计算广告技术专栏

导 读

作为互联网健康发展的经济基础,在线广告在互联网时代扮演着重要的角色.根据易观智库商业信息服务平台提供的《中国互联网广告市场趋势预测 2012—2015》中的数据,2012 年中国互联网广告市场规模为人民币 724.8 亿元,2013 年将到达 1060 亿元.

计算广告(Computational Advertising)研究以计算的方式来提高在线广告的效率,是曾任雅虎实验室首席科学家的 Andrei Broder 在 2008 年召开的第 19 届 ACM-SIAM 离散算法学术讨论会上提出的.计算广告技术涉及大规模文本分析、信息检索、机器学习、统计模型和微观经济学等学科,研究的目标是实现语境、广告和受众三者的最佳匹配.计算广告技术试图针对三种主要的在线广告形式(搜索广告、显示广告和上下文广告),以及对广告系统的四个主体(受众、广告主、媒体和广告联盟网络),进行优化,从而使得广告投放活动的综合收益最大化.美国芝加哥大学的经济学家 Phillip Nelson 认为:广告也是一种信息,计算广告的目的是提高用户和相关广告的匹配,从而实现广告信息的精准投放.

华东师范大学软件学院在海量数据处理、数据挖掘、Web 数据管理和数据流技术等领域有着长期的技术积累;聚胜万合信息技术(上海)有限公司是国内从事精准营销及数字营销的专业广告技术服务机构,在互联网广告投放技术领域有着丰富的开发经验和领先的技术成果.双方共同创建了"华东师大-聚胜万合计算广告技术联合实验室",建立长期的合作关系,充分发挥双方各自优势,对计算广告技术进行联合研究;组织产学研结合的科研团队,推进双方在计算广告技术领域的研发工作.联合实验室通过相关的课程建设、研发课题以及研讨会等形式在学术界和工业界之间建立沟通的桥梁,来推动计算广告技术的发展和应用.

2012 年 5 月,来自复旦大学、中国人民大学和美国新泽西州立大学等国内外十多所高校的专家学者,以及来自微软、IBM 和百度等近十家企业的科研人员,共计四十余人参加了联合实验室组织的计算广告技术研讨会.研讨会的目的是促进有关高校在计算广告研究这一新兴领域的交流和合作,为互联网广告企业和高校之间的产学研联合提供平台,以推动科研成果快速转化成生产力.与会专家就广告投放系统架构、广告点击率预测、用户定向和移动推荐等问题进行了讨论.会后,在联合实验室的组织下,参会的学者将讨论的成果整理成文.经过有关专家的认真审查,这些论文被推荐在《华东师范大学学报》(自然科学版)上发表;以期进一步推动计算广告技术在国内广泛传播和深入发展.

文章编号:1000-5641(2013)03-0002-13

广告点击率估算技术综述

纪文迪[1], 王晓玲[1,2], 周傲英[1,2]

(1. 华东师范大学 软件学院 上海市高可信计算重点实验室,上海 200062;
2. 复旦大学 上海市智能信息处理实验室,上海 200433)

摘要: 计算广告是根据给定的用户和网页内容,通过计算得到与之最匹配的广告并进行精准定向投放的一种广告投放机制.广告的点击率预测是指利用点击日志预测的点击率,其结果受到广告的自身性质、广告位置、页面信息、用户性质,以及广告主信誉等诸多因素的影响.有效地预测广告的点击率,对于提高广告投放的效率有着至关重要的作用.本文介绍了广告点击率预测的常用模型,包括历史数据丰富的广告点击率预测模型、新广告和稀疏广告的点击率估算模型和点击率预测的优化模型,并通过真实数据集举例说明了其实现的方法.

关键词: 计算广告; 点击率估算; 逻辑回归模型; 贝叶斯方法

中图分类号: TP391 **文献标识码**: A **DOI**:10.3969/j.issn.1000-5641.2013.03.001

Techniques for estimating click-through rates of Web advertisements: A survey

JI Wen-di[1], WANG Xiao-ling[1,2], ZHOU Ao-ying[1,2]

(1. *Software Engineering Institue, Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai* 200062, *China*;
2. *Shanghai Key Laboratory of Intelligent Information Process, Fudan University, Shanghai* 200433, *China*)

Abstract: Computational advertising is a kind of advertising mechanism which has the capability to find the most suitable ads for given users and web content, so as to advertises them accurately. Therefore, estimating click-through rate (CTR) precisely makes significant difference in the efficiency of advertising on the Internet. Ad click-through rate prediction is to estimate CTR with click log, which is influenced by the nature features of ad, the position, the page information, user properties, the reputation of advertisers and such other factors. This paper is aimed to illustrate useful CTR prediction models, including CTR models for ads of abundant history data, CTR models for rare ads or new ads and some optimization models. Finally, the implementation methods with real data set were demonstrated as examples.

收稿日期:2013-03

基金项目:工信部核高基项目(2010ZX01042-002-003-004);国家自然科学基金重点项目(61033007);国家 973 课题(2010CB328106);教育部新世纪人才支撑计划(NCET-10-0388);创新研究群体科学基金(61021004)

第一作者:纪文迪,女,硕士研究生,研究方向为数据库. E-mail:51111500007@neu.edu.cn.

通信作者:王晓玲,女,教授,博士生导师,研究方向为 Web 数据管理环境. E-mail:xlwang@sei.ecnu.edu.cn.

Key words: computational advertisement; click-through rate; logistic regression; Bayes method

0 介　　绍

计算广告(CA, Computational Advertisement)是根据给定的用户和网页内容,通过计算得到与之最匹配的广告并进行精准定向投放的一种广告投放机制. 采用该机制可以大幅度地提高广告主所投放广告的点击率(CTR,Click-Through Rate),增加广告所投放网站的访问量,帮助用户获取优质信息,从而构建良性和谐的广告投放产业链[1].

