浙江理工大学学报(自然科学版),第 33 卷,第 6 期,2015 年 11 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 33, No. 6, Nov. 2015

文章编号: 1673-3851 (2015) 06-0851-07

广告点击率预估技术综述

陈巧红, 余仕敏, 贾宇波

(浙江理工大学信息学院,杭州 310018)

摘　要: 广告点击率的预估是计算广告学领域的重要研究内容,准确的广告点击率预估可以提高真实的广告点击率,增加收益。逻辑回归模型、支持向量机模型、贝叶斯模型、神经网络模型等模型适用于历史广告点击数据丰富的情况,适用无历史广告点击数据和广告点击数据稀疏的模型包括层次聚类模型、相似项预估模型、因子分解机等模型,而时间空间模型、层次模型则适用上述所有广告点击数据的情况。根据不同的广告数据特征,采用不同的模型,可以获得很好的预估效果。

关键词: 广告点击率;预估模型;神经网络;因子分解机

中图分类号: TP181　　**文献标志码**: A

0　引　言

广告点击率(click-through rate, CTR)是指在广告显示中广告被用户点击的概率,广告点击率的预估就是根据广告数据和用户数据来预估广告点击率。现在很多搜索公司例如百度联盟和 Google AdSense 都是采用点击付费(cost per click, CPC)[1],点击付费是现在主流的付费方式,这类付费机制最适合交易型广告,此类广告的收益就是点击次数和每次点击的付费金额的乘积。研究显示,用户点击广告的概率性与广告的投放位置有很大的相关性[2],要获得最大的收益就是要把点击率大的广告投放在靠前的位置。根据精确的 CTR 预测来确定投放的顺序,在线地在返回页面中投放广告[3]。

为了预测广告的点击率,要充分考虑影响广告点击率的因素,例如广告自身的影响和广告浏览者的影响相关性,上下文内容的相关性等因素,从而进一步提高广告的点击率,使点击次数和每次点击的付费金额的乘积变大,以此扩大搜索引擎的收益。

1　在线广告点击率预估流程

图 1 所示广告点击率预估流程,数据包括广告日志数据以及用户数据,根据不同模型的要求提取相应的特征数据,这些特征数据通过归一化或是规范化后,输入到点击率预估模型训练,通过预估出的点击率再进行排序,确定广告的投放位置,提高真实的点击率,从而扩大收益。

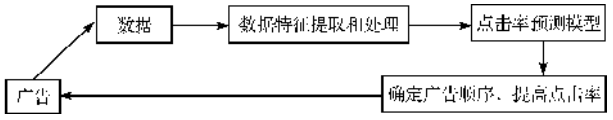

图 1　点击率预估系统流程

广告点击率预估模型就是利用机器学习算法模型以及概率统计模型去预估广告的点击率,如图 2 所示。其中基于历史广告数据丰富的预估模型,本文分别介绍逻辑回归模型[4]、贝叶斯模型[5-8]、基于决策树模型[9-10]、递归神经网络模型[11]、支持向量机模型[12-13]、混合模型[14]和 COEC 模型[15]。基于稀疏广告数据和新广告数据的预估模型,本文介绍基于层次的预估模型[16-18]、相似项点击率预估模

收稿日期: 2014－11－13

基金项目: 浙江省自然科学基金项目(LQ13F020015)

作者简介: 陈巧红(1978－),女,浙江临海人,副教授,主要从事计算机辅助设计及机器学习技术方面的研究。

通信作者: 余仕敏, E-mail: ywy2130635@163.com

型[19]、基于先验概率的实时点击预估模型[20]、时间空间模型[21]]和因子分解机模型[22]。

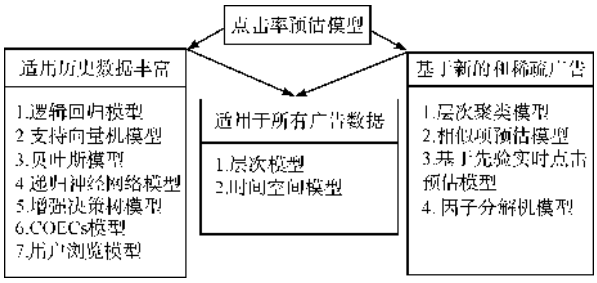

图 2 广告点击率的预估模型

图 3 所示的是广告点击率预估常用的评估方法，常用的有 KL 距离(KL-Divergence)[20]和 ROC 曲线下面积(area under curve，AUC)方法[21]。

