搜索引擎是如何最大化关键字广告收益的

通过什么样的规则来安排哪个广告给哪个关键字，才能最大化当天的收益呢？此问题可以抽象成“On-line带权二部图最大匹配问题”。

■ 文 / 陈文雄

Google的AdWords，或者其他搜索引擎的关键字广告，使用的基本都是“关键字竞价”(或者称“关键字拍卖”)的机制，对每个用户搜索的关键字，挑选为它竞价的广告来显示。

用户搜索的关键字到达搜索引擎的次序无法预知，每个竞价者为一个关键字出的价钱也千差万别，竞价者还会对每天的花费总额有一个封顶的预算，超过了这个预算，即使有合适的关键字，竞价者也不希望为它多花钱了。

搜索引擎们是通过什么样的规则来安排哪个广告给哪个关键字，以最大化当天的收益的呢？

此问题表述为：

有N个关键字竞价者，每个竞价者设定了一个当天的最大预算bi。
Q是一个关键字的集合。
每个竞价者i对一个关键字q∈Q，指定一个出价c_{iq}。
竞价开始后，关键字序列$q_1q_2...q_M$ ($q_j \in Q$)实时到达，每个q_j必须实时分配给某个竞价者i的广告以赚取收益c_{iqj}。

问题的目标是：在满足竞价者对关键字匹配要求的基础上，使当天总收益最大。

按照常识，可能会有这样的猜想：是不是只要在关键字到来时，选择对此关键字出价最高的广告显示，就可以达成当天收益的最大化？

这样的算法被称为贪婪算法，本文第5节我们会看到，使用贪婪算法，当天的收益可能仅为理论最大化收益的一半。

要解决此问题，本文先由简单的模型：二部图最大匹配问题开始介绍，在第4节抽象出与该竞价问题完全相同的模型：On-line带权二部图最大匹配问题。

此模型，本文提供了3种解决方法，将重点介绍Google员工提到的折中算法。

此折中算法的最优化结果，在搜索引擎“仅在广告被用户click的情况下才有收益”的假设下，依然可以达到Competitive Ratio下限为1-1/e（第3节会详细解释什么是Competitive Ratio）。

本文的第6节会在了解此算法的基础上，对实际的广告竞价进行简单模拟，以给使用竞价服务的广告主提供参考。

1. 二部图最大匹配

二部图(Bipartite Graph)又称作二分图，是图论中的一种特殊模型。设G=(V,{R})是一个无向图。如顶点集V可分割为两个互不相交的子集，并且图中每条边依附的两个顶点都分属两个不同的子集。则称图G为二部图。

典型的例子，如《相约星期六》的男嘉宾和女嘉宾，就可以组成一个二部图。(今天的中国，《相约星期六》这样的节目还不支持男男配和女女配，因此“每条边依附的两个顶点都分属两个不同的子集”可以完全满足。)

二部图的最大匹配，是指找到一个子集M，使M的边集{E}中的任意两条边都不依附于同一个顶点，并且这样的边数最大。

再拿《相约星期六》举例：

例1

女1对男1、男2都有兴趣，女2仅仅对男1有兴趣。
如果把男1同时配对给女1和女2，两位女嘉宾的要求都能满足（不考虑争风吃醋的情况），但因为有“两条边(女1男1边、女2男1边)都依附于同一个顶点(男1)”的情况，所以不是一个匹配。
此二部图有且仅有以下2种匹配：
1. 男1 配 女1；女2没人配
2. 男1 配 女2；男2配女1
显然，匹配2是此图的最大匹配，它满足了边数最大（2条边）。

由例1看出，求一个图最大匹配的一种简单算法是：先找出全部匹配，然后保留匹配数最多的。但是这个算法的时间复杂度为边数的指数级函数。因此，需要寻求一种更加高效的算法。

在1955年，Harold Kuhn在两位匈牙利数学家Dénes Kőnig和Jenő Egerváry的研究基础上，提出匈牙利算法(Hungarian Algorithm)，通过找增广路径(Augment Path)，

