互联网广告的点击率预测
Click-through Rate Prediction

You Ouyang

2013/05/16

本节课内容大纲

- 背景介绍
- 相关研究
  - 搜索广告领域
  - 展示广告领域
- 经验分享

背景介绍

互联网广告中的点击

- 广告点击的定义

- 两种情况
  - 正常点击
  - 异常点击

点击率预测

- 点击率 (Click-through Rate，CTR)
  - 曝光产生点击的概率
  - 点击率=点击总数/曝光总数*100%
    - 广告位的点击率、一段时间的点击率等等

- 点击率预测的目的
  - 评价广告吸引力的重要指标
  - 直接影响按点击计费模式的收入
    - 曝光量固定的情况下，收入=CTR*CPC
  - 预估交易曝光的效果

影响点击率的因素

- 广告自身的影响
  - 广告类型：文字、图片、富媒体、……
  - 广告内容：颜色、构图、语言、……
- 上下文环境的影响
  - 广告位属性：媒体、类型、位置、尺寸、……
  - 曝光属性：发生时间、停留时间、……
- 广告浏览者的影响
  - 人群属性：性别、年龄、兴趣爱好、……
  - 历史行为：浏览过此广告几次、浏览过同品牌广告几次、……

互联网广告的点击率预测

搜索广告的点击率预测

- 点击售卖模式
  - 用户的搜索结果页上的一系列广告位置
  - 一系列赞助广告及各自的CPC报价
  - 为每个位置分配最合适的赞助广告
  - 被展示广告被用户点击时，搜索引擎根据CPC收取费用
- 即时竞价模式
  - 收费模式

搜索广告的点击率预测

- 点击率预测的意义
  - 搜索引擎
    - 搜索引擎的收入 = CPC * CTR
    - 按照CPC * CTR排序来决定位置分配
  - 客户
    - 提高固定投入下的广告效果

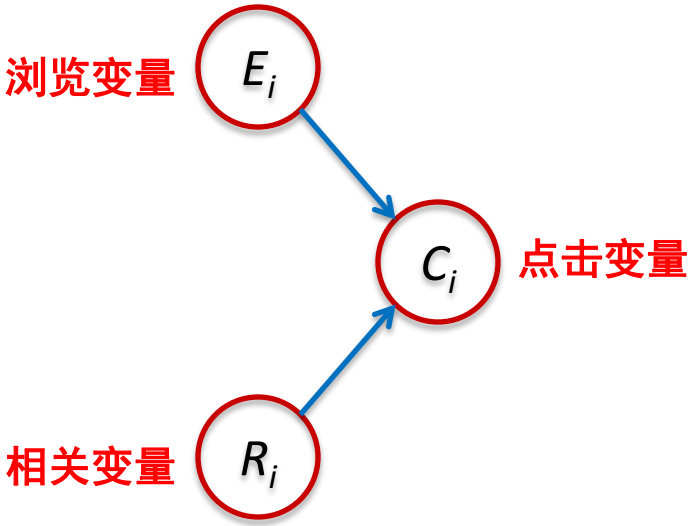
- 基本预测方法
  - 基于历史数据预测点击率
  - 点击率 = 历史点击/历史曝光

概率模型

- 历史数据估计方法的局限性
  - 预测未投放过的位置+广告组合的点击率
- 用概率统计模型对用户行为进行建模和预测
- 常见概率模型
  - Position model
  - Cascade model
  - User browsing model
  - Dynamic Bayesian network model

