杨焜 (/users/22738) 2016-05-30 14:31:22 发表于： 阿里妈妈技术团队 (/teams/12) >> 算法 (/teams/12?cid=63)

知识体系： 机器学习 (/articles/?kid=201) 修改知识体系

文章标签： 算法 (/search?q=算法&type=INSIDE_ARTICLE_TAG) 修改标签

标签历史 (/articles/55562/tags/history)

附加属性： **内部资料请勿外传** **作者原创** **热门推荐**

直通车搜索广告Matching的蜕变之路

在过去的一年里，直通车搜索广告Matching方向经历着翻天覆地的变化，在QR改写上我们提出了新的改写算法Nodebidding完全替换了旧改写算法，在粗排上我们从单库单通道演进到了多库多通道，这些优化持续输出结果助力搜索广告业务的发展。这篇文章讲述了我们的心路历程，其中包含两篇外链详细介绍了算法的细节部分。

背景

在直通车搜索广告场景下，广告的召回可以分为两个阶段，第一个阶段从query改写到bidword，这个阶段称为query rewrite；第二个阶段从bidword的倒排链条里召回ad，这个阶段一般称为粗排(ad selection)。在KGB系统中粗排又分成两个阶段：Pre ranking(海选)和First Ranking(初选)。如图1所示，海选阶段是直接对倒排链的排序，一般输入规模为5－10万ad，对性能要求较高，只能做有限的计算，海选后输出700个ad进入初选，初选可以做一些更复杂的排序。

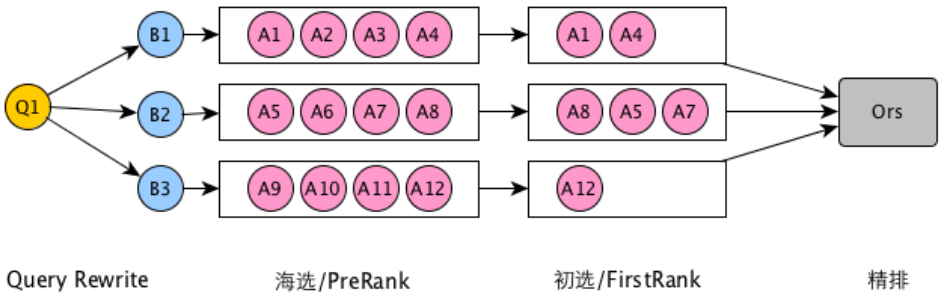

图1 广告召回路径（不包含Ors部分）

Query Rewrite可分为两种：精确匹配和模糊匹配。精确匹配要求query和bidword完全相同，模糊匹配只要query和bidword相关即可，如图2所示。在召回环节，算法可以发挥力量的环节的主要在模糊匹配、海选和初选。

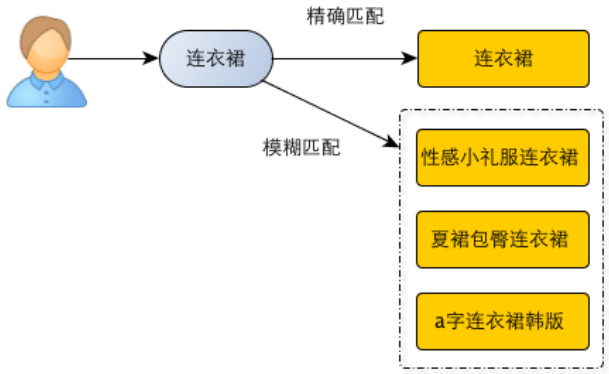

图2 qurey rewrite两种匹配方式示意图

Query Rewrite 常用方法

Query Rewrite是一个成熟的研究领域，从2000年以后开始有大量的研究，在2006-2009年有个高峰，近几年来研究热度有所下降。Query Rewrite要解决的问题是尽量召回更多更优质的候选集合，在网页搜索场景下，这个候选集是网页摘要和链接，在搜过广告场景下，这个候选集是广告。