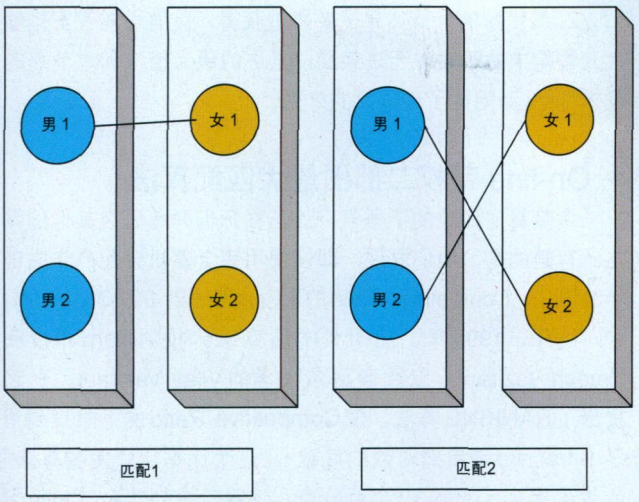

图1 例1附图

并将之取反后加到原匹配中的方法，得出最大匹配。

目前匈牙利算法是二部图最大匹配问题的最有效解法，其详细步骤可参见http://tinyurl.com/HungarianAlgorithm，本文不再赘述。

二部图匹配问题虽然应用广泛，甚至应用在图像识别领域，但它是怎么与“竞价排名”最优化联系起来的呢？这就要使用复杂一点点的带权二部图最大匹配模型 。

2. 带权二部图最大匹配

带权二部图，是指二部图中连接每个顶点的边，都有一个权值。

设想以下二部图：顶点子集U包含所有软件部员工，子集V包含软件部要完成的工作，某个员工完成某项工作的效益为C_{UV}(权值)，每个C_{UV}都不一定相等。此二部图为带权二部图。

带权二部图的最大匹配问题，是指找到一个匹配，使二部图总的权值和最大（或最小）。用上文软件部员工的例子，就是要求出如何分配员工工作，使总的工作效益最大。

我们来看，此问题已经与广告的关键字竞价排名非常相似了（后文会详细说明为什么不是“完全一致”）：

例2

广告1为关键字“手机”出价￥1.00，为关键字“照相机”出价￥1.50

广告2为关键字“手机”出价￥2.00，为关键字“照相机”出价￥1.00

用户搜索“照相机”、“手机” 2个关键字。

假设用户每搜索一个关键字时，仅出现一个广告。并假设每个广告对此用户仅出现一次（基于内容丰富性考量）。

此二部图的匹配有且仅有：

1. 照相机 -> 广告1；手机 -> 广告2。此时搜索引擎总收益：￥3.50

2. 照相机 -> 广告2；手机 -> 广告1。此时搜索引擎总收益：￥2.00

因此，匹配1为此带权二部图的最大匹配，因为其总收益最大。

计算一个带权二部图的最大匹配，可以像例1那样，使用穷举法。同样，其时间复杂度非常高，为O(n!)。我们同样需要一个高效的算法。

1957年，James Munkres提出了KM算法(Kuhn-Munkres Algorithm)，通过逐次调整每个顶点可行顶标的方法，将计算最大匹配的时间复杂度降到O(n3)。KM算法的详细步骤和C++源码请参阅http://tinyurl.com/KMAlgorithm。

