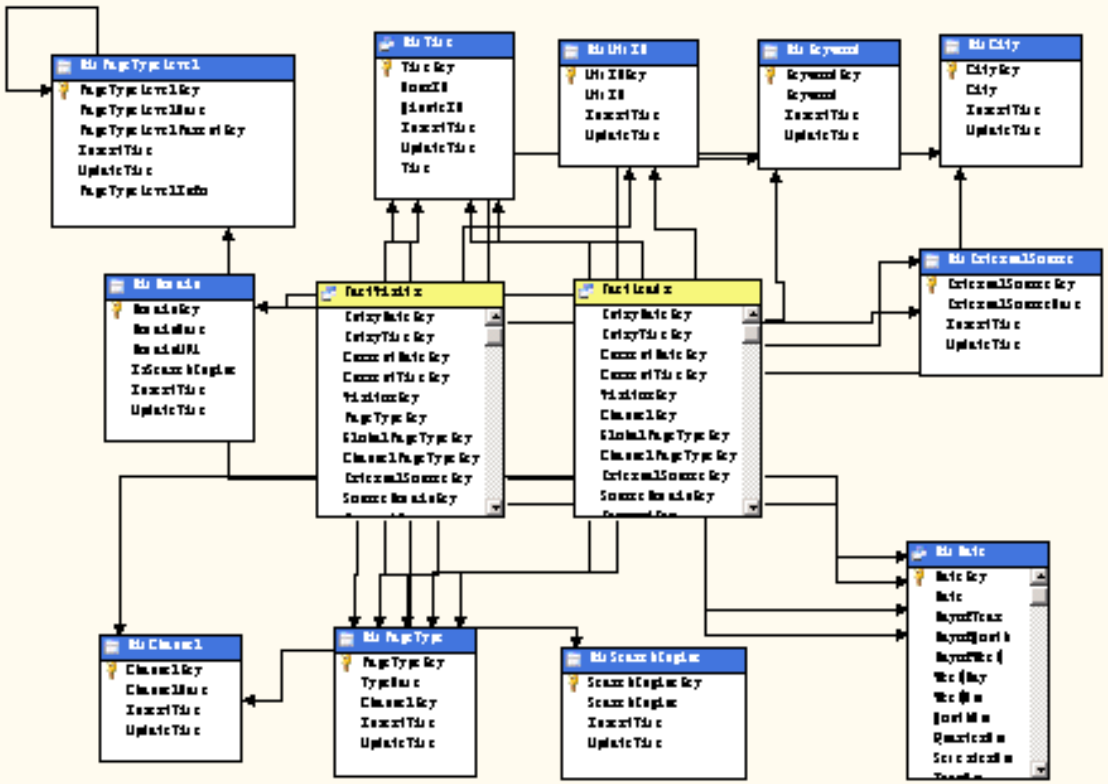
数据统计 (4)

- 重要特性
  - 支持流量的实时查看，最慢为小时级
  - 支持频道间交叉流量的统计
- 对数据安全的考虑
  - 浏览器安全证书：不可仿冒，不可抵赖
  - 详细的审计日志

数据分析

- 例一：
  - 利用SQL Server Analysis Services 的 OLAP （联机分析处理）解决方案，分析SEM投放的投入产出比
- 工作流程
  - 建立事实表和维度表
  - 创建多维数据集
  - 进行ETL操作

事实表和维度表 (1)

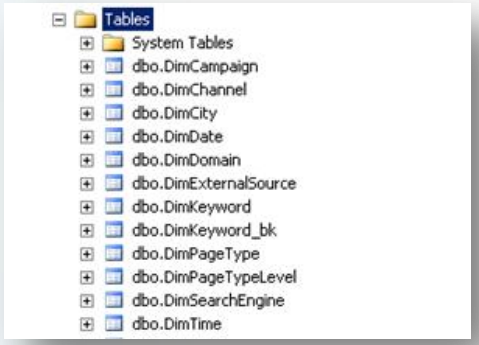


事实表和维度表 (2)