计算广告的一种形式是赞助商搜索(Sponsored Search),其广告投放的目标位置是搜索引擎所返回的搜索结果页面. 在赞助商搜索的场景中,搜索引擎既充当了网络媒体,也充当了广告网络. 因此,赞助商搜索便成为广告主、用户和搜索引擎三方的一个博弈过程,博弈的目标是要使三方的总收益(Payoff)最大[1]. 其中,广告主要求更多用户来点击广告,搜索引擎要求得到最多的收益(Revenue),用户要求被推荐更适合自己的广告来提高其自身的体验[3-4].

另一种形式是情景广告(Contextual Advertising)或叫内容匹配(CM, Content Match),是指在一般的网页上根据上下文投放商业文本广告或图片广告. 通常广告联盟(Ad-Network)会参与到这种广告投放形式中,充当第三方的角色,其作用是通过广告选择实现提高收益和用户的满意度.

各种广告投放形式中的广告投放者的愿望都是提高自己的收益,这就需要提高广告的 CTR. 对于 CTR,我们关注对它的预测,只有对 CTR 有准确的预测,才能及时地在查询返回页面投放相应顺序的广告. 以搜索引擎为例,搜索引擎对用户可能查询的关键词进行竞拍(Auction),广告主根据自己的具体经营情况来竞拍这些关键词. 目前广告主的主要付费方式为点击付费(Pay Per Click);若单位点击的付费额记为 CPC(Cost Per Click)[1-2],则搜索引擎的收益(Revenue)是 CTR×CPC. 研究显示,用户点击广告的可能性按广告的排放位置快速递减,最高可达 90%[6];搜索引擎想获得最大的收益就需要把 CTR×CPC 大的广告投放在靠前的位置,并依据相乘的结果对广告在查询返回页面上进行排序[5]. 因此,CTR 的预测作为计算广告中的一个关键问题,具有研究的必要性、一定的理论意义和实际价值.

广告的点击率预测就是要通过广告的历史点击记录,预测对于给定的查询用户的点击概率是多少. 这里就需要使用点击日志(Click-Through Log). 分析点击日志是一个预测和优化的双向过程:通过正确的预测来实现广告的正确排序,通过广告的正确排序来提高广告的点击率. 通过对点击日志的分析,不但可以预测和优化广告的点击率,还可以优化搜索引擎的排序结果以及估计用户的满意度. 因此,点击率预测是一个互联网许多领域都需要解决的问题,包括搜索引擎的排序结果以及推荐系统. 广告点击率只是点击率预测的一个应用,但这一应用是全球网络公司的一个重要收入来源,因此有着重要的商业价值和学术研究价,已经成为了近几年学术界和产业界的一个重要研究领域.

为了预测广告的点击率,需要考虑广告的内容与用户查询的相关性,可以表示为用户的观察相关性(User-Perceived Relevance). 这种相关性在信息检索的过程中一般被描述为文本相似性. 在点击率预测的过程中,这种相似性通过点击日志进行计算:用户点击一个广告

说明用户认为广告与自己的查询相关,即可以通过点击记录计算出用户的观察相关性.许多研究领域用点击日志研究用户的偏好,用于提高搜索引擎的排序质量[7-9].

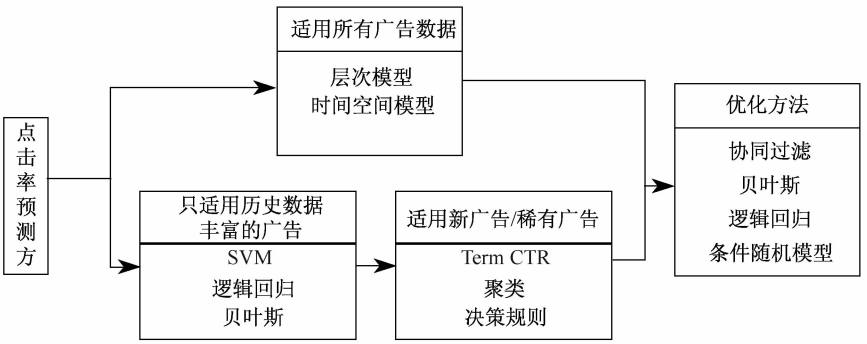

图 1 点击率预测方法框架图

Fig. 1 Framework of click-through rate prediction functions

利用点击日志预测广告的点击率,是一个高效简便的方法,可以用到的方法如图 1 所示.但同时需要解决许多问题.广告的位置对广告的点击率有着重要的影响,显著位置的广告被用户检查到的几率远大于其他位置的广告,这种偏差被定义为位置偏差(Position Bias).文献[10]利用视线跟踪试验证明,根据点击日志计算出的 CTR 不能直接用于衡量广告的内容与用户查询的相关性.文献[11,12,21]提出了基于位置模型的点击率估算方法.文献[13-20,22]提出了基于检验假设(Examination Hypothesis)的方法:假设用户点击一个广告的前提是看到这则广告,从而对用户点击行为进行不同方式的建模,例如用户浏览模型 UBM[14]、动态贝叶斯网络模型 DBN[14]和整页点击模型 WPC[20].这些方法都是基于贝叶斯方法,其共同缺陷是无法处理稀疏广告或新广告.解决这类问题的方法主要分两种:一是通过历史记录丰富的广告对稀疏广告或新广告进行预测[12,23,24],例如关键词聚类[23];另一类方法是采取新的模型对点击数据进行建模,例如层次结构[25-27]以及时间空间模型[28].文献[29-35]提出了一些优化点击率预测准确质量的方法,其中包括增添个性化信息[31]和查询意图信息[33,34].文献[36,37]提出了广告点击率估计的两种应用,分别是点击率欺诈检验和最小曝光数估计.

