大数据挖掘技术及其应用
Big Data Mining Technology and Applications
第8章
Web广告

内容概要

教学大纲要求

- 在线广告概念
- 在线算法
- Adwords问题
- Adwords的实现

教学基本要求：掌握在线广告基本概念和在线贪心算法基本模型，了解Adwords问题和Adwords算法的实现。

Adwords

Web广告的现状

- 各种有趣的Web应用能够通过广告而不是用户订阅来维持生计
- 广播和电视行业已设法将广告作为它们的主要收入来源
- 大部分媒体，如报纸和期刊，却不得不采用混合策略，即同时从广告和订阅中获得收入。

Web广告的现状

- 广告信息平台（直投）
  - 58同城，赶集网，安居客
- 媒体网站广告（显示）
  - 同一个用户相邻两次访问，显示的广告不同
- 站内推荐
  - 还买了什么，猜你喜欢
- Adwords广告
  - 搜索引擎广告

8.1 在线广告相关问题

8.1.1 Web的常见场景

- 一些网站，如eBay, 新浪门户等，允许广告商以免费、付费或委托方式直接投放广告。
- 很多Web网站上的展示广告(display ad)。广告商按照每展示一次(某个用户下载一次网页则认为该网页上的广告被展示一次)的固定费率付费。通常，即使是同一个用户对网页的第二次下载，也会导致一个不同的广告展示。
- 诸如Amazon的在线商店在很多上下文中都显示广告。这些广告并非由广告商品的生产者来付费，而是由在线商店选出，以最大化顾客对商品感兴趣的概率。
- 搜索广告(search ad)包含在搜索结果中。广告商要为某些查询进行投标以获得在搜索结果中展示广告的权利，但是他们只在广告被点击的情况下才付费。显示广告的选择过程非常复杂。

显示式广告

- 新闻/媒体网站上的广告
- 按显示付费
  - CPM： Cost per thousand impressions
  - 和电视/杂志广告类似
- 问题
  - 读者和广告的匹配
  - 每次观看，只值几分钱
- 改进
  - 网站内容专门化，提高广告和读者的匹配程度。
  - 汽车网上，放汽车广告，价格就提上来了。

显示式广告的优化

- 根据用户历史，分析用户兴趣，提高广告的针对性
- 怎么获取用户历史数据?
  - 用户登录
    - Gmail
    - 微信
  - Cookie
    - 淘宝
  - 浏览器
    - 360
  - 网络爬虫

搜索广告的问题

- 按点击付费
  - Overture发明，付费排名（百度）
  - Google Adwords改进（搜索结果和广告分开）
- 模式：
  - 广告主竞标搜索关键字
  - 用户搜索问题，提供广告
  - 广告主预算

8.2 在线算法

- Off-line算法

  离线算法(off line algorithms)，是指基于在执行算法前输入数据已知的基本假设，也就是说，对于一个离线算法，在开始时就需要知道问题的所有输入数据，而且在解决一个问题后就要立即输出结果。
- On-line算法
  - 在线算法是指它可以以序列化的方式一个个的处理输入，也就是说在开始时并不需要已经知道所有的输入。
  - 执行算法时，不知道所有的输入
  - 类似第4章中的流
  - 例：
    - 淘宝推荐商品
    - 买滑板还是租滑板？

8.2 在线算法

- Off-line算法
- On-line算法
- 例 8.1
- 仿古家具制造商A对词项“chesterfield”的投标价格是10美分；
- 厂商B同时为“chesterfield”和“sofa”付出的投标价格是20美分；
- 他们两家的月广告预算都是100美元，并且没有其他厂商对这两个词投标； A：1000个，B：500个
- 如果投放机会低于500个，选B
- 如果投放机会不低于500个，如何选？
- 结论：无法保证在线算法和离线算法的效果总是一样好。

8.2 在线算法

- 8.2.2 贪心算法

定义：又称贪婪算法，是一种在每一步选择中都采取在当前状态下最好或最优（即最有利）的选择，从而希望导致结果是最好或最优的算法。比如在旅行推销员问题中，如果旅行员每次都选择最近的城市，那这就是一种贪心算法。

贪心算法在有最优子结构的问题中尤为有效。最优子结构的意思是局部最优解能决定全局最优解。简单地说，问题能够分解成子问题来解决，子问题的最优解能递推到最终问题的最优解。

8.2 在线算法

- 8.2.2 贪心算法

例8.2 8.1中所描述场景下的一个明显的贪心算法就是，将查询分配给还有预算的出价更高的广告商。对于上例的数据，前500个“sofa”，或“chesterfield”查询会分给B。此时，B的预算被花完，从而不会再分配给B任何查询。这之后，剩下的1000个“chesterfield”查询会分给A，而之后的“sofa”不会产生任何广告，因此它不会给搜索引擎带来任何收入。