过去研究的Query Rewrite方法我们大致分为四大类如表1所示。第一类是基于log挖掘的方式，利用日志构建query的关系图，再加上图的扩展和约束，得到最终关系，这类方法是工业界主力采用的方法。例如基于session log的query flow graph方法、基于click log的simrank/simrank++方法等；第二类是基于relevance feedback的方法，包括explicit/implicit/pseudo feedback，explicit方式一般采用人工标准作为样本，通过machine learning方式构建模型，pseudo方式会引入搜索结果作为反馈，采用迭代的方式构建改写集合；第三类是基于语义的方式，例如同义词替换、增加/删除部分词、提取中心词等，这也是业界常用的方法；第四类是基于隐式空间的方式, 分别把query和bidword映射到特征空间，然后通过cosine距离的方式来计算相似度，例如LSA/LSI，WordToVector类型的各种变种，以及一些最新的Deep Learning技术在这个领域的应用。另外还有一些比较小众的方法，例如translation model的应用等，这里不再细列。

表1 query rewrite常见方法

类型	典型方法
基于 **log** 挖掘方式	Query flow graph, simrank, simrank++
基于 **relevance feedback** 方法	Explicit/implicit/pseudo feedback
基于语义方式	同义词替换，增删词，中心词提取等
基于隐式空间方式	LSA/LSI, WordToVector, DeepLearning

Nodebidding亮点

前面提到的种种解法，多数是应用在网页搜索场景下，往往主要聚焦在query和bidword的相关性上，然而在搜索广告场景下，仅仅考虑相关性是不够的，还需要考虑平台的收益，我们的目标是相关性保证的前提下，尽量最大化平台的收益。因此，我们提出了一种比较新颖的改写算法Nodebidding，在最近一年多里在这个目标的指导下我们陆续进行了六个大版本的优化，都取得了大两位数rpm增长的不错效果。如图3所示。

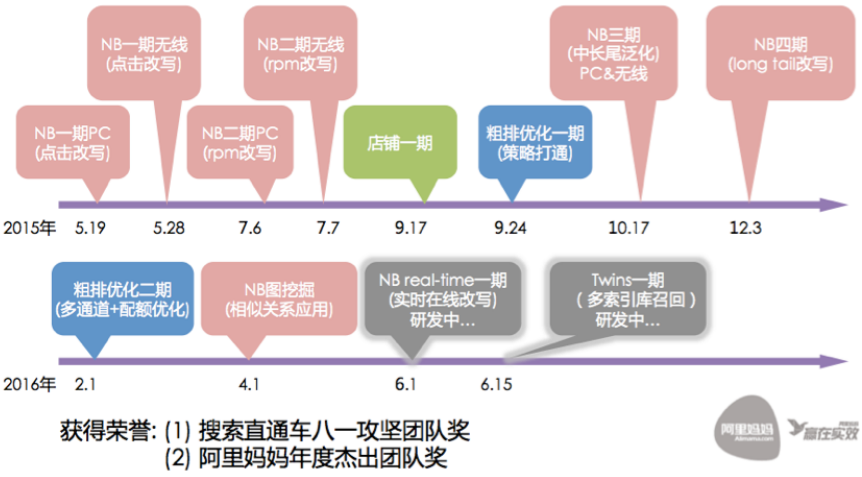

图3 NB&粗排优化milestones

Nodebidding思想

在Nodebidding的算法体系里，我们对QueryRewirte问题进行了抽象，问题求解定义如下：

从一个三层图 G=<Q, A, B>里构建出一个两层子图 G'=<Q', B'>，使得在满足一定约束条件下，平台能够获取最大的 rpm 收益。

符号说明：

Q：表示一天全部 query 集合，天级别为 3 千万，周级别为 1 亿。

B：表示所有的候选 bid word 集合，数量约 7 千万。

A：表示所有的 ad(广告)集合，数量约 4 千万。

Q'：表示改写算法需要覆盖的 query 集合，其中$Q' \subseteq Q$

B'：表示 Q' 的 rpm 最优改写集合，其中$B' \subseteq B$

$E_{q,a} \subseteq Q \times A$ ：表示 query 和 ad 的关系边，边含义表示 query 下可以展示的 ad。

$E_{a,b} \subseteq A \times B$ ：表示 ad 和 bidword 的关系边，边含义表示 ad 购买的 bidword。

B_a：表示 ad $a \in A$ 购买的 bidword 集合。

$r_q(a)$ ：表示对于 query $q \in Q$ 下展示 ad $a \in A$ 的预期收益，

其计算如下：$r_q(a) = ctr(q,a) * \max_{b \in B_a}\{price(a,b)\}$， $r_q(a)$ 在图中的可以确定一条<q,a,b>路径。