本文的第 1 节介绍广告点击率估计的应用背景.第 2 节介绍基于位置模型和检验假设的点击率估算方法.第 3 节介绍针对稀疏广告或是新广告的点击率估算方法.第 4 节介绍广告点击率估计的优化方法.第 5 节介绍模型实现和评估方法.第 6 节总结全文并展望点击率预测将来的工作.

1 广告点击率估计的应用背景

典型的互联网广告的投放流程分为三步,如图 2 所示.

请求分析(Request Parsing):广告投放主要分两类,另一类是搜索引擎的赞助商广告,一类是在普通的页面上投放条幅广告.请求分析阶段就是对用户的一次请求进行分析,进而获取相关的信息.对于搜索引擎,即是提取查询 Query 的信息及上下文信息;对于条幅广告,即是提取页面信息、上下文 Cookie 信息、用户的地理信息等.

广告检索(Ad Retrieval):这一阶段的任务是根据请求分析阶段得到的信息,从整个广

告空间中检索出一个与请求相关的子集作为广告投放的待选集合. 这一步骤与搜索引擎的排序过程(Ranking)相似，根据公告的排序得到与请求最相关的一组待选广告. 之后，根据地理信息、上下文信息、黑名单及频率限制对待选广告集合进行筛选. 如果筛选的结果依然过大就要进行第二次筛选，由于二次筛选的代价过高，应尽量避免.

广告选择(Ad Selection)：广告选择的任务是对广告检索所得到的待选集合进行排序，将适当的广告放置于适当的广告位上以实现收益最大化. 这一阶段的第一个任务就是 CTR 预测(Click-Through Rate Prediction)，目的是预测出每个广告放置于每个广告位时，广告的点击率是多少. 第二个阶段，根据 CTR 结果计算出 RPM(广告点击千次的收益)，并根据 RPM 计算结果对待选广告进行排序. 最后一个阶段，通过竞价得到最后需要投放的广告，以及广告的投放位置.

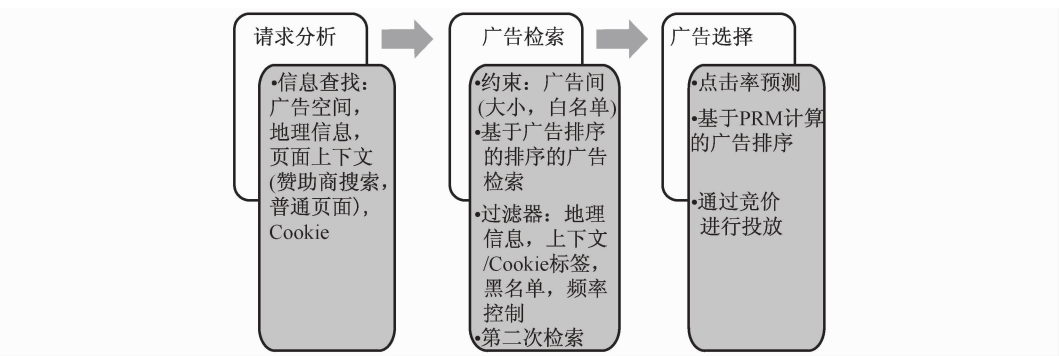

图 2　互联网广告投放流程

Fig. 2　The procedure of advertising on the Internet

广告点击率预测是广告选择过程中的一个重要步骤. 点击率估计是指在给定网页和用户的情况下，估计所投放的广告被点击次数占总展示次数的比例. 互联网广告的点击率从 20 世纪 90 年代起一直呈下降趋势，目前平均点击率在 0.2%～0.3%，0.2%的广告点击率即被视为非常成功的广告投放. 随着广告计费方式的改变，广告点击率估计在广告投放过程中占有越来越重要的地位，估计的结果直接影响到广告检索结果的排序，进而影响到用户、网络媒体和广告主的效用. 据统计，所有广告的展示频率和点击率均呈幂率分布[27]，搜索关键词频率也按幂率分布[24]. 大量的广告和查询的点击日志都是稀疏的，稀疏的数据不利于预测模型的训练，也较难进行广告点击率的准确估计，特别是针对最新投放的广告进行估计.

最原始的点击率预测出现于搜索引擎的排序结果的优化. 因此除了广告排序，点击率预测还有两个重要的应用领域：搜索信息检索和推荐系统. 在计算广告学其他方向的研究中也可以用到点击率预测的结果，如：文献[36]针对计算广告中的按点击率付费(Pay Per Click)模式中的点击欺诈(Click Fraud)问题，提出了一类学习算法，叫做基于点击算法(Click-Based Algorithms)，用来解决点击欺诈中的短期损失(Short Term Loss)和长期受益(Long Term Benefit)的问题；文献[37]针对广告的点击数过于稀少而不能有效地支持广告相关参数的推断问题，阐述了广告的点击行为与搜索结果的点击行为的相关性，进而用搜索结果的点击数据来近似推断每个查询相应的理想的广告数量，以使从最顶部广告获得的收益最大.

2 点击率估算方法

由于位置偏差,广告的点击率无法通过广告的点击日志进行直接计算.处于最显著的位置的广告,用户会最先注意到,因此点击率最高的广告与查询的相关性并不一定最强.针对这一问题,可以建立以下两种模型,如图 3 所示.

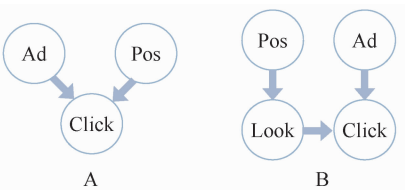

图 3 位置模型和检验假设模型

Fig. 3 Position model and examination model

A 图表示基于位置模型的点击率估算方法:将广告的自身性质与位置信息作为统一级别的两个系数来考虑.假设同一页面上的每条广告的点击事件都可看作一个独立的事件,点击率与广告的相关性和位置有关,不同位置的点击概率依次递减.