KM算法还有一些扩展应用，例如计算如何使带权二部图总权的积最大（不是和最大），可参考http://www.byvoid.com/blog/match-km/。

根据本节的例子，KM算法看来是解决如何最大化搜索引擎收益的出路了。但是在实际应用中，工程师们发现了一个更加复杂的问题：

用户的搜索关键字，到达的次序是不可预知的。而且每个关键字，在到达时必须至少分配一个广告来响应，而且这种分配不可逆转。

解决这个问题，需要先了解On-line / Off-line问题。

3. On-line / Off-line 问题的区别与研究

二部图匹配的On-line问题，是指具有两个条件限制的二部图最大匹配问题：

1. 顶点、和顶点需要匹配的属性，到达次序是不可预知的。

2. 在指定两个顶点之间的匹配关系后，此匹配操作不可撤销、不可逆转。

满足这两个条件的相反情况的二部图最大匹配问题，即为Off-line问题。

On-line二部图，由于在求最优解的时候缺少对“全局”的把握，无法预知增广路径，因此不但KM算法不适用，就连其“最优解”（实际上是最优解的近似解）的权值和也总是小于（或等于）Off-line Solution下的权值和。

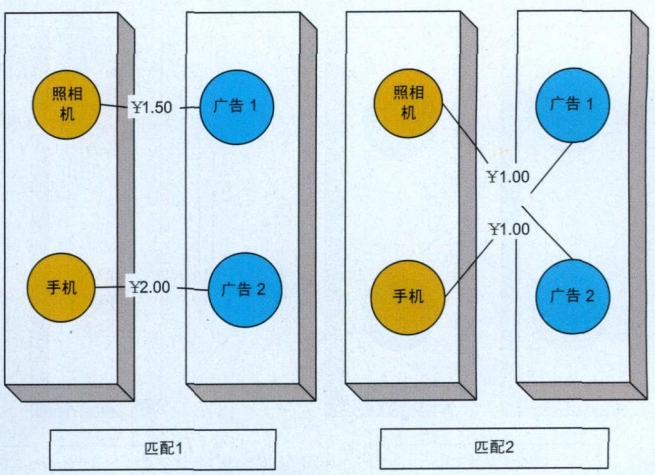

图2 例2附图

我们使用贪婪算法，为关键字分配广告，作为实例来看一下为什么会小于等于：

例3

广告1为关键字“手机”出价¥2.00，为关键字“照相机”出价¥3.00。
广告2仅为关键字"照相机"出价¥2.00。
用户搜索“照相机”、“手机”2个关键字。
假设用户每搜索一个关键字时，仅出现一个广告。并假设每个广告仅出现一次（基于内容丰富性考量）。
此带权二部图的最优匹配为：广告2 配 照相机；广告1 配 手机。此时搜索引擎的收益为最大收益：¥4.00
设想，如果使用搜索引擎的用户，第一时间先搜索关键字“照相机”。此时，因为搜索引擎并不知道之后是否会有其他用户搜索“手机”，无法事先排好最优解，所以只能使用其他算法来选择这时应该显示哪个广告作为响应。
此处使用贪婪算法，即挑选为此关键字出价最高的广告。广告1因为出价高，得以显示。
之后，该用户在很短的时间间隔后再次输入“手机”关键字，基于内容丰富性考量，广告1短时间内不能再显示（第5节会详细说明为什么这个限制不存在也不会对结果有影响）。但是，广告2并未对“手机”关键字出价，因此此时无法显示广告2。
这种情形下，搜索引擎的收益仅为：¥3.00，小于¥4.00。

经过研究发现，On-line问题的近似最大收益$C_{(Online)}$与使用相同参数Off-line问题的最大收益$C_{(Offline)}$总是满足如下等式：

$$C_{(Online)} \leq c \cdot C_{(Offline)} + b$$

式中，b为常数，c被称为Competitive Ratio。

可以通过简单的证明，得出在使用贪婪算法时，Competitive Ratio的下限是1/2。换言之，即使使用贪婪算法，On-line问题优化以后的“最大收益”，可能只有实际最大收益的一半左右。

Competitive Ratio在其他On-line优化问题中也存在，如K服务器问题(K-Server Problem)、Steiner Tree(斯坦纳树)。其出现的原因主要有两点：