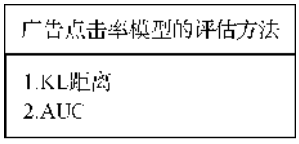

图 3 广告点击率模型的评估方法

2 广告点击率预测的模型

2.1 基于历史数据丰富的预估模型

在广告本身的历史数据点击数据丰富的情况下，预测该类型的广告要充分利用广告的历史点击数据，基于逻辑回归的预估模型、基于贝叶斯网络的预估模型、基于支持向量机的预估模型、基于神经网络的预估模型和基于决策树的预估模型等模型在历史数据丰富的情况下可以得到很好的训练，最终获得很好的预估效果。

2.1.1 基于逻辑回归预估模型

Richardson 等[4]采用逻辑回归模型(logistic regression model)去预估广告点击率，目前很多公司的广告点击率预估都是基于逻辑回归的预估模型。该模型的特点就是简单且非常适合解决概率预估问题，其核心计算公式为：

$$CTR = \frac{1}{1+e^{-Z}} Z = \sum_i w_i f_i(ad) \qquad (1)$$

式(1)中的 i 是广告的第个特征的数值，是广告的第 i 个特征的学习权值。

该文使用 L-BFGS(limited-memory broyden-fletcher-goldfarb-shanno)方法(该方法是拟牛顿方法的一个优化算法)来训练逻辑回归模型，损失函数使用零均值和标准差的正态分布的交叉熵函数，每一个广告特征都归一化为期望值为 0、单元标准差的数，该归一化也应用于之后的训练和测试广告特征数据集。模型效果评估采用的是 KL 距离，文中的 KL 距离是模型预测的 CTR 和真实的 CTR 的距离。KL 距离简化了 log 似然模型，忽略测试数集的熵。作者还增加一个均方差(mean squared error，MSE)作为一个评估指标。由于逻辑回归模型采用最大似然估计，需要大量数据以保证性能，所以不适合对稀疏广告数据的预估。

2.1.2 基于支持向量机算法的预估模型

Joachims[5]提出了从 web 搜索引擎日志中挖掘点击数据，利用支持向量机实现对广告点击率的预估。支持向量机(support vector machine，SVM)利用核函数将一个向量映射到比其自身更高维的空间，在高维空间建立一个最大间隔的超平面。在分隔超平面两边各有一个与之平行的超平面，最大化平行超平面之间的间隔，平行超平面距离越大，分类效果越好。基于核函数，支持向量机可以处理多维非线性数据。该文利用点击数据通过支持向量机来预估点击率，从而提高搜索引擎的检索能力，在没有明确的反馈信息和没有人工参数优化的情况下，该模型可以自动适应一些特殊的参数选择。

2.1.3 基于贝叶斯网络的预估模型

Chapelle 等[6]提出动态贝叶斯网络模型，作者介绍了满意度的概念，利用这个概念去分别模拟登陆页面的相关性和搜索结果页面可感知的相关性。动态贝叶斯网络模型是用来模拟用户浏览行为，并且认为只有用户看到链接并且认为该链接与用户所要获得的信息有关的情况下才去点击这个链接，用户基于文档观察相关性决定是否要点击和通过结果做出一个线性横向选择。如果用户不满意点击的链接他们会选择点击下一个链接(基于真实的相关性)。

Guo 等[7]提出基于贝叶斯结构的点击链模型(click chain model in web search)，类似链表结构，所以该模型具有很好的扩展性。将文档内容的相关性和用户点击下一个链接的概率作为相关性后验参数来建立模型的。

Graepel 等[8]提出在线贝叶斯概率回归模型(online bayesian probability regression)，该模型基于特定广告特征，所以很难准确做到个性化推荐。