B 图表示基于检验假设的点击率估算方法:假定广告的位置不直接影响广告的点击事件,而是决定用户是否能看到这则广告,当且仅当用户看到了一则广告才会检查广告的相关性,这个过程就是对用户行为的一种建模.检验假设模型就是先判断广告是否被用户看到,再计算它的相关性.对整个页面而言这是一个递推的过程,可以通过贝叶斯方法进行建模.

2.1 基于位置模型的点击率估算方法

文献[8]首次提出使用点击日志计算搜索结果的点击率,并结合搜索引擎查询日志和用户点击日志,自动优化搜索引擎的检索质量.通过分析用户在当前返回的排序结果中点击链接的日志,使用支持向量机算法(SVM, Support Vector Machine)最大化 Kendall 相关系数,从而达到排序结果接近最佳排序的目的.同一作者的文章[9]对查询搜索中隐含反馈的可靠性进行了研究,分析了用户选择和谷歌搜索排序的相互影响以及用户的选择行为与排序结果相关性的关系.文献[9]包括两个方面的研究:① 通过眼睛的移动轨迹分析用户在谷歌搜索结果页面的选择行为,如,用户是否从上到下地浏览结果,用户在点击前需要阅读多少网页摘要等.② 通过轨迹和行为分析,提供一些生成反馈信息的策略,并和明确的反馈信息进行对比,评价隐含反馈信息的准确度.进而,综合研究结果,就可以通过隐含日志分析更好的优化搜索引擎排序的排序结果.

文献[11]为页面的每个位置设定了一个固定点击率,利用文档的实际点击率与预先设定的点击率的比值作为文档与查询的相关型,成为 COEC(Clicks Over Expected Clicks)模型.这个模型的缺点是相关性的预测结果可能大于 1.

文献[12,19]使用一种逻辑回归模型(Logistic Regression Model),将广告查询相关性和位置因素作为参数.这种方法的点击率是由逻辑回归方程计算得出,不再具有统计学的物理意义,但这种方法的优点是易于优化,所得结果不会超出[0,1]区间.文献[21]一方面利用基于返回页面单词的逻辑回归模型对"广告与查询返回页面内容的匹配"方法进行改进,另一方面是独立于用户查询而进行点击率的计算和预测,与文献[12]基本一致.

2.2　基于检验假设的点击率估算方法

基于检验假设的点击率估算方法的种类很多,是搜索引擎点击率和推荐系统点击率估计的重点研究方向.这类方法假设用户从上到下依次浏览页面内的文档(广告).如图 4 所示,对于一条广告 $A_i(i\in\mathbf{N})$,E_i 表示用户是否检验广告 A_i,R_i 表示用户是否认为广告 A_i 与查询相关,C_i 表示用户是否点击了广告 A_i,其条件是用户已经检验了广告 A_i[15].由此可以看出,广告是否被点击与广告的相关性和是否被用户检验到有关,广告是否被检验(是否被用户看到)与上一条广告的相关性和点击情况有关.基于检验假设的点击率估算就是对上述过程进行进一步的细化,进行不同的用户行为假设,以实现准确的预测.下面简单介绍几种相关方法.

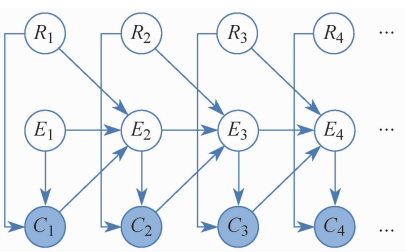

图 4　用户浏览过程,C_i 为唯一可观察到的数据

Fig. 4　In the user browsing process, C_i is observed click data

用户浏览模型(UBM,User Browsing Model)

文献[13]提出了基于位置的两种浏览模型和一种逻辑回归模型,用于根据点击日志预测文档的点击率.假设同一个查询的每一次点击都是一个独立事件,将文档的相关性和位置影响因素作为参数.该文作者利用 EM(Expectation Maximization)算法计算出这两个属性,并利用交叉检验的方法进行性能评估.

动态贝叶斯网络模型(DBN,Dynamic Bayesian Network)

文献[14]提出了一种基于动态贝叶斯网络的点击率预测模型,通过对用户点击过程的建模,分别估计出文档的观察相关性与实际相关性.假设用户点击一个文档当且仅当用户看到了这个文档并且文档的观察相关性(Perceived Relevance)达到了用户的要求;如果用户接下来继续点击了之后的文档,这说明该文档的实际相关性(Actual Relevance)未能达到用户的要求.有时用户在查询的初始阶段并没有明确的查询对象,而是在查询过程中逐步明确查询的目标,这一过程可称作探索性查询.该文的一个强假设是,如果文档的实际相关性达到了用户的需求用户就会终止这次查询,但这在探索性查询中是不现实的.

点击链模型(CCM,Click Chain Model in web search)

文献[15]提出了基于一种贝叶斯框架的点击链模型,提供了一种可扩展的、增量的点击率预测方法.该文假设用户在一个查询会话(Session)中,会依次浏览整个查询结果,并且点击行为仅与文档的位置和文档相关性有关.该文作者将文档的相关性和是否查看下一文档的概率属性设为后验参数,对整个点击过程进行建模.

在线贝叶斯概率回归模型(OBPR,Online Bayesian Probability Regression)

文献[22]提出了一种实现二分预测(Binary Prediction)的算法,即 OBPR 算法,用于在赞助商搜索广告(Sponsored Search Advertising)情景下对 CTR 进行预测.但是 CTR 的预

测仍是以广告特征和网页特征为主,因此很难做到以用户为中心的个性化广告推荐.