$r_{q,d}(A')$：表示对于 query $q \in Q$ 下展示 ad 集合 $A' \subseteq A$ 的最优质 d 个广告的收益，

其计算如下：$r_{q,d}(A') = \max_{A_s \subseteq A', |A_s|=d} \sum_{a \in A_s} r_q(a)$。

$B_{q,d}(A')$：表示 $r_{q,d}(A')$ 对应的 bidword 集合。

$A_{q,d}(A')$：表示 $r_{q,d}(A')$ 对应的 adgroup 集合。

约束条件：

(1) 出于检索性能的考虑，每个 query 所能改写的 bidword 数量是有限制的，我们通过 bidword 的总 ad 深度来控制。表示为

$|\{a | a \in B_{q,d}(A')\}| \le f(q)$ ，其中 f(q)是根据 q 热门度的一个函数

(2) 每个 ad 的每天预算是有限的，预算消耗完后会下线，ad 不能无限次被用于计算。表示为

$\sum_{q \in Q, a_i \in A_{q,d}(A')} r_q(a) \le buget(a_i)$ ，其中 $buget(a_i)$ 为 a_i 一天的投放预算

目标函数：

$\forall Q' \subset Q$，$R(Q')$定义为 Q'集合改写可能获得的收益，求解目标为使得$R(Q')$最大，

表达如下：

$$R(Q') = \sum_{q \in Q'} r_{q,d}(A') \left\{ \left| \{a | a \in B_{q,d}(A')\} \right| \leq f(q), \quad \sum_{q \in Q', a_i \in A_{q,d}(A')} r_q(a) \leq buget(a_i) \right\}$$

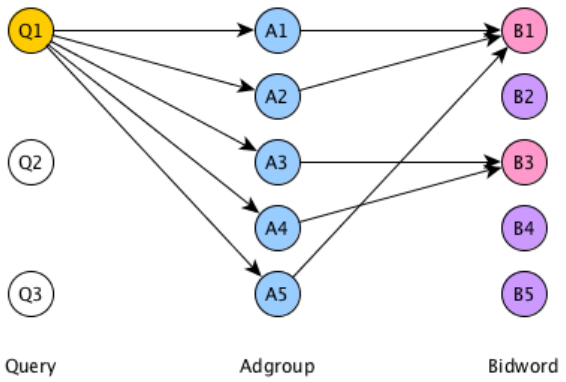

图4 G=<Q,A,B>示意图

考虑到真实环境下ad-bid价格、购买关系的动态变化、地域投放、分时折扣等因素，上述问题的求解和计算量是非常复杂的。因此我们进行了一定简化，另外针对长尾和中高频Query采用了不同的处理方式。Nodebidding系列算法都是在对问题简化的基础上，不断逼近所能达到的最优解。

详细求解方法请参考这篇文章：

Nodebidding：一种高效的搜索广告查询改写方法 (http://www.atatech.org/articles/55560)

粗排的优化

搜索广告是白盒广告系统，广告主可以为ad任意购买bidword，这也导致bidword下ad可能会有各种情况，有可能并非完全与bidword相关。在现有的KGB引擎架构下，倒排链中每个ad的payload存储的信息非常有限，出于检索性能和存储容量的限制，在ad selection阶段只能进行比较有限的计算，无法在这个阶段考虑更多的因素，例如无法实时计算ad和query的品牌匹配度信息。此外排序过程中使用<bidword, ad>预估分和<query, ad>预估分也有偏差。因此，目前的ad selection机制的效率并不高。

对粗排的优化我们始终本着一个出发点：尽量减少这个环节的信息损失。采用的手段主要是多通道方式，包括海选多通道和初选多通道。海选多通道出于性能的问题合并到多库索引里解决。优化前后的示意图如5所示。