Dupret 等[9]提出了一种用户浏览模型的点击率估算方法，利用点击日志预测文档的点击率，假设每次用户点击行为都是互相独立，将日志内容的相关性和位置距离作为参数，利用 EM(expectation maximization)算法迭代计算出所有参数的最大似然估计，再利用交叉检验的方法进行性能评估。

基于贝叶斯网络的模型,当数据发生变化时模型必须重新训练,耗时过长,且对新广告数据无法预估。

2.1.4　基于神经网络的预估模型

Zhang 等[10] 提出了利用递归神经网络(recurrent neural networks,RNN)来预测搜索广告的点击率问题。递归神经网络是在多层神经网络的中间层建立一种组合的递归网络,用于取代一般的多层网络,并依次对被控对象的动态特性进行直接的学习,通过调整其中有关参数,以获得所需的最优控制输入,过去的一些工作只是将单独的广告曝光作为输入去预测点击概率,并没有考虑到不同广告曝光的依赖性,而且过去在时间序列的分析常常也只是关注于构建数据序列趋势或是周期性模式。近些年来的一些研究利用 RNN 来解决数据的时间依赖的问题。例如 RNN 语言模型[11]成功的利用大量语言库中的大跨度连续的信息,获得了比传统神经网络语言模型更好的效果。RNN 由于它的特殊的递归神经网络的结构使其有很大的能力去利用数据间连续的依赖关系。Michael Auli 等认为每个用户的广告浏览历史可以作为一个序列,从而产生了固有的内部依赖关系。神经网络的输入量必须是二元值,基于 BPTT(back propagation through time)算法的 RNN 框架如图 4 所示,实验结果表明,比起神经网络和逻辑回归模型该模型的预测广告点击率更加准确。模型的评测标准采用的是 AUC(area under roc curve)和 RIG(relative information gain)。

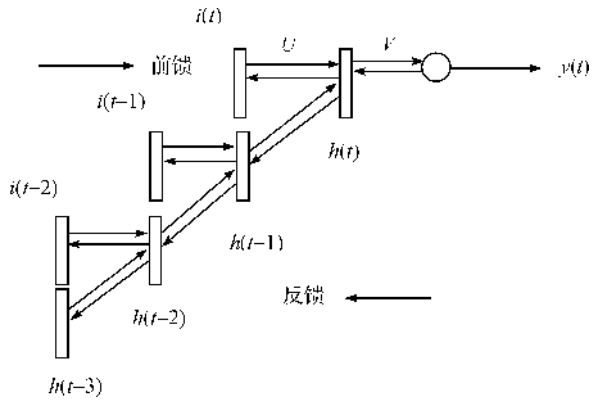

图 4　RNN 结构框架

2.1.5　基于决策树的点击率预估模型

Dave 等[12]在文中从广告数据中提取相似性特征,再利用梯度增强决策树(gradient boosting decision tree, GBDT)作为一个回归模型训练相似特征来预估广告点击率。GBDT 不像传统的决策树模型只有一棵决策树,而是由多棵决策树组成的。Boosting 的基本思想是将一系列弱分类器组合起来,构成一个强分类器,也就是让每棵树不需要学太多的东西而是学一点点,再将每棵树学到的知识累加起来组合成一个强大的模型。它的思想起源于 Valiant 提出的 PAC(probably approximately correct)学习模型。

Rofimov 等[13]提出了基于强分类决策树(boosting tree)的一种机器学习的算法 MatrixNet,它是梯度提升机器模型(gradient boosting machine, GBM)算法中采用随机上升(stochastic boosting)的修改版算法。MatrixNet 算法继承了 GBM 的优点,而且在 GBM 中采用随机上升可以进一步提高准确性和功能性。MatrixNet 从 GBM 算法继承以下 3 个主要的超参数,分别是上升步长 M,正则化率 v 和最大树高度 H。作者采用均方误差(MSE)作为效果评测标准。