1. 未来的输入参数不可预期，因此可能“有敌意”的“对手”输入先出现，阻扰了最优解的产生。

2. 最优解的计算如果需要累进获得，就有可能在安排输入元素需求的时候，无法兼顾“累进的最大值”和“节点的需求”，从而阻扰了最优解的产生。

4. On-line 带权二部图最大匹配算法

贪婪算法如此的不争气，让工程师们开始研究其他的算法。有趣的是，他们发现，即使使用完全随机分配广告关键字的做法，Competitive Ratio的下限也能达到1/2+O(log n/n)。

终于，1990年，加州大学伯克莱分校的Karp和校友Umesh Vazirani，以及康耐尔大学的Vijay Vazirani，一起提出了RANKING算法，使Competitive Ratio的下限提高到了1-1/e。(e是自然对数的底数)。这个比率比“贪婪算法”的效果提高了大约12个百分点。该篇论文以“An Optimal Algorithm for On-line Bipartite Matching”为题发表在ACM的官方网站上，有兴趣的读者可以自行下载阅读，本文不作重点介绍。

我们持续变换的世界再次给我们的算法研究工程师发出了挑战：我们的广告投放者希望能设置自己当天的最高预算，在此预算被安排用尽之后，即使再有合适的关键字到来，我也不希望再安排我的广告显示了。Kalyanasundaram和Pruhs在2000年，提出并解决了这个On-line b-matching问题：每个广告主每日最高预算设定为b，他们给每个关键字出价为0或1。他们的算法，BALANCE算法，把每个到来的关键字分配给当日预算余额还剩下最多的那个对此关键字感兴趣的广告。最终他们证明了，当预算b趋向于无穷大时，Competitive Ratio的下限也是1-1/e。

5. Google 的算法

2007年8月，Google的员工Aranyak Mehta与斯坦福大学的Amin Saberi，以及当初提出RANKING算法的两个Vazirani(此时的Vijay Vazirani已经进入了乔治亚理工大学)，在BALANCE算法的基础上，结合Factor-Revealing线性规划方法，共同提出了他们自己的折中算法。该算法的关键是：在关键字出价和未使用预算余额（比例）之间，找到一个最佳的平衡点。

该算法的具体描述为：

最新到来的关键字，应该匹配给V值最大的那个广告。
其中：
$V = c(i) \times \Psi(T(i))$
其中：
$c(i)$ = 该广告为该关键字i的出价
$\Psi(x) = 1 - e^{(x-1)}$ (e为自然对数)
$T(i) = mi / bi$
其中：
mi：目前为止该广告的竞价者当日已经实际用掉多少钱
bi：该广告的竞价者当日总预算

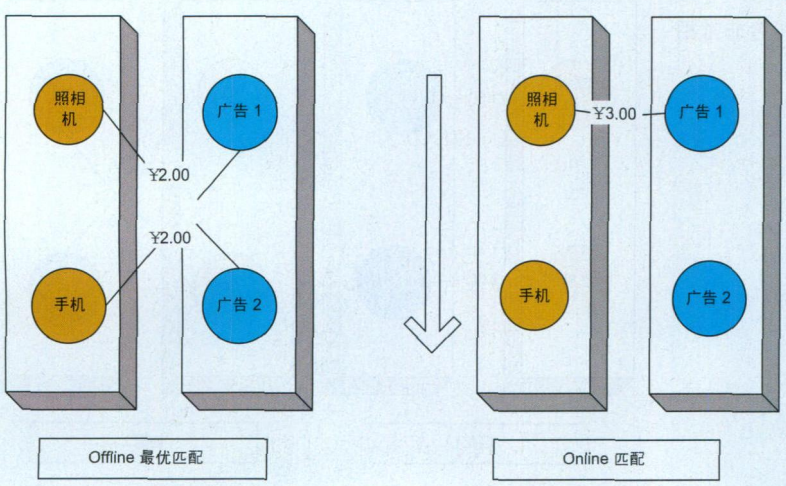

图3 例3附图

细心的读者可能发现，在例3中，我们假设“每个广告仅出现一次”的原因是“基于内容丰富性考量”。这确实是一个合理的考量，同一个广告如果在1分钟内反复出现，这将成为信息轰炸，使广告的效应大打折扣。但是同样，我们假设这样一种情况：每个广告在此出价后，它的所剩余额都将全部用完。换言之，可以将这次匹配视为该广告的最后匹配机会。 这样的考量同样也能得到“每个广告仅出现一次”的限定条件。