8.2 在线算法

- 8.2.3 竞争率
- 在线算法不如最佳的离线算法效果那么好。
- 定义竞争率（competitive ratio）：存在某个小于1的常数c，使得对于任一输入，一个具体的在线算法的结果至少是最优离线算法结果的c倍。

  常数c如果存在的话，将被称为在线算法的竞争率
- 立足于最差情况

  例8.3

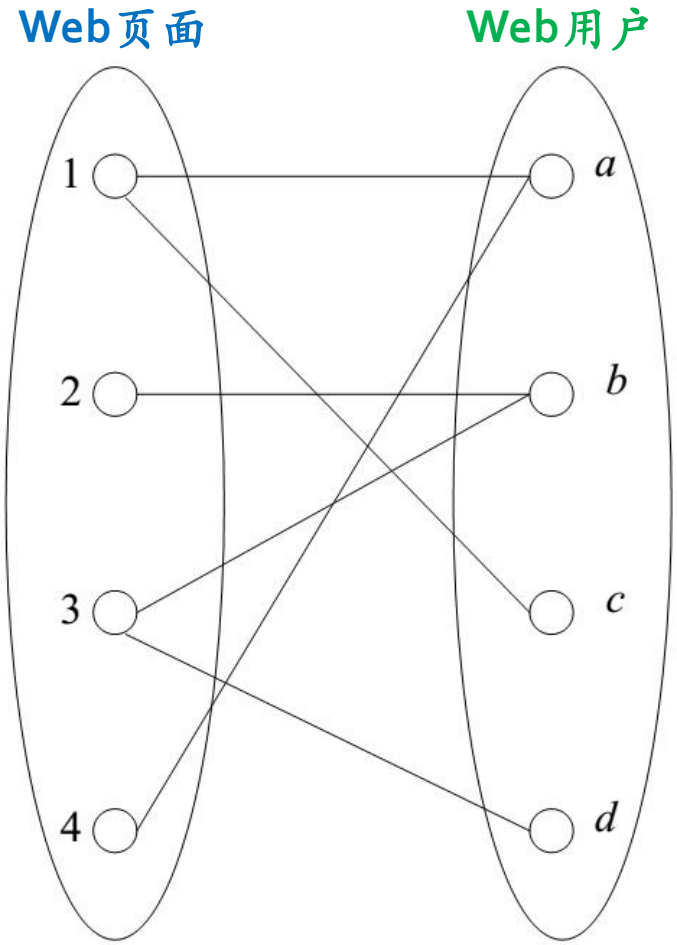
  贪心算法： 100元

  最优算法： 150元

  竞争率： 2/3

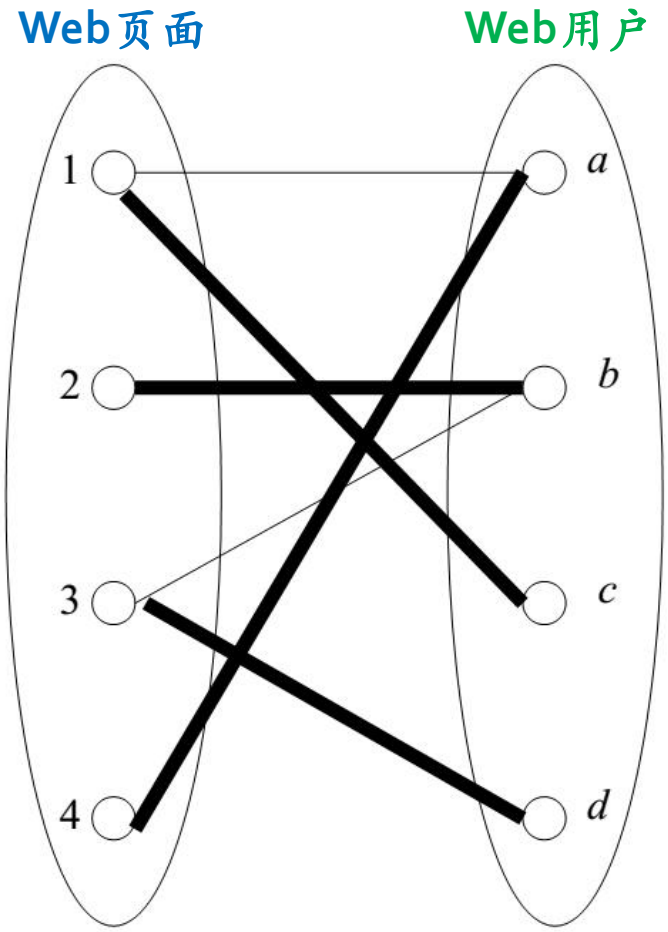