Position model

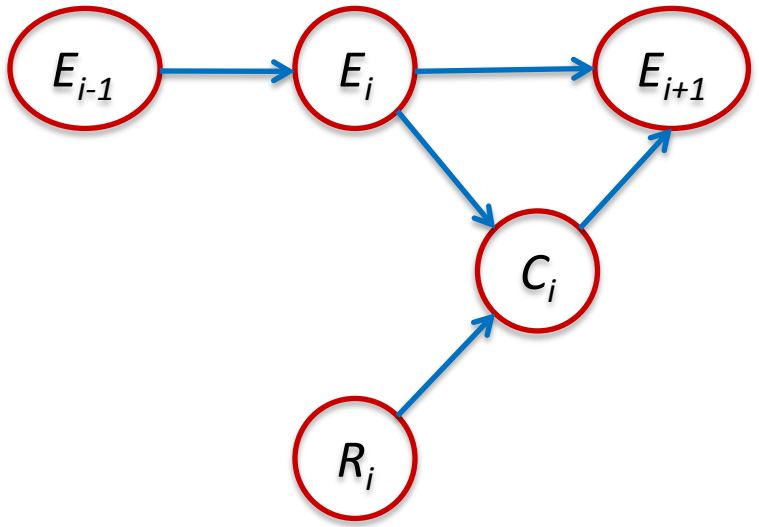
- 主要考虑广告位置对浏览概率的影响



- Examination hypothesis
  - 点击发生须满足
    - 发生浏览行为
    - 用户认为文档相关
  - 模型
    - 预测用户浏览第i个位置的文档d_j的点击率
    - $C_i = 1 \Leftrightarrow E_i = 1, R_i = 1$
    - $P(C_i = 1) = P(E_i = 1) * P(R_i = 1 \mid E_i = 1) = P(E_i = 1) * P(R_i = 1) = \lambda_i * r_{di}$
    - E、R为无法直接观测的隐藏变量，EM算法估计参数

Cascade model

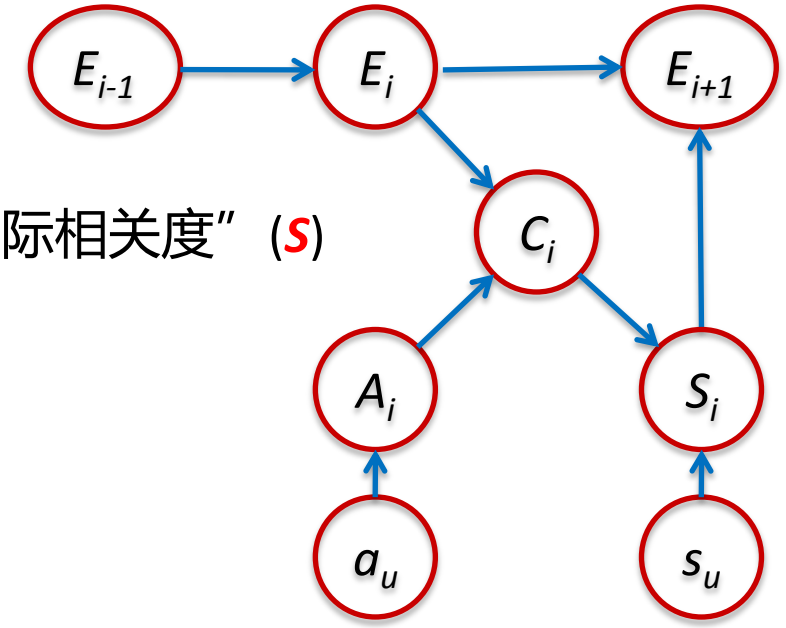
- Cascade hypothesis
  - 用户由上至下地检视搜索结果，点击后放弃整个搜索结果
    - $P(E_{i+1} = 1 \mid E_i = 0) = 0$
    - $P(E_{i+1} = 1 \mid C_i) = 1 - C_i$
- Dependent Click model
  - 推广到多次点击
  - $P(E_{i+1} = 1 \mid E_i = 1, C_i = 1) = \lambda_i$



Models with multiple clicks

- Click chain model
  - 被点文档的相关度影响继续浏览的可能性
    - $P(E_{i+1} = 1 \mid E_i = 1, C_i = 0) = \alpha_1$
    - $P(E_{i+1} = 1 \mid E_i = 1, C_i = 1) = \alpha_2 * (1 - r_{di}) + \alpha_3 * r_{di}$

- Dynamic Bayesian network model
  - 区分链接的“预计相关度”(***A***)和“实际相关度”(***S***)
  - $P(E_{i+1} = 1 \mid E_i = 1, C_i = 0) = \alpha$
  - $P(E_{i+1} = 1 \mid E_i = 1, C_i = 1) = \alpha * (1 - s_{di})$



Models with multiple clicks

- User browsing model
  - 用户浏览某位置的概率跟此位置距离上次点击位置的距离有关
  - $P(E_i = 1) = \gamma_{i,i-li}$
  - $P(C_i = 1) = r_{di} * \gamma_{i,i-li}$