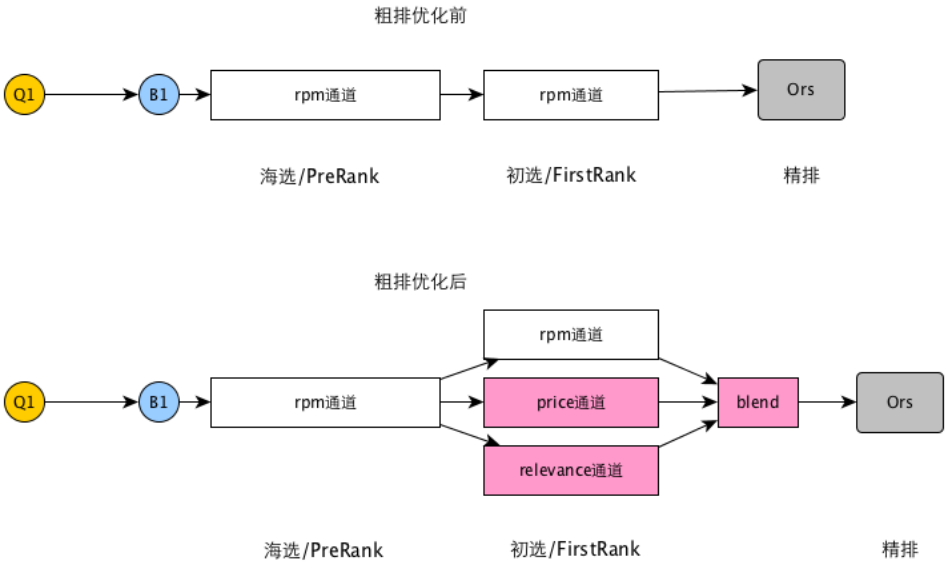

图5 粗排优化前后的示意图

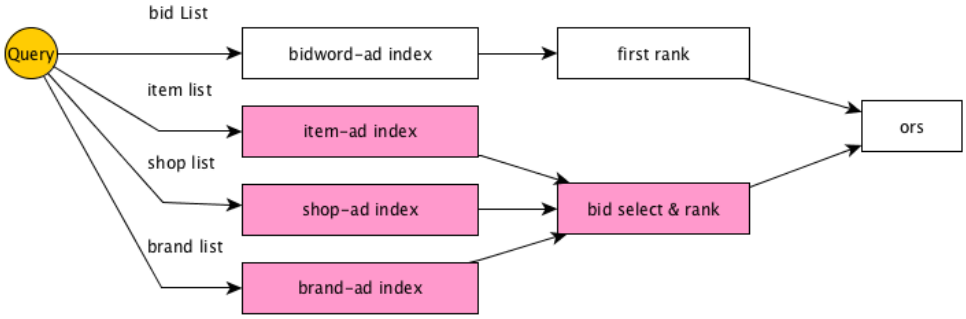

图5 多库召回示意图

多库和初选优化细节可以阅读这篇文章

KGB Matching粗排的优化历程 (http://www.atatech.org/articles/55561)

未来的Match

业务总是在不断的发展，需求不停的涌现，在最近半年内我们的努力方向主要在个性化、实时和多库。个性化是指在召回阶段考虑用户的近期和长期个性化行为；实时是指改写算法依赖信息更新做到实时化，未来数据更新流有三种组成：天级别的offline更新，小时级别的nearline更新和实时的realtime更新；多库本质上是一种多通道机制，除了bidword库，会增