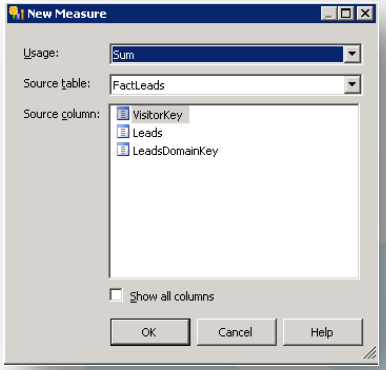
- 事实表
  - Visits
  - Leads
- 维度表
  - ChannelPageType 首次到达某频道的页面类型
  - CurrentDate 当前时间
  - EntryDate 此次访问所在Visits开始时间（该visits的第一次访问时间）
  - GlobalPageType 首次到达酷讯的页面类型
  - Lead Domain 跳往下游网站的主域名
  - Keyword 搜索关键词

创建多维数据集的过程

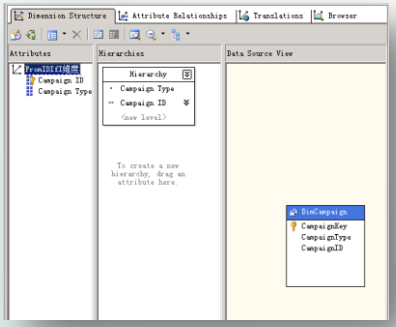
1. 在数据库层新建事实表和维度表



2. 建立度量信息



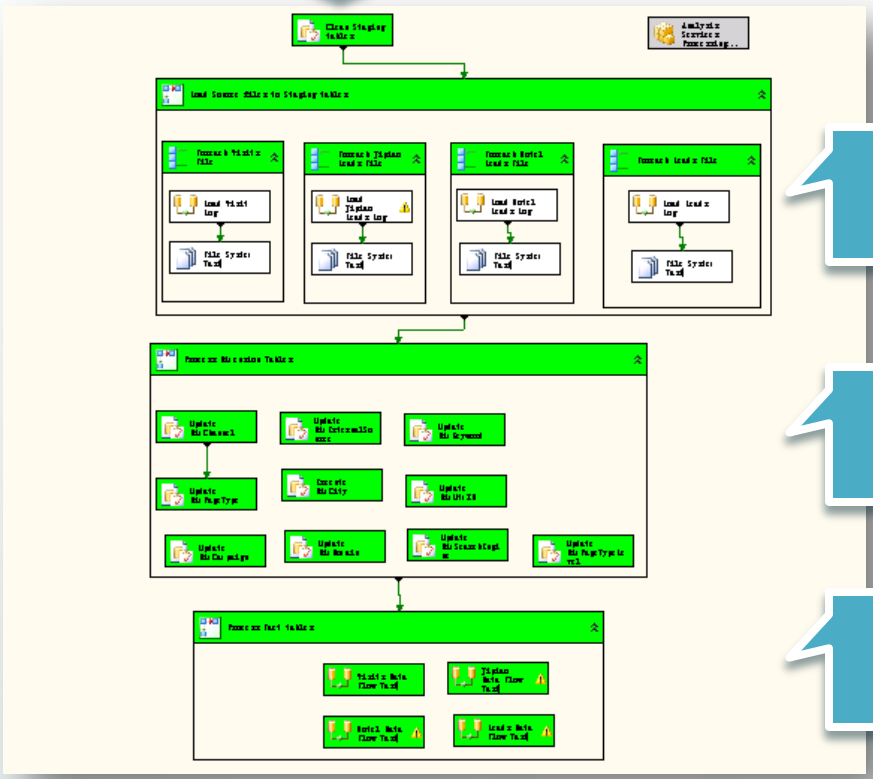
3. 建立维度



4. 将度量值与维度相关联

ETL流程图

清除上次执行时产生的临时文件

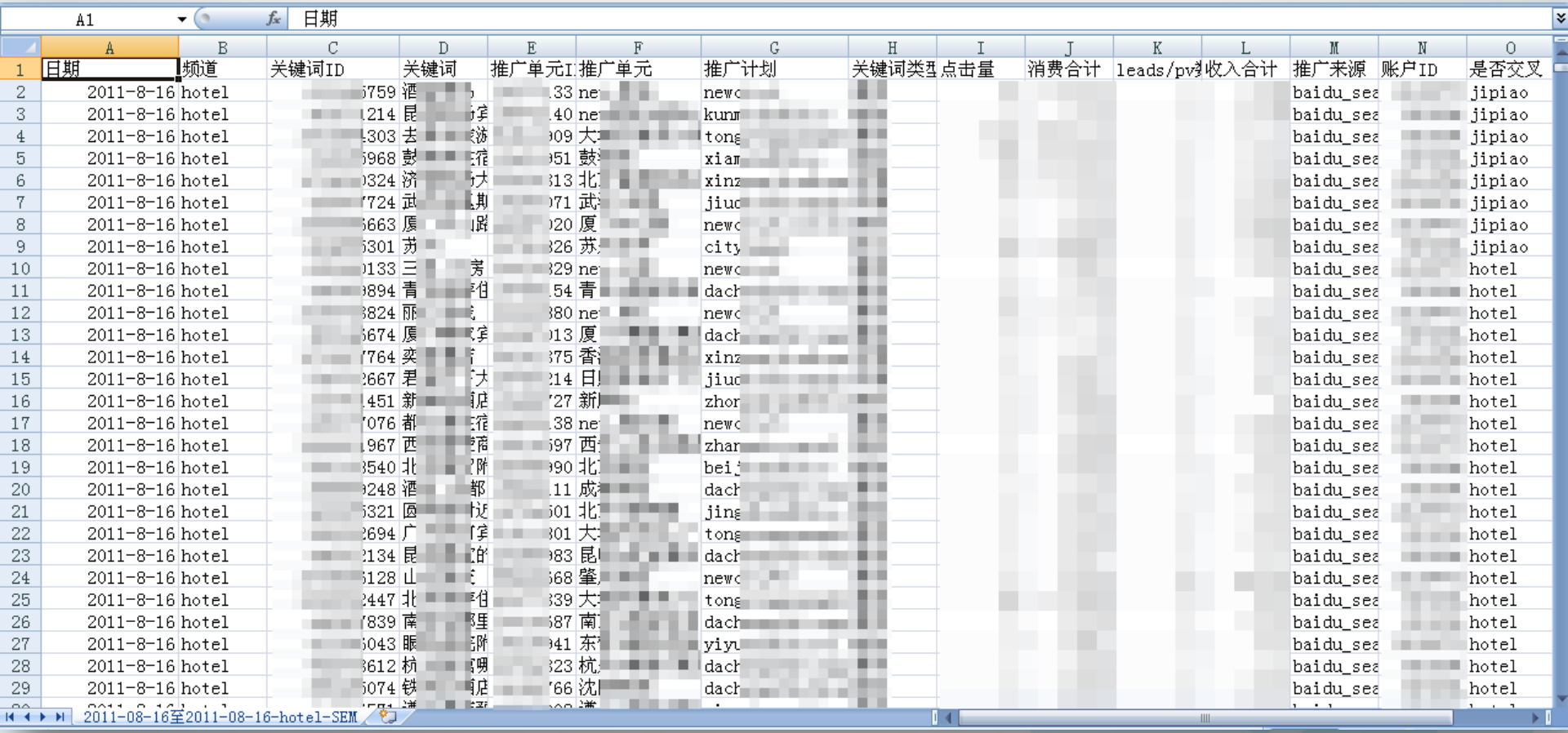


E - 抽取

T - 转换

L - 加载

处理结果



日期	频道	关键词ID	关键词	推广单元I	推广单元	推广计划	关键词类型	点击量	消费合计	leads/pv	收入合计	推广来源	账户ID	是否交叉
2011-8-16	hotel	[illegible]759	酒[illegible]	[illegible].33	ne[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]1214	昆[illegible]宾	[illegible].40	ne[illegible]	kunm[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]4303	去[illegible]游	[illegible]909	大[illegible]	tong[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]5968	鼓[illegible]宿	[illegible]951	鼓[illegible]	xiam[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]0324	济[illegible]大	[illegible]813	北[illegible]	xinz[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]7724	武[illegible]期	[illegible]071	武[illegible]	