8.3 广告匹配问题

- 最大匹配是一个涉及二部图的问题。
- 二部图：由左右两个节点集合组成的图，每条边连接的都是左集合的一个节点和右集合的一个节点。

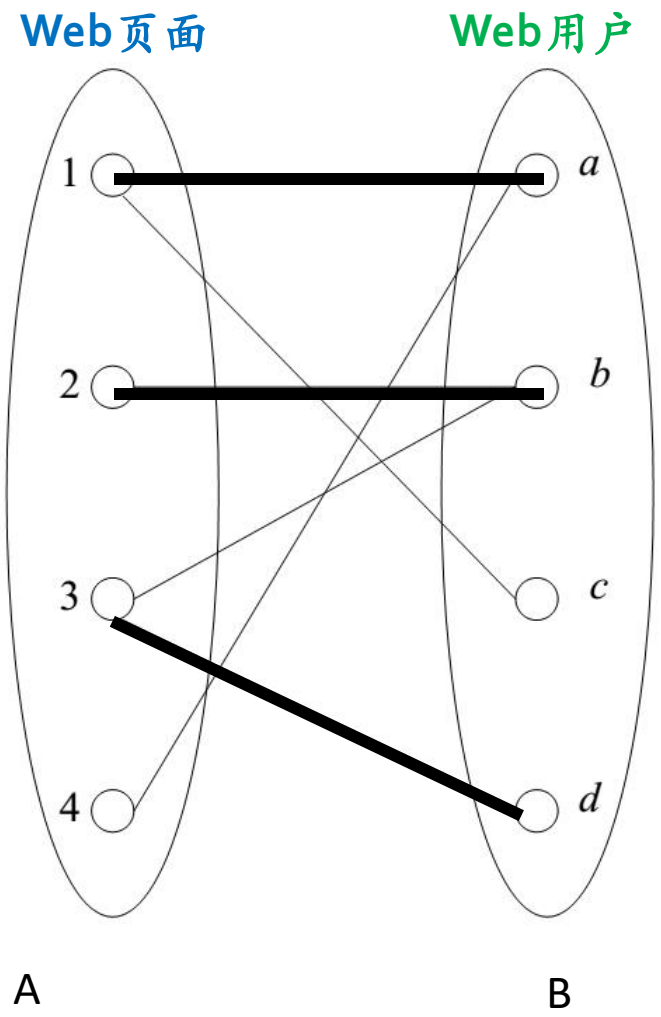


8.3.1 完美匹配

- 匹配：一个由边构成的子集，且**任何一个节点都不会同时是两条或多条边的端点**。
- 完美匹配：每个节点都在另一边找到对象
- 右图匹配数：4
- 最大匹配
  - 最大配对数所对应的那组匹配



8.3.2 最大匹配贪心算法



从左至右的策略:

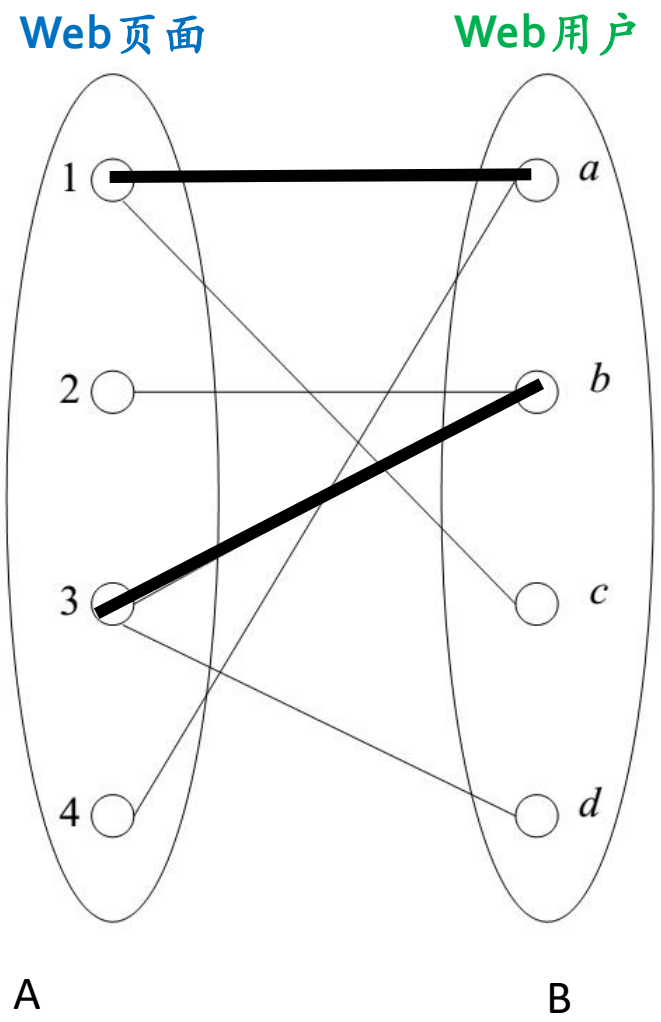
1 -> a

2 -> b

3 -> d

共3对

8.3.2 最大匹配贪心算法



从右至左

a -> 1

b -> 3

2对

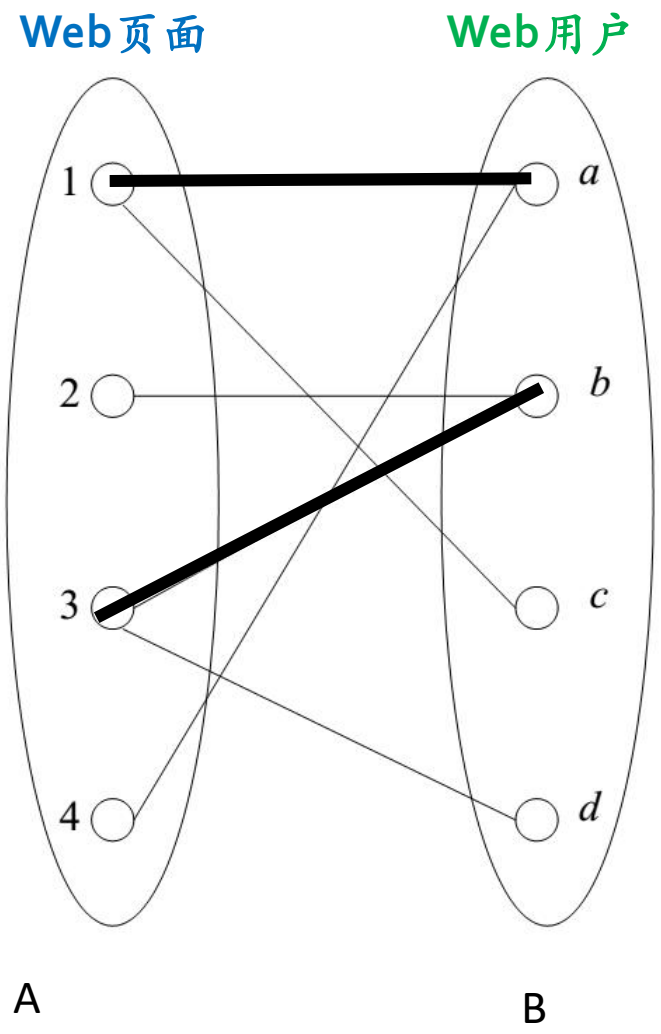
为什么这么差?

1）a,b先选。

2）她们有两个可选对象

3）可她们选择的，却是另外的组唯一可选的

8.3.2 最大匹配贪心算法



还有比这更差的吗?

如果没有，竞争率 = 1/2

8.4 adwords 问题

- 8.4.1 搜索广告的问题

问题描述

- 对于每条查询，谷歌显示的广告数目有限；
- Adwords系统的用户会指定一个预算，即他们愿意在一个月内为其广告的所有点击所付的费用；
- 并不是简单地按照广告商的出价来排序，而是按照其对每条广告的期望收益来排序

8.4 adwords 问题

- 8.4.2 Adwords问题的定义

已知：

- 众多广告商为搜索查询设定的投标价格集合。
- 每个广告商一查询对所对应的点击率。
- 每个广告商的预算。我们假定预算的周期为一个月，当然实际中任意时间单位都有可能使用；
- 每个搜索查询所显示的广告数目上限

约束：

- 反馈不会超过上述每条查询所显示的广告数目的上限
- 该集合中的每个广告商都对本条搜索查询出价
- 每个广告商必须剩余足够的预算来为广告的点击付费

8.4 adwords 问题

8.4.3 Adwords问题的贪心算法

- 对每条查询只显示一个广告;
- 所有广告商的预算都相等
- 所有广告的点击率都相等
- 所有的出价不是0就是1 或我们可以假设每个广告的价值(出价和点击率的乘积)相等。

8.4.4 Balance 算法

含义： 它将查询分配给出价最高且剩余预算最多的广告商。如果多个广告商的剩余预算相等，那么可以随意地选择其中的一个。