加item库、shop库和品牌库等。目前初步的框架已经形成，借用@撼林同学的架构图来表达，如图7所示。

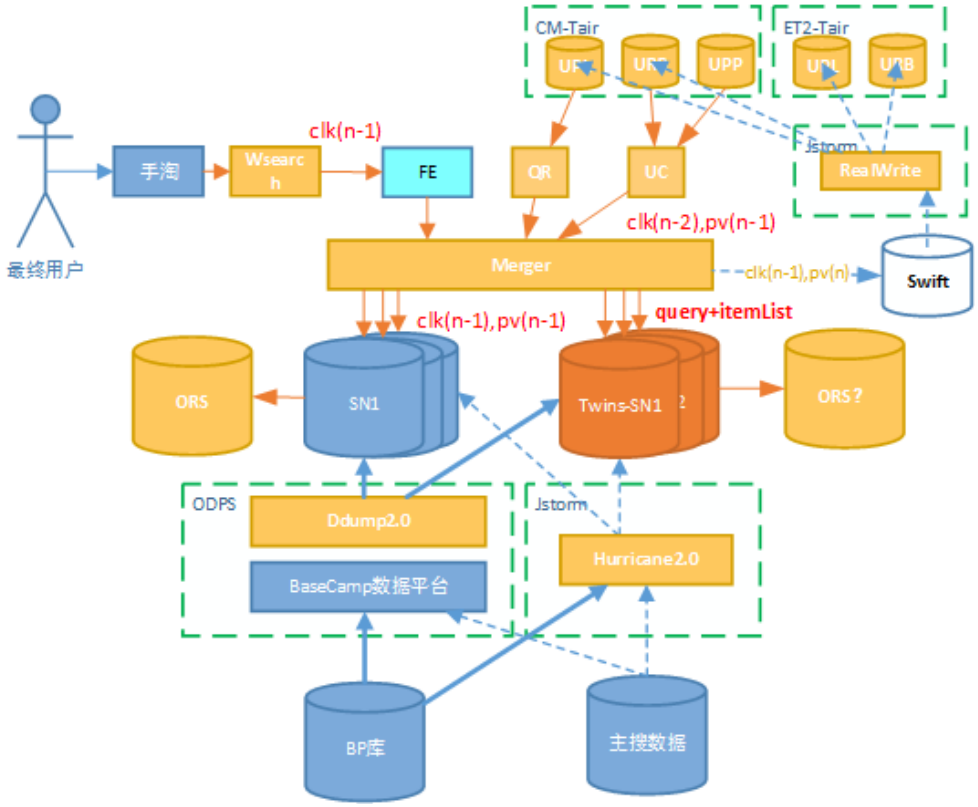

图7 实时化、个性化和多库框架

再远的未来，我们还没有细致的计划，不过有一点可以预见是要做的：全流量上的全局优化。目前我们的优化算法都是基于单个query的考虑，可以理解为贪心方式，未来需要全局优化求解，这是一个困难问题。

致谢

一路走来，我们并不是一帆风顺，踩过坑、经历过低谷，可贵的是我们坚持下来了，并最终拿到了不错的结果。特别感谢老大吴波和治平的指导和支持，感谢我们项目团队的每一个人！感谢见独团队衍悔同学在多库项目中不分昼夜的提供索引支持，感谢模型组闻风为我们离线预估的帮助，感谢K2小致团队为我们提供高质量的item2ad数据。

为NB和多库项目工作的成员如下：

算法：治平、士铎、国俊、杨焜

工程：洛歌、叶卿、撼林、衍悔

产品：万程、芙嫄

评测：红梅、我型

PE: 宗道、柄风

QA: 刑志、飞索、依颜、韩萱

运营：依人、蔡妍

相关文章

Nodebidding：一种高效的搜索广告查询改写方法 (http://www.atatech.org/articles/55560)
KGB Matching粗排的优化历程 (http://www.atatech.org/articles/55561)

他们赞过该文章

乐迪 (/users/6185) 西风 (/users/9439) 关冕 (/users/10123) 万程 (/users/12689) 赵印 (/users/13740)
兵乙 (/users/16072) 闻风 (/users/20059) 放歌 (/users/22970) 潮汐 (/users/25356) 依颜 (/users/30627)
沧睿 (/users/30737) 东垣 (/users/34829) 无当 (/users/64400) 雕侠 (/users/64571) 刑志 (/users/65025)
荐轩 (/users/65560) 石士 (/users/66183) 随空 (/users/66754) 藤木 (/users/67095) 撼林 (/users/67644)
胡玲 (/users/67665) 淮阴 (/users/68186) 决辰 (/users/68300) 鸿祺 (/users/68391) 口肃 (/users/70089)
文樵 (/users/73814) 知煜 (/users/74832) 一鸥 (/users/75369) 五津 (/users/75772) 筱洋 (/users/78109)
庙算 (/users/80602) 宵夙 (/users/82205) 士铎 (/users/82330) 积流 (/users/92316) 霁光 (/users/92886)
飞索 (/users/99661) 静渔 (/users/130432) 淑菡 (/users/152627) 思召 (/users/153560)
慕隆 (/users/166781) 策羽 (/users/185363) 橦录 (/users/199401) 学正 (/users/207024)
旺京 (/users/210406) 弘照 (/users/254740) 北封 (/users/385665)