jiuc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]5663	厦[illegible]路	[illegible]020	厦[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]5301	苏[illegible]	[illegible]826	苏[illegible]	city[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	jipiao
2011-8-16	hotel	[illegible]0133	三[illegible]芳	[illegible]829	ne[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]8894	青[illegible]伯	[illegible].54	青[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]3824	丽[illegible]	[illegible]380	ne[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]5674	厦[illegible]宾	[illegible]013	厦[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]7764	奕[illegible]	[illegible]375	香[illegible]	xinz[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]2667	君[illegible]大	[illegible]214	日[illegible]	jiuc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]1451	新[illegible]店	[illegible]727	新[illegible]	zhon[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]7076	都[illegible]宿	[illegible].38	ne[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible].967	西[illegible]商	[illegible]597	西[illegible]	zhan[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]3540	北[illegible]附	[illegible]990	北[illegible]	bei[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]9248	酒[illegible]都	[illegible].11	成[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]5321	圆[illegible]近	[illegible]601	北[illegible]	jing[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]2694	广[illegible]宾	[illegible]301	大[illegible]	tong[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]2134	昆[illegible]的	[illegible]983	昆[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]5128	山[illegible]	[illegible]568	肇[illegible]	newc[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]2447	北[illegible]伯	[illegible]339	大[illegible]	tong[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]7839	南[illegible]里	[illegible]687	南[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]6043	眼[illegible]附	[illegible]941	东[illegible]	yiyu[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]3612	杭[illegible]明	[illegible]323	杭[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel
2011-8-16	hotel	[illegible]5074	铁[illegible]店	[illegible]766	沈[illegible]	dach[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	baidu_sea	[illegible]	hotel