3 新广告和稀疏广告的点击率预测方法

广告和用户的信息都有长尾特性,即点击次数高或曝光次数高的广告往往是少数,绝大多数广告的点击和曝光都是稀疏的;发布大量广告的广告主只占整体的极少数,大多数的广告主仅发布少量广告.由此看出,广告数据的特点是非均匀性,存在大量稀疏数据.对于刚刚进入系统的广告而言,更是没有可参考的历史信息.同时,越来越多的网页都采用动态的方式生成,广告也以一定的速度在更新.

本文第 2 节介绍的点击率预测模型,使用的是传统的统计学方法,一般会采用频繁的数据作为样本,对历史数据丰富的头数据(Head)的拟合是有效的,但无法实现对新广告和稀疏广告这些尾数据(Tail)的有限预测,可这些尾数据每年创造着数十亿美元的产值.针对这一问题,可以利用已知广告点击率(头数据),也可以建立适用于新广告和稀疏广告的点击率预测模型来解决.

3.1 利用已知广告点击率

文献[12]对于新广告,利用与其包含相同或相似项(Term)的已知广告来预测其点击率.对于一个 Term,利用保护这个 Term 的已知广告点击率和所有已知广告的平均点击率,估算出这个 Term 的点击率(Term CTR),作为逻辑回归的一个输入变量.同时,计算出相似的一组 Term 的点击率(Related Term CTR),又可以作为逻辑回归的一个输入变量.经实验证明,Term 的点击率和相似 Term 的点击率不但可以预测缺乏历史数据广告的点击率,并且可以极大地优化已知广告点击率的预测结果(相对熵减少约 13%~20%).文献[23]针对用户历史数据不足时的点击率预测问题,给出(基于"关键词—广告主"矩阵)层次聚类(Hierarchical Clustering)方法,其中的聚类通过计算广告的文本相似度来评估.文献[24]利用基于总体决策规则模型(Ensemble of Decision Rules)来预测新的或者很少被展示的广告的 CTR.

3.2 适应用新广告和稀疏广告的点击率预测模型

Deepak Agarwal 等针对设计适应用新广告和稀疏广告的点击率预测模型这一问题,提出了基于层次结构的预测模型[26,27]和基于时间空间模型[28].这是一类很有代表性的方法,利用模型的特点实现对稀疏广告(包括新广告)的直接预测.

3.2.1 基于层次结构的预测模型

文献[25]提出两种不同的平滑计算方法对层次模型进行改进,一种是基于 Empirical Bayes 的自然数据分层(Natural Data Hierarchy)方法,另一种是基于数据一致(Data Continuity)的方法来解决这种稀有事件中的稀疏性问题.(这一段与下面的衔接突兀)

文献[26]中提出基于事先已有的不同粒度的 Web 页面和广告的概念层次模型.针对点击率低和覆盖面稀疏的问题,根据已知数据计算每个层次中不同区域所对应的广告点击率,并以此来对长尾分布中广告的点击率进行估计,可以获得较高的估计准确率.过程分为两个阶段:第一个阶段对样本进行预处理,包括采样和样本补差;第二个阶段采用树状马尔可夫模型(Tree-Structured Markov Model),通过兄弟节点之间的相互关系,在不同抽象层次预测点击率.

文献[27]的研究焦点是预测高纬度、多类别数据中的稀有时间的发生频率.这些数据中

的一些维度是分层的,互联网的广告数据正好有这样的特征,而且在这种层次结构中细粒度层次往往是稀疏的,例如广告的点击率或转化率.该文针对上述数据特性提出了一种多层结构的 Log 线性模型 LMMH(Log-linear Model for Multiple Hierarchies),用于处理 Map-Reduce 框架中大规模的训练数据和预测指标.此模型利用了多层次数据的整体相关性,利用能够提供稳定预测结果的粗粒度层次,来提高由于数据稀疏的细粒度层次的频率预测的准确率.为了提高准确性和扩展性,该文提出了一种基于尖峰和平板回归(Spike and Slab Prior)的筛选步骤(Screening Procedure),用于忽略无用预测指标,简化模型,提高计算效率.此模型不同于文献[26]的地方,在于它建立的是一个有向无环图,提高了通用性.

3.2.2　基于时间空间模型

文献[28]提出一个时空模型(Spatio-Temporal Model),用来计算内容推荐(Content Recommendation)背景下的点击率.通过动态伽马泊松模型模型(Dynamic Gamma-Poisson Model)计算不同位置上文档的点击率;通过动态线性回归模型(Dynamic Linear Regressions),比较不同位置上同一文档的点击率;根据同一文章的重复展示特性,调节用户的疲惫指标.此模型也可以应用于个性化推荐系统:一种方法是建立用户分类,利用 CRT 预测方法计算出每个分类中最热门的文档;另一种方法是建立基于用户特性和 CTR 预测方法的回归模型.

4　广告点击率估计的优化方法

点击率估计的优化一直是研究中的一个重点,其目的是根据用户、广告和页面等的特性,调节广告点击率的预测结果,提高预测效率,从而增加收益和用户满意度.用户的点击行为受很多因素的影响,第 2 节的"点击率估算方法"和第 3 节的"新广告和稀疏广告的点击率预测方法"都只考虑到广告的相关性和位置因素的影响,而除此之外还有很多因素可能影响广告的点击率.例如,广告的定向性或页面内广告之间的差异性.广告的定向性是指广告是否指向一个单一的实体.例如:用户的查询是鞋,广告如果指向一个鞋店,这就为单一定向(Unique Targeted);广告如果指向一个综合网络商城,那就为综合定向(Generated Targeted).文献[12]经统计证实,单一定向广告的点击率高于综合定向广告的点击率;然后将这一特性加入逻辑回归模型中,优化了点击率的预测结果.

除此之外,还有很多优化方法可以提高点击率的预测效果,下面择要介绍.