相似文章

- 阅读思考及摘要之”计算广告简介” (/articles/5919)
- NodeBidding : 搜索广告QR的浴火重生 (/articles/39144)
- OCPM：CPM计费方式下的智能调价技术 (/articles/73548)
- 计算广告及搜索广告简介 (/articles/26812)
- KGB Matching粗排的优化历程 (/articles/55561)
- 基于i2q的底纹个性化推荐 (/articles/76353)

上一篇：NodeBidding : 搜索广告QR的浴火… 下一篇：搜索广告个性化匹配之路 (/articles…

1F 博浩 (/users/75295) 2016-05-30 15:13:07

NB！

0 (/comments/92945/voteup) 0

2F 撼林 (/users/67644) 2016-05-30 16:52:27

NB项目果然NB！

0 (/comments/92966/voteup) 0

3F 乐田 (/users/5895) 2016-05-30 17:04:05

几经易稿，终于吐出了真货，大家可以来围观啦～

0 (/comments/92971/voteup) | 0

4F 怀人 (/users/72657) 2016-05-30 17:05:01

创新的想法，出色的效果，大赞！

0 (/comments/92973/voteup) | 1

探微 (/users/33443) 2016-06-02 11:32:48

被醍醐灌顶了吧～

0 (/comments/92973/subcomments/29224/voteup)

写下你的评论...

5F 止善 (/users/24907) 2016-05-30 17:26:05

先顶后看

0 (/comments/92982/voteup) | 0

6F 叶卿 (/users/20238) 2016-05-30 17:33:16

NB！

0 (/comments/92986/voteup) | 0

7F 沧睿 (/users/30737) 2016-05-30 18:10:34

强烈顶！

0 (/comments/92993/voteup) | 0

8F 刑志 (/users/65025) 2016-05-30 19:04:27

焜总提纲挈领，Matching大变样

0 (/comments/92999/voteup) | 0

9F 橦录 (/users/199401) 2016-05-30 19:09:42

赞

0 (/comments/93004/voteup) | 0

10F 墨丞 (/users/68284) 2016-05-30 19:10:50

nb的项目！

0 (/comments/93006/voteup) | 0

11F 决辰 (/users/68300) 2016-05-31 09:04:06

UP

0 (/comments/93029/voteup) | 0

12F 里仁 (/users/21473) 2016-05-31 09:05:27

已读完！赞！持续学习中

0 (/comments/93032/voteup) | 0

13F 淡昆 (/users/75778) 2016-05-31 09:39:43

NiuBility，学习

0 (/comments/93050/voteup) | 0

14F 协明 (/users/67100) 2016-05-31 10:38:11

厉害了！！！！学习！！！！！

0 (/comments/93076/voteup) | 0

15F 娇娇 (/users/3449) 2016-05-31 12:48:26

NB！期待后面的进一步发功，加油↖(^ω^)↗

0 (/comments/93088/voteup) | 0

16F 飞桐 (/users/10621) 2016-05-31 14:27:21

项目名就NB啊

0 (/comments/93096/voteup) | 0

17F 五津 (/users/75772) 2016-05-31 16:45:35

顶！

0 (/comments/93129/voteup) | 0

18F 见独 (/users/6460) 2016-06-01 15:53:57

精品

0 (/comments/93227/voteup) | 0

19F 雕侠 (/users/64571) 2016-06-01 16:07:21

牛！

0 (/comments/93236/voteup) | 0

20F 兵乙 (/users/16072) 2016-06-01 16:13:54

焜哥，NB！

0 (/comments/93240/voteup) 0

◀ 1 2 (/articles/55562/comments?p=2) ▶ (/articles/55562/comments?p=2)

写下你的评论...

评论

ATA社区常见问题汇总 (/articles/67826)