2011-08-16至2011-08-16-hotel-SEM

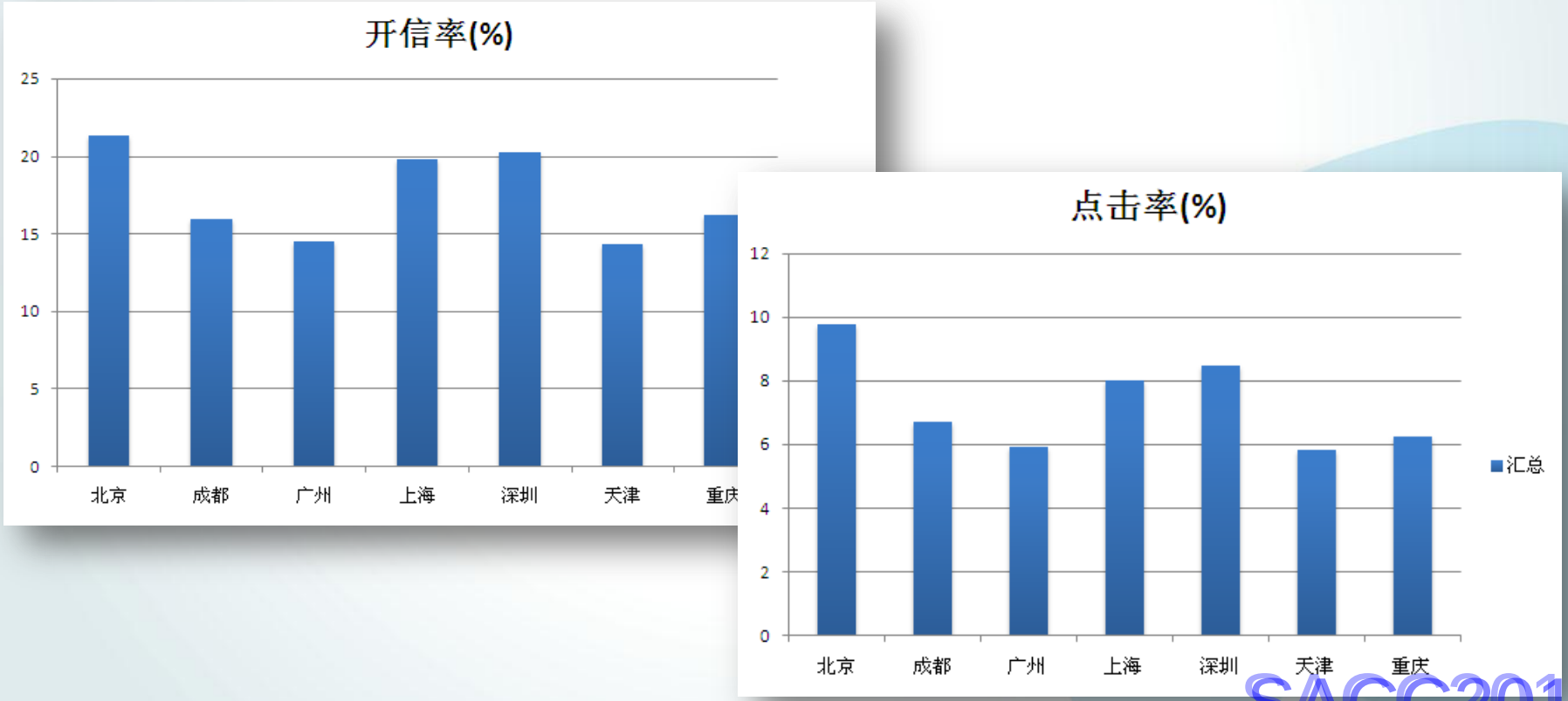
数据分析

- 例二：
  - 分析不同城市用户邮件营销的开信、点击效果
- 工作流程
  - 数据准备
  - 进行ETL操作

工作列表

- 数据准备
  - 行政区划数据库
  - IP至城市对应关系的数据库
- ETL
  - 抽取：将开信日志和点击日志导入数据库
  - 转换：将开信日志表和点击日志表中的IP转换成城市
  - 加载：将转换后的城市和对应的email插入email和城市对应关系表中

邮件营销的分析结果


开信率(%)
25
20
15
10
5
0
北京
成都
广州
上海
深圳
天津
重庆
点击率(%)
12
10
8
6
4
2
0
北京
成都
广州
上海
深圳
天津
重庆
汇总