增强分类决策树的优点:a)防止过度拟合;b)高阶交互处理;c)接近不连续函数;d)在大多数情况下不需要功能转换。

2.1.6　COEC 预估模型

Zhang 等[14]提出了 COEC(clicks over expected clicks)模型,COEC 定义为预先设置一个期望的点击率值,再利用实际点击率与之前设置的期望点击率的比值作为目标函数,它具有排序标准化的好处。

2.1.7　混合模型

Wang 等[15]提出了将 4 种模型(在线贝叶斯概率回归(online bayesian probit regression)、支持向量机(support vector machine)和因子模型(latent factor model)和基于最大似然估计的模型(maximum likelihood estimation)组合起来使用的一种混合模型,该模型可以更好地模拟用户点击行为,进而预估广告的点击率。首先使用基于一种综合特征的方法去抽取和产生描述特征数,然后使用上述的描述特征数,将四种模型应用于训练数集中,使用基于 MLE 的一些模型方法去模拟经常在训练集中出现的实例,是为了充分利用训练数集,最后提出了一种基于排序的集成学习方法,该方法可以规范化 4 种模型方法的结果并产生最后的结果。图 5 就是作者所使用的多种模型结构图。特别需要指出的是针对多种模型使用了两组特征数,一组是原始特征数(包括离散特征和连续特征),一组是合成特征数(将任意两组原始离线特征数结合起来作为一个合成特征数)。最后的评估标准采用的是 ROC 曲线下面积。

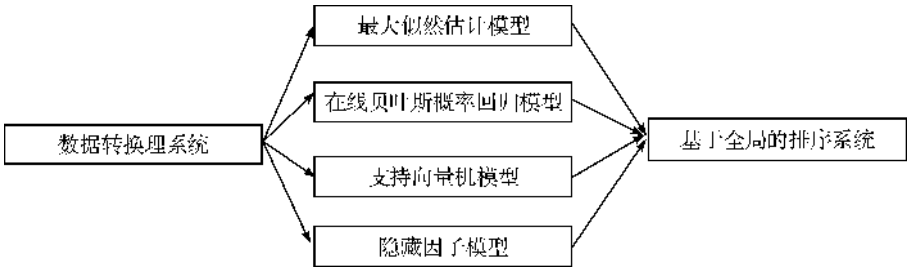

图 5 多模型结构

2.2 基于稀疏数据广告和新广告数据的预估模型

具有丰富历史点击数据的广告毕竟是少数，大多数的点击数和曝光数都是很稀疏的，特别是新投放到平台的广告更是没有历史点击数据的参考，所有需要在线地评估。2.1 节所介绍的预估模型对历史数据丰富的情况下，能获得很好的预估效果，但是对上述问题不奏效，针对上述问题，下面介绍了基于层次的预估模型、基于相似项的预估模型、基于时间空间的预估模型和基于先验概率的实时点击的预估模型来解决广告数据稀疏和新广告的预估问题。

2.2.1 基于层次结构的模型

Regelson 等[16] 提出了一种层次聚类(hierarchical clustering)方法，在历史数据不足缺少或者没有历史数据情况下，用广告的文档相似度来预估点击率，这种使用历史数据分层聚合的方法可以获得更准确的估计。

Agarwal 等[17]提出使用稀疏数据预先存在的层次结构，解决稀疏事件及其稀疏数据的出现率估算问题，主要解决针对 web 网页、广告的点击率的预估，这些网页和广告都可以在不同粒度中获取广泛的上下文信息来按层次分类。典型的情况是点击率非常低的和层次覆盖面比较稀疏的问题，为了解决这些问题，该文作者采用的抽样方法是分析那些从训练集中选取的特别样本。该模型的预估点击率模型可以分为两个阶段，第一个阶段就是调整样本偏差，第二个阶段就是采用树形结构的马尔可夫模型(tree-structured markov model)，通过同一级节点的相关性来达到对该层次点击率的预估。

Agarwal 等[18]提出了一种针对稀疏事件广告数据具有高维多元可分层特征的预估方法模型，该模型叫做多层次 Log 线性模型(log-linear model for multiple hierarchies)，这种模型可以处理在 Map-Reduce 框架的大