文献[31]提出了一种基于协同过滤的个性化点击率预测模型.以往的点击率预测模型往往只考虑文档属性和位置信息,而实际用户的个性化信息也对点击率预测有重要影响.该文建立了三种协同过滤模型,分别是:基于查询和文档关系的(MFCM),基于用户、查询和文档三者关系的(PCM)以及综合这两种的(HPCM).这三种协同过滤模型都可以结合基于位置的预测模型使用.通过实验得出第三种模型综合了前两种方法的优点,可以更好地实现点击率预测.

文献[33]提出了一种结合用户搜索意图的点击率预测优化方法.利用点击日志的点击率预测模型,不但受到位置因素的影响,而且还受到用户查询的真正意图和用户的实际查询语句偏差的严重影响.该文作者通过实验得出,96.6%的查询受到意图偏差的影响.针对这一问题,通过贝叶斯方法对用户意图假设进行建模,可以提高以往的基于检验假设的点击率

预测模型的准确性.原有的点击率预测模型,检验(Exam 或 Look)是用户点击事件的一个隐含条件,而这里将用户真实目的也作为一个隐含条件加入贝叶斯模型中,利用 EM 方法进行计算.文献[34]也是研究用户的真实意图,但其着手点为查询:它首先对历史查询进行分类,然后基于已分类的历史查询建立 Intent aware Model,以提高 CTR 预测的精确度.

文献[35]提出了一种利用广告相似性预测广告点击率的方法.广告的点击率不仅取决于广告的自身性质,同时也受到同一页面其他广告的影响.该文认为,同一页上的一组广告的相似性越强,那么每一则广告的辨识度就越低,点击率也就会越低.也就是说,在查询一定时,一条广告的点击率与其同周围广告的相似度成反比.利用广告的自身性质和广告的相似度,采用 CRF(Condition Random Function,条件随机函数)模型,结合最大似然估计方法训练出预测模型.通过实验表明,这种方法的准确性要好于不考虑广告相似性的方法.

文献[30]认为,传统的推荐系统或广告投放系统关注用户的浏览顺序以及广告和页面的特性等方面,而用户所花费的时间和最终的收益数据也十分重要.该文把优化一次点击以及之前相关点击的推荐链的过程定义为点击形成(Click Shaping),实现了一种多目的的优化算法 MOP(Multi-Objective Programming).

还有许多方法可以用于点击率估算的优化,其中包括定向技术(Targeting).定向技术一般用于用户广告投放的检索阶段,但也可以用来提高 CTR 预测的准确率.其中,行为定向(Behavioral Targeting, BT)是一种提高在线广告的效果的技术.文献[29]阐述了行为定向技术对搜索引擎中的在线广告的效果上的影响,并得出了三个结论:① 点击同样广告的用户在网络上具有类似的行为;② 通过赞助商搜索中适当的行为定向广告,可以普遍提高点击率,最高可达 670%;③ 对于行为定向,使用用户的短期行为来代表用户,比用长期用户的行为来代表更有效.除此之外的定向方法还有很多,如地理信息定向(GEO Targeting)、年龄性别定向(Age-Gender Targeting)等.

5 模型实现和评估方法

总结前面所介绍的点击率预测方法,常有的模型共有四类:SVM 模型[8,9],逻辑回归模型[11,12,19,21],贝叶斯模型[13,14-18,20,22,31-33,35]以及其他模型(目前包括层次模型[25-27]和时间空间模型[28].由此也可总结出,逻辑回归模型和贝叶斯模型是研究的热点,近些年的许多优化算法也是基于这两个模型实现的.逻辑回归模型的优点是易于实现和优化,而贝叶斯模型的特点是能够模拟用户的行为特点,有利于个性化推荐.下面结合这两种模型介绍点击率预测的实现和性能评估方法.

5.1 模型实现

评估广告点击率预测的效果,需要使用到真实的点击日志数据.一般的点击日志数据集包括:用户信息(年龄、性别和地理信息等),查询,广告信息(广告主、标题、关键字、描述以及 URL 等)和上下文信息(网页上广告总数和位置).表 1 所示为 kddcup2012[38] 的测试数据集,其中每条记录表示某一用户,一次查询所获得的一个广告的相关信息,包括广告的点击次数(Click)、曝光次数(Impression)、广告链接(DisplayURL)、广告 ID(DisplayID)、广告主 ID(AdvertiserID)、广告所在页面的广告总数(Depth)、广告位置(Position)、查询 ID(QueryID)、查询关键字 ID(KeywordID)、广告标题 ID(TitleID)、广告描述 ID(DescriptionID)和用

户 ID(UserID). 其中，后五项分别对应四个 Token 文件和一个 Age-Gender 文件. 通过这些数据，我们可以利用前面三节提到的方法进行广告点击率预测. 通常情况下，真实数据集中的属性是很多的，例如表 1 的基本训练数据就有 12 个属性，此外还关联 5 个文件. 因此，模型属性(变量)的选择就变得十分重要.