数据挖掘

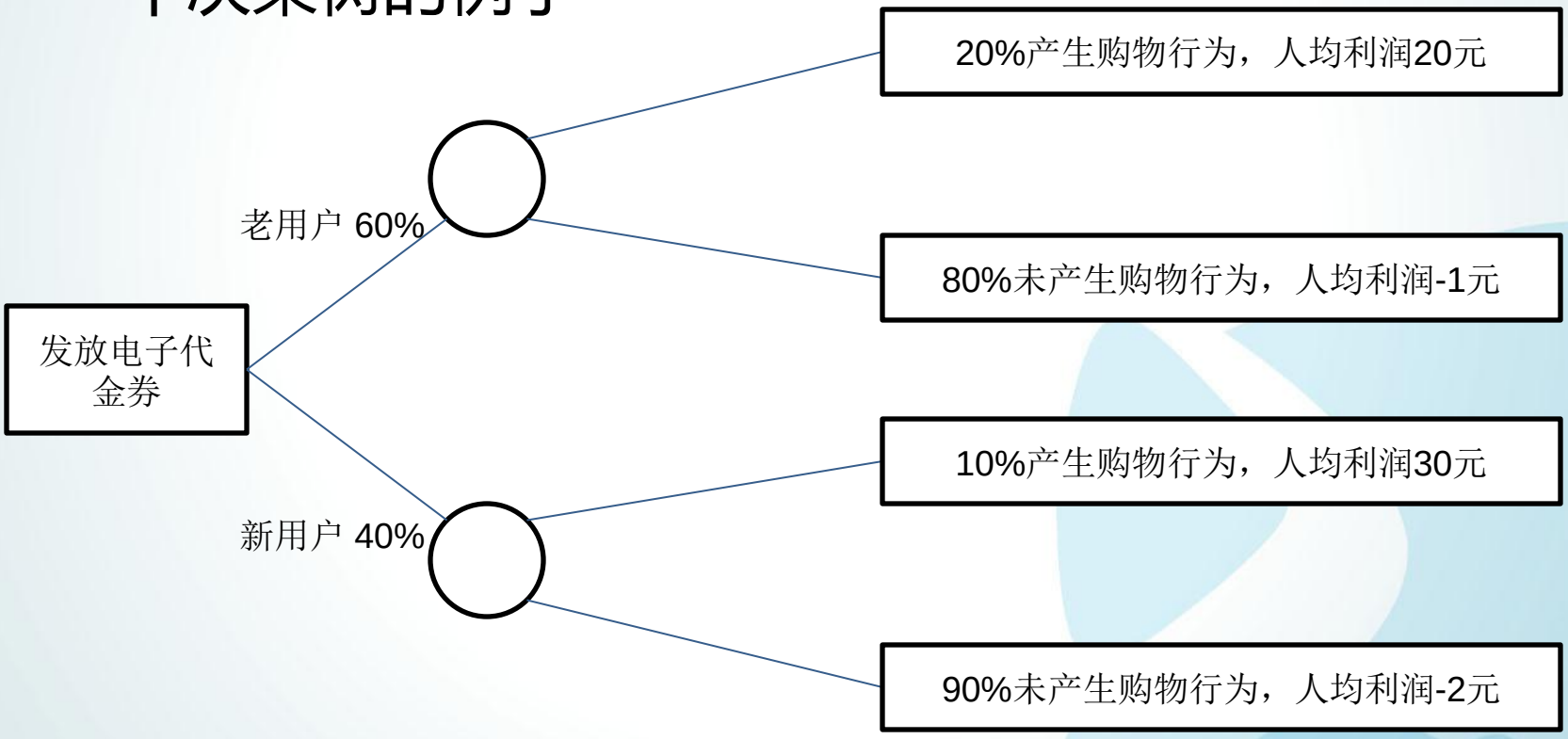
- 四种任务
  - 聚类分析
  - 预测建模
  - 关联分析
  - 异常检测

聚类分析 (1)

- 发现紧密相关的观测值组群，使得同组的相似性越大，不同组的差别越大，以达到较好的聚类效果
- 根据聚类得到的不同观测值组，做出决策树，为业务部门提供决策支持

聚类分析 (2)