- Bayesian browsing model
  - UBM中的参数Bayesian化

对稀疏数据的处理

- 点击率预测中的数据稀疏问题
  - 很多关键词对应的曝光/点击较少
  - 新关键词缺少历史数据

- 解决方案
  - 基于时间、类型等维度将不同的数据关联到一起

数据关联方法

- 根据时间平滑
  - 用前一天的曝光/点击总数对后一天的数据进行平滑
    $\hat{C}_j = \gamma C_j + (1-\gamma)\hat{C}_{j-1}$
  - 用平滑之后的曝光数和点击数来计算后一天的点击率

- 根据类型平滑
  - 同类型关键词的点击率具有相关性
  - 汇总同类型关键词的数据进行预测

识别关键词类型

- 关键词分类
  - 基本类型分类
    - 浏览查询、学术查询、网址url查询、……
  - 关键词的层次结构
    - 旅游 - 节假日旅游 - 元旦旅游
  - 关键词中隐藏的用户特征
    - 普通浏览、购买搜索、……
- 关键词聚类
  - 根据关键词内容、搜索结果页等特征对关键词进行聚类
  - 聚类 vs 分类

用户属性对点击率的影响

- 用户之间的区别
  - 用户在点击上存在喜好区别
    - 兴趣爱好
    - 上网习惯
  - 不同用户对于同个东西的认识可能不一样
    - Opera？Apple？
- 不同用户点击同样广告的概率不一样