表 1　广告点击日志

Tab. 1　Ad click log

Click	Impression	DisplayURL	AdID	AdvertiserID	Depth	Position	QueryID	KeywordID	TitleID	DescriptionID	UserID
0	1	4298118681424644510	7686695	385	3	3	1601	5521	7709	576	490234
0	1	4860571499428580850	21560664	37484	2	2	2255103	317	48989	44771	490234
0	1	9704320783495875564	21748480	36759	3	3	4532751	60721	685038	29681	490234
0	1	13677630321509009335	3517124	23778	3	1	1601	2155	1207	1422	490234
0	1	3284760244799604489	20758093	34535	1	1	4532751	77819	266618	222223	490234
0	1	10196385171799537224	21375650	36832	2	1	4688625	202465	457316	429545	490234

对于逻辑回归模型可以用公式(1)表示：ad 表示一条广告记录，$f_i(\mathrm{ad})$表示广告的第 i 个预测变量(predictor)($i=1,2,\cdots,n$)，w_i 表示广告第 i 个预测变量的系数(Coefficient)，CTR 表示广告的点击率的预测值. 首要的两个预测变量是广告的点击率(Click/Impression)和广告的位置(Position，Depth)，这里的 Click/Impression 为点击率的观察值，CTR 为点击率的预测值. 通过这两个属性就可以训练出历史几率丰富的广告的预测模型，其他属性(例如广告标题的单词数量或用户的年龄性别分布)可以作为额外参数对模型进行优化. 这些属性直接带入到模型，并通过最大似然估计的方法计算预测变量的系数，通常情况下预测变量系数和预测值 CTR 都是未知的，这就要使用最大后验的估计方法(Maximum a Posteriori Probability, MAP)计算模型的系数，例如 L-BFGS 方法.

$$\mathrm{CTR}=\frac{1}{1+\mathrm{e}^{-Z}},\quad Z=\sum_{iz1}^{n}w_i f_i(\mathrm{ad}) \tag{1}$$

对于贝叶斯模型，以上的数据是不够的，因为该模型模拟的是用户的一次查询行为，或者说是用户提交一次查询的会话(Session)，这就还需要知道查询结果页面上的所有广告以及广告的点击情况，以此来对用户的行为进行建模. 以动态贝叶斯网络模型[14]为例，对于一个 Session 中的一个广告(如图 4)，与图 5 所描述的通用模型相比较，增加了一个预测变量 S_i，S_i 表示用户是否对文档(广告)的 Landing Page 满意. 模型假设，如果用户对文档的 Landing Page 满意，用户就会终止本次 Session，不会再浏览其他文档；如果用户对 Landing Page 不满意，用户会以一定的几率继续浏览后续文档. 通过这一假设就可以计算出两个变量 a_u 和 s_u. a_u 表示文档的观察相似性，当且仅当 a_u 大于一定值时，用户才会点击广告(Click-Through)；s_u 表示广告的实际相关性，当且仅当 s_u 小于一定值时用户才会检验后续广告. 通过以上过程，建立贝叶斯模型(a_u 和 s_u 为先验变量)，通过 EM 算法进行计算.

5.2　评估方法

点击率预测有许多评测方法. 文献[19]利用逻辑似然得分(Log Likelihood Score)，都是通过与标准估计式的比较评价估计结果. 大多数文献都是用相对熵(KL-divergence)和均方差(MSE)的方法计算误差值，评估估计的准确性. 在用于排序结果(Ranking)优化时，也

可以使用 DNCG[14](Compute the Normalized Discounted Cumulated Gain)方法,对优化效果进行评价.使用 ROC 图(ROC curves)对比各个模型的性能也是一个常用的方法,能够体现各模型在不同能性能指标上的一个 Tradeoff 关系,例如均方差和对错误率.

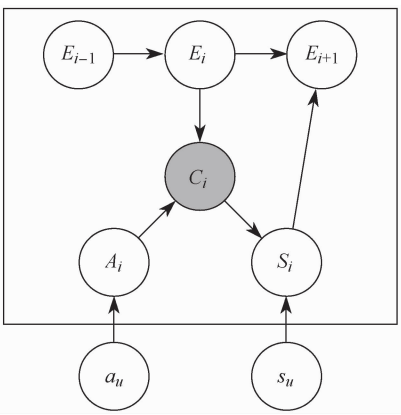

图 5 动态贝叶斯网络,C_i 是唯一的观察值

Fig. 5 Dynamic Bayesian Network, C_i is the unique obvious variable

6 总 结

本文介绍了计算广告学的一个重要研究领域:广告点击率预测.广告点击率预测是广告排序的第一个步骤,其过程是利用广告点击日志,预测广告在不同用户、不同上下文环境下的点击率.根据广告历史记录的丰富程度,广告可以分为是历史数据丰富的广告和(曝光次数和点击次数)稀有广告(包括新广告).针对历史数据丰富的广告,可以使用 SVM 模型、逻辑回归模型和贝叶斯模型来预测广告点击率;针对稀有广告或新广告,可以结合历史记录充分广告的 CTR,使用 Term CTR、层次聚类和决策规则的方法进行估算.此外,还有一些方法,例如基于层次模型和事件空间模型的点击率预测方法,适用于所有广告.目前,逻辑回归模型和贝叶斯模型普遍应用于各商业互联网广告投放领域,也是学术的研究重点.逻辑回归模型的特点是易于实现和优化,但不能模拟用户行为,扩展性差.贝叶斯模型对用户的浏览行为进行建模,可以通过多种角度进行拓展(例如用户查询的目的性、用户的查询习惯等).点击率预测不但可以用于与互联网广告投放,还被广泛应用于信息检索排序优化和推荐系统,对提高收益和用户满意度都有着重要作用.因此,广告点击率预测是一个非常值得研究的领域.

接下的研究工作有两个重点:个性化推荐和在线更新算法.个性化推荐是广告投放领域一个研究重点,个性化推荐可以针对不同的用户进行广告投放,由此提高广告的点击率,此类问题可以结合定向技术实现.同时广告点击数据的海量和实时更新的特点给预测模型的性能和更新带来了巨大挑战,如何利用分布式的计算方法实现模型的在线更新算法也是一个待解决的问题,同时也是工业界的一个迫切需求.

[参 考 文 献]

[1] 周傲英,周敏奇,宫学庆.计算广告:以数据为核心的 Web 综合应用[J].计算机学报,2011,34(10):1805-1819.

[2] BRODER A, JOSIFOVSKI V. Introduction to Computational Advertising [M/OL]//The Course of Computation Advertising, Stanford University. 2011[2013-03-21]. http://www.stanford.edu/class/msande239.