- 一个决策树的例子



结论：发放电子代金券的人均利润为60%*(20%*20+80%*(-1))+40%(10%*30+90%*(-2))=2.4元，值得一做。

预测建模

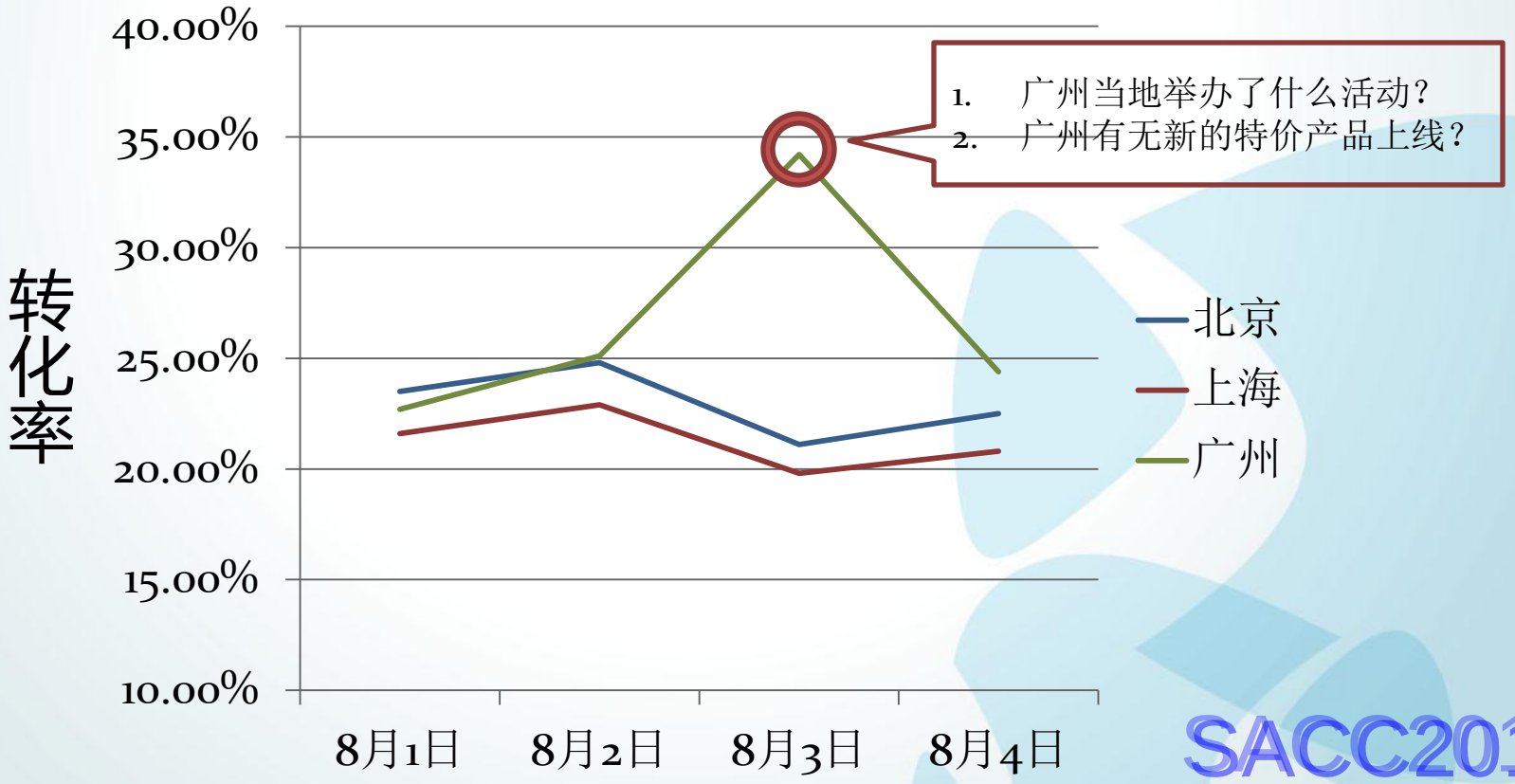
- 以自变量函数的方式为目标建立模型
- 分类：预测离散的目标变量
  - 例：在过去5年内，早上10点比下午4点的流量均高出20%，可以预测未来一段时间也是这个比例。
- 回归：预测连续的目标变量
  - 一元线性回归
  - 多元线性回归
  - 非线性回归

关联分析

- 用户在预定机票的同时预定了什么？

异常检测

- 识别其特征显著不同于其他数据的观测值（异常点，离群点）



回顾

- 需要明确用户行为的衡量指标体系
- 用户行为统计
  - 不同来源的流量质量差异明显
- 采集与清洗
  - 数据存储的格式要利于查询
  - 需要处理好与上游数据商的关系
  - 将足够的资源投入数据清洗工作
- 分析与挖掘
  - 数据分析的两个例子：SEM投入产出比、邮件营销效果
  - 数据挖掘的四类工作

Q&A