但是，该算法不会仅限应用于“在预算快用完”，或者“每个广告只能出现一次”的情况。折中算法对以下情况均适用：

1. 广告竞价者各自有不同的当天总预算。

2. 本次最佳匹配并不会用完广告竞价者的所有当天总预算余额。

3. 广告竞价者可以在不同的时期加入游戏，提出竞价。

4. 每一个关键字可以匹配n（n>1）个广告，使n个广告同时出现。（只需在关键字到来时选择V值最高的n个广告显示即可，V越高排位越前。）

5. 每个竞价者只在有用户点击他的广告后才产生费用。

最后一点非常重要，无论是Google还是百度的竞价广告，广告主只有在有用户真正点击广告链接后才记一次成功出价。这个特性是使得搜索引擎提供的广告比传统广告性价比更高的最根本原因。

要实现这一点上的最优匹配，需要在原算法上做一个小补充：

```
V = c(i) × ψ(T(i)),  c(i) = 该广告为该关键字的出价；此处的c(i)替换成c(i)',
其中：
c(i)' = c(i) × CTR，其中：
CTR = 该广告历史点击量 / 该广告历史显示总次数
（注：新进入系统、从未播放过的广告，在第一次候选时会有“冷起点”的问题，此时可以简单将CTR定为50%或更高。设得越高，越容易帮助新广告更快融入系统。）
```

即使在以上5点状况下，使用了该算法后，Competitive Ratio的下限也能达到1-1/e。(证明略，有兴趣的读者可以email笔者索要证明方法。)虽然性能上比BALANCE算法并没有任何提升，但是毕竟，此折中算法提供的实现方案，其实用价值高多了。

6. 给广告关键字竞价者的一些参考

基于第5节的公式，笔者模拟了一些数据。最后一列的V值越大，越容易赢得展示的机会（此处不考虑关键字的语义匹配程度）。

表1 基于第5节公式的数据模拟

竞价者	关键字出价	当日已使用金额	当日总预算	CTR	V
A	¥1.00	¥10.00	¥100.00	50%	0.29671517
B	¥1.00	¥30.00	¥100.00	50%	0.251707348
C	¥3.00	¥10.00	¥100.00	50%	0.89014551
D	¥3.00	¥66.00	¥100.00	50%	0.432344516
E	¥1.50	¥10.00	¥100.00	50%	0.445072755
F	¥0.10	¥0.10	¥1,000.00	50%	0.031604188
G	¥2.00	¥2.00	¥100.00	50%	0.624688901
H	¥1.00	¥10.00	¥100.00	80%	0.474744272

通过分析此表数据，可以得出以下几个结论

1. 比较竞价者A和B，可以看出：在相同关键字出价的情况下，当日预算余额多的广告优先得到展示的机会。（此点可以理解为：之前花多了这次就先帮你省着点吧，万一一会儿还要再用你钱呢。）

2. 比较竞价者A和C，可以看出：在当日预算所剩比例相同的情况下，关键字出价高的广告优先得到展示机会。

3. 比较竞价者D和G，可以看出：即使关键字出价低（不能太低，参照结论4），只要当日预算的余额足够大，还是有希望打败关键字出价高的广告赢得优先展示的机会。

4. 比较竞价者F和B，可以看出：在当日预算所剩比例即使相差再悬殊（3000倍），关键字出价高（10倍）还是更容易赢得广告优先展示的机会。

5. 比较竞价者A和H，可以看出：其他变量保持不变的情况下，广告历史的CTR数据是决定是否再次被展示的决定性因素。

6. 比较竞价者E和H，可以看出：即使关键字出价不如别人，广告的历史CTR表现也会为自己赢得更好的被展示的机会。（基金经理总是喜欢说：该基金的历史表现不代表它的今后表现。在Google看来可能不是那么回事。）