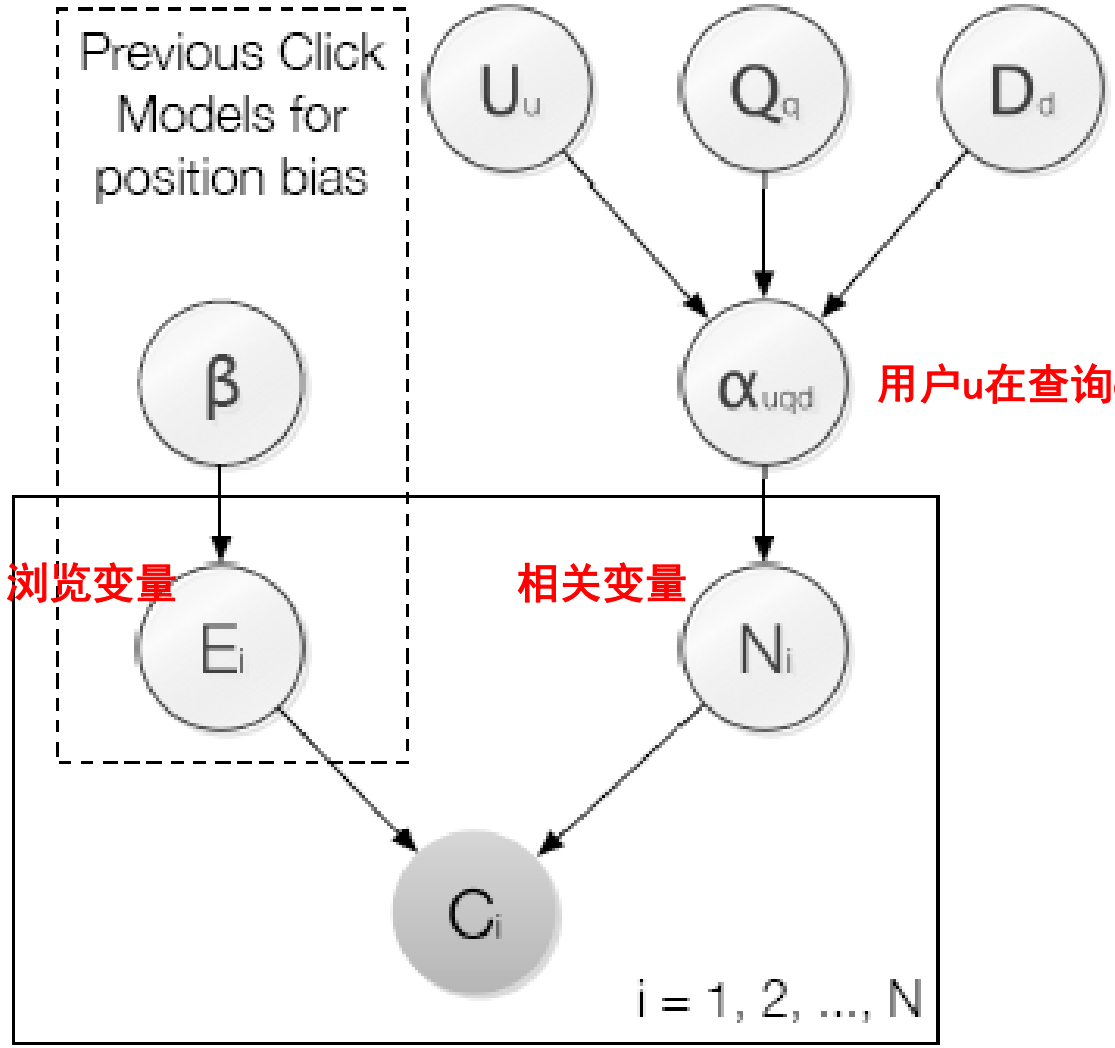
VPI用户浏览模型

- VPI (varying persistence and initial motivation)
  - 考虑了用户的最初搜索满意程度和耐心度
  - 最初考虑整个搜索结果的概率 $P(E_1 = 1) = u^q$
  - 未点击时放弃整个搜索结果的概率 $P(E_{i+1} = 1 \mid E_i = 1, C_i = 0) = \lambda^q$
- 模型中的五组变量
  - U 用户最初的搜索动机，E 是否浏览，C 是否点击，A 文档是否相关(表面)，S 文档是否相关(实际)
  - 五组变量以概率关系建模，只有C是可见变量
  - 仍然可以用EM算法求解

基于协同过滤的用户兴趣分析

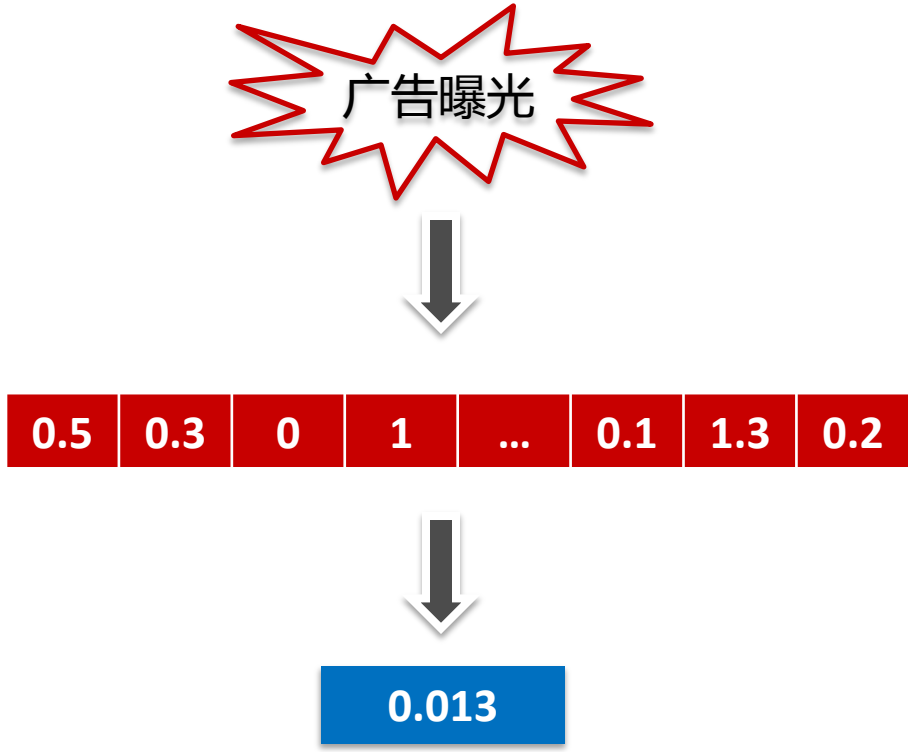
- 基于协同过滤的方法来引入用户的影响
  - 用Matrix Factorization方法找出用户(U)、关键词(Q)、文档(D)之间的关联
  - 解决数据稀疏问题
- 结论：同时考虑UQD的模型结果优于只考虑QD的模型
- 可以跟更复杂的浏览模型结合