[3] GABRILOVICH E. An Overview of Computational Advertising [R/OL]. Yahoo Research. 2008[2013-03-21]. http://research.yahoo.com/pub/2915.

[4] AGARWAL D, CHAKRABARTI D. Statistical Challenges in Online Advertising [R/OL]. Yahoo Research. 2008[2013-03-21]. http://research.yahoo.com/ pub/2430.

[5] GRAEPEL T, BORCHERT T, HERBRICH R, et al. Probabilistic Machine Learning in Computational Advertising [R/OL]. Microsoft Research, 2009[2013-03-21]. http://research. microsoft.com/en-us/um/beijing/events/mload-2010.

[6] DID-IT, ENQUIRO, EYETOOLS. Eye Tracking Study [R/OL]. 2007[2013-03-21]. http://www.enquiro.com/ eye-tracking-pr.asp.

[7] AGICHTEIN E, BRILL E, DUMAIS S. Improving web search ranking by incorporating user behavior information [C]//Proc 29th SIGIR, 2006:19-26.

[8] JOACHIMS T. Optimizing search engines using clickthrough data[C]//Proc KDD, 2002.

[9] JOACHIMS T, GRANKA L, PAN B, et al. Accurately interpreting clickthrough data as implicit feedback[C]//Proc SIGIR, 2005.

[10] GRANKA L A, JOACHIMS T, GAY G. Eye-tracking analysis of user behavior in www search[C]//Proc SIGIR'04, 2004.

[11] ZHANG V, JONES R. Comparing click logs and editorial labels for training query rewriting[C]//Query Log Analysis: Social And Technological Challenges. Banff: WWW'07, 2007.

[12] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting Clicks: Estimating the Click-Through Rate for New Ads[C]//Banff: WWW'07, 2007.

[13] DUPRET G E, PIWOWARSKI B. A user browsing model to predict search engine click data from past observations.[C]//SIGIR'08, 2008.

[14] CHAPELLE O, ZHANG Y. A dynamic Bayesian network click model for web search ranking[C]//WWW'09, 2009.

[15] GUO F, LIU C, KANNAN A, et al. Click chain model in web search[C]//WWW'09, 2009.

[16] LIU C, GUO F, FALOUTSOS C. Bbm: Bayesian browsing model from petabyte-scale data[C]//KDD'09, 2009.

[17] ZHU Z A, CHEN W, MINKA T, et al. A novel click model and its applications to online advertising[C]//WSDM'10, 2010.

[18] DUPRET G, LIAO C. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine[C]//WSDM'10, 2010.

[19] BECKER H, MEEK C, CHICKERING D M. Modeling contextual factors of click rates[C]//AAAI'07, 2007.

[20] CHEN W Z, JI Z L, SHEN S, et al. A Whole Page Click Model to Better Interpret Search Engine Click Data [C]//AAAI'11, 2011.

[21] CHAKRABARTI D, AGARWAL D, JOSIFOVSKI V. Contextual Advertising by Combining Relevance with Click Feedback[C]//WWW'08, 2008.

[22] GRAEPEL T, CANDELA J Q, BORCHERT T, et al. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine[C]//ICML'10, 2010.

[23] M Regelson, D C. Fain. Predicting Click Through Rate Using Keyword Clusters[C]//EC'06, 2006.

[24] DEMBCZYNSKI K, KOTŁOWSKI W, WEISS D. Predicting Ads' ClickThrough Rate with Decision Rules[C]//WWW'08, 2008.

[25] WANG X, LI W, CUI Y, et al. Click Through Rate Estimation for Rare Events in Online Advertising[C/OL]//Online Multimedia Advertising: Techniques and Technologies, 2011, Chapter 1[2013-03-21]. http://labs.yahoo.com/node/434.

[26] AGARWAL D, BRODER A Z, CHAKRABARTI D, et al. Estimating rates of rare events at multiple resolutions [C]//KDD'07, 2007.

[27] AGARWAL D, AGRAWAL R, KHANNA R, et al. Estimating rates of rare events with multiple hierarchies through scalable log-linear models[C]//KDD'10, 2010.

[28] AGARWAL D, CHEN B C, ELANGO P. Spatio-Temporal Models for Estimating Click through Rate[C]//WWW'09, 2009.

[29] YAN J, LIU N, WANG G, et al. How much can Behavioral Targeting Help Online Advertising? [C]//WWW'09, 2009.

[30] AGARWAL D, CHEN B-C, ELANGO P, et al. Click shaping to optimize multiple objectives[C]//KDD'11, 2011.

[31] SHEN S, HU B, CHEN W Z, et al. Personalized click model through collaborative filtering[C]//WSDM'12, 2012.

[32] CHEN W Z, WANG D, ZHANG Y C, et al. A noise-aware click model for web search[C]//WSDM'12, 2012.

[33] HU B, ZHANG Y C, CHEN W Z, et al. Characterizing search intent diversity into click models[C]//WWW'11, 2011.

[34] ASHKAN A, CLARKE C L A, AGICHTEIN E, et al. Estimating Ad Clickthrough Rate through Query Intent Analysis[C]//WI-IAT'09, 2009.

[35] XIONG C, WANG T, DING W, et al. Relation Click prediction for sponsored search[C]//WSDM'12, 2012.

[36] IMMORLICA N, JAIN K, MAHDIAN M, et al. Click Fraud Resistant Methods for Learning[C]//WINE'05, 2005.

[37] GOLLAPUDI S, PANIGRAHY R, GOLDSZMIDT M. Inferring Clickthrough Rates on Ads from Click Behavior on Search Results[C]//WSDM'11, 2011.

[38] TENCEN T. KDD CUP[EB/OL]. 2012[2013-03-21]. http://www.kddcup2012.org/c/kddcup2012-track2.