参考资料

1. Aranyak Mehta, Amin Saberi, Umesh Vazirani, Vijay Vazirani, R.2007. AdWords and Generalized On-line Matching

2. Alexa Sharp. Thoughts on the Competitive Ratio

3. Richard M. Karp, Umesh V. Vazirani, Vijay V. Vazirani, R 1990. An Optimal Algorithm for On-line Bipartite Matching

4. Wei, Shih-Yi. On-line Problem

5. 佚名. BipartiteGraph.ppt

作者简介

陈文雄（Vincent Chen），Cereson.com董事、软件设计总监。从事人工智能、商业智能、云计算、DVD自动租售机等方面的设计和研究。邮件地址为vincent.chen@cereson.com。

■ 责任编辑：郭晓刚 (guoxg@csdn.net)

Contents

http://www.programmer.com.cn

Programmer 程序员

2009年09月刊 总第203期

主管：中国社会科学院
主办：中国社会科学院文献信息中心
出版：《程序员》杂志社
网址：http://www.programmer.com.cn
国际刊号：ISSN 1672-3252
国内刊号：CN11-5038/G2
邮发代号：2-665
广告经营许可证号：京东工商广字0188号

总编：黄长著 Editor-in-chief：Huang Changzhu
社长/常务副总编：张悦校 President：Zhang Yuexiao
副社长：蒋 涛 Vice President：Jiang Tao
编委会：黄长著 张悦校 陈洋彬 蒋涛 曾登高 韩磊
Editorial Member：Huang Changzhu Zhang Yuexiao Chen Yangbin Jiang Tao Zeng Denggao Han Lei

执行主编：孟迎霞 Executive Editor-in-chief：Meng Yingxia
编辑部主任：孟迎霞（兼） Director：Meng Yingxia
编辑部副主任：欧阳璟 Deputy Director：Ouyang Jing
责任编辑：郑柯 周至 李雨来 郭晓刚
Editors：Zheng Ke Zhou Zhi Li Yulai Guo Xiaogang
特邀编辑：方梁 高昂 常政 赵健平 彭一凡 吕娜
Contributing Editors：Fang Liang Gao Aang Chang Zheng Zhao Jianping Peng Yifan Lv Na
美术设计：纪明超 Art Designer：Ji Mingchao
美术编辑：吴志民 Art Editor：Wu Zhimin
Tel：010-64351458
Email：editor@csdn.net

发行部 Distribution Dept.010-64351431
Email：sales@csdn.net

广告总代理：北京创新乐知广告有限公司
Sole Advertising Agency：Bejing CSDN Co.,Ltd
Tel：010-64376055
Email：ad@csdn.net
Marketing Dept：010-51661202（ext 149）
Email：market@csdn.net

读者服务部
Readers service Dept.
网上订购：http://book.csdn.net/programmer/
读者信箱：reader@csdn.net
地址：北京市朝阳区酒仙桥路14号兆维工业园B区3楼2门1层
Address: B3-2-1F, Zhaowei Industry Park, No. 14 Jiuxianqiao Road, Chaoyang Dist, Beijing
邮政编码：100015
电话：010-64351436
传真：010-64348545

法律顾问：北京中润律师事务所 王杰
Law Consultant：Beijng Hengsheng Lawyer Firm
印刷：北京盛通印刷股份有限公司
Print: Beijing Shengtong Printing Co., Ltd.
出版日期：每月1日
Publication Date：the first day per month
零售价：RMB 15.00元 新台币 390元 HK $ 35.00（港、澳）US $ 9.00（海外）
Retail Price：RMB 15, NT$390, HK $ 35.00,US $ 9.00