基于协同过滤的用户兴趣分析(图示)



基于特征的方法

- 要素
  - 特征
  - 目标函数
- 用特征向量描述广告相关信息
  - 广告内容、用户信息、上下文环境等
- 基于特征向量预测点击率
  - 经验模型
  - 机器学习模型
    - 分类、回归、Learning-to-rank等方法

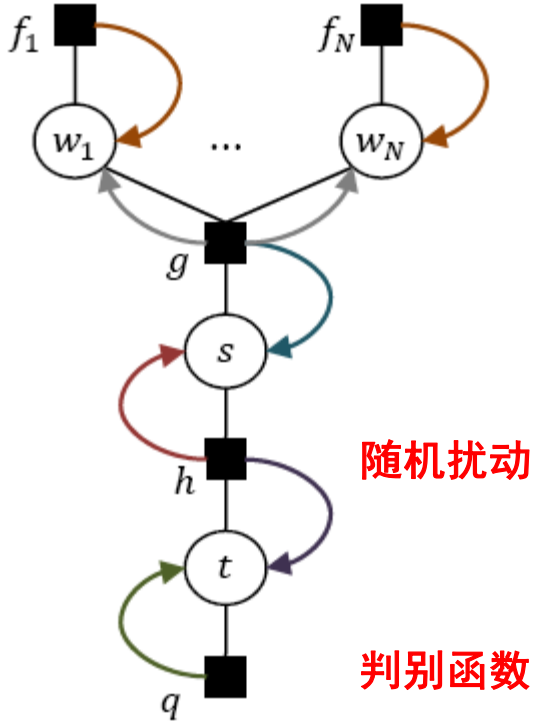


Bayesian CTR prediction

- 特征
  - 基于查询词的特征
  - 基于结果页内容的特征
- 模型
  - 判别函数
    - Probit link function加上Gauss noises

$$p(y|\boldsymbol{x},\boldsymbol{w}) := \Phi\left(\frac{y \cdot \boldsymbol{w}^T\boldsymbol{x}}{\beta}\right)$$

  - 在线学习，并行计算



Estimating CTR for new keywords

- 设计不仅仅依赖关键词本身的特征
  - 关键词、长度
  - 竞标词、相关关键词
  - 广告吸引力、商户信用度、落地页质量、相关度
  - 特定类别指向度
- 模型
  - 逻辑回归

垂直搜索和点击率预测

- 影响垂直搜索点击率的分析
  - 分析数据
    - 位置对垂直搜索结果和普通搜索结果的影响有区别
    - 垂直结果会影响周围的普通结果的点击率
  - 总结结论
    - 垂直结果更引人注意、容易让人满意，但普通网页结果往往是必须的

- 根据结论设计点击率预测模型
  - 在基础概率模型上加上attention bias和exploration bias
  - Attention bias (A) 垂直结果带来的周围浏览概率增加

Bias的影响 $P(E_i = 1 \mid A = 0) = \phi_i \quad P(E_i = 1 \mid A = 1) = \phi_i + (1 - \phi_i)\beta_{dist}$

Bias的出现概率 $P(A = 1 \mid pos_v) = h_{pos_v} \quad P(A = 1 \mid d) = u_d$

  - Exploration bias (D) 点击垂直结果导致的搜索进程结束概率

Bias的影响

$$P(E_i = 1 \mid V_i = 1 \vee (D = 0, V_i = 0)) = \phi_i$$

$$P(E_i = 1 \mid D = 1, V_i = 0) = 0$$

展示广告的点击率预测

- 展示广告跟搜索引擎广告的区别
  - 更丰富的广告位类型
  - 更丰富的广告类型
  - 更复杂的用户点击倾向
  - 更复杂的属性交互影响
- 基于特征的方法更为常用
  - 特征对结果的影响非常巨大

广告位置对点击率的影响

- 相似位置点击率相近
  - 对位置进行分类/聚类，一类广告位同时估计点击率
- 自动学习类别信息
  - 定义大量位置信息特征
  - 每个位置的特征值形成位置向量Z_c
  - 分类函数为

Global参数　Campaign自身的参数

$$f(\boldsymbol{x}_u, \boldsymbol{z}_c, c) = \boldsymbol{x}'_u D \boldsymbol{z}_c + \boldsymbol{x}'_u \boldsymbol{\beta}_g + \boldsymbol{x}'_u \boldsymbol{\beta}_c$$

每个位置特征下的参数

  - 最优化

$$\min_D \sum_{u,c} L(y_{u,c}, \boldsymbol{X}'_u D \boldsymbol{z}_c + \boldsymbol{x}'_u \boldsymbol{\beta}_g + \boldsymbol{x}'_u \boldsymbol{\beta}_c)$$

总误差

$$+ \lambda \|D\|_p + \lambda \|\boldsymbol{\beta}_g\|_p + \sum_c \lambda_c \|\boldsymbol{\beta}_c\|_p$$

正则化因子

单位置点击率研究

- 针对一个位置分析
  - 多屏滑动的广告位置中，各个位置的点击率
- 数据分析
  - 对比两类人群(热门/随机)
  - 考察时间、地域、重复观看等
- 模型
  - 考虑首看CTR和后续耐心度

$$\theta_{u,ilt} = \theta_{0ilt} \exp\{g(R_u)\}$$

某用户在某时间某地域的点击率　随着次数增加点击率的衰退

  - 在训练语料上Fit模型估计参数

广告内容对点击率的影响

- 图像对点击率的影响
- 基于特征的研究结果
  - 特征
    - Global features，Local features，Advanced features
    - 灰度，亮度，颜色分布，颜色和谐，支配色，分块特征，OCR结果中的object数等
  - 模型
    - Support Vector Regression (支持向量回归)
- 结论
  - 高对比度、组成个体不过量、组成个体居中的图像较好

广告内容对点击率的影响

- 产品搜索中展示图片的影响
- 特征
  - 物品介绍、查询词、卖家信息
  - 图片的长宽比、亮度、动态、对比、背景
  - 图片的颜色、纹理、形状特征

- 模型
  - Restricted Boltzmann machine
  - Logistic regression

- 结论
  - 除卖家信息和运输费用等少数特征外，图片特征是决定用户是否点击的关键特征

用户对点击率的影响

- 行为定向 (behavior targeting)
  - 用户的历史浏览记录
- 人群定向 (demographic targeting)
  - 人口属性/兴趣爱好
- 需要考虑的两个问题
  - 如何得到信息
  - 如何利用信息

投放经验分享

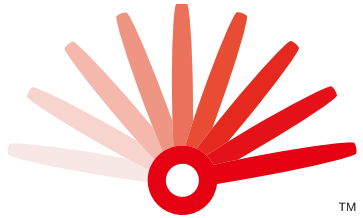
- 广告位
- 广告创意
- 定向
  - 时间、地域等
  - 频次、用户等

参考文献

- Bbm: bayesian browsing model from petabyte-scale data
- Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine
- Unsupervised Learning of Result Page Context for Clickthrough Analysis in Sponsored Search
- Spatio-Temporal Models for Estimating Click-through Rate
- Predicting ClickThrough Rate Using Keyword Clusters
- Predicting Ads' ClickThrough Rate with Decision Rules
- Classifying Web Search Queries to Identify High Revenue Generating Customers
- Beyond Ten Blue Links: Enabling User Click Modeling in Federated Web Search
- A Dynamic Bayesian Network Click Model for Web Search Ranking
- Web-Scale Multi-Task Feature Selection for Behavioral Targeting

参考文献

- Visual Appearance of Display Ads and Its Effect on Click Through Rate
- The Impact of Images on User Clicks in Product Search
- Search engine advertisements: The impact of advertising statements on click-through and conversion rates
- Position-Normalized Click Prediction in Search Advertising
- Personalized Click Model through Collaborative Filtering
- Modeling Browsing Behavior for Click Analysis in Sponsored Search
- Learning to Predict the Cost-Per-Click for Your Ad Words
- Impact of query intent and search context on click-through behavior in sponsored search
- Click-Through Rate Estimation for Rare Events in Online Advertising
- A User Browsing Model to Predict Search Engine ClickData from Past Observations



秒针系统
Miaozhen Systems™

Data 数据有乾坤！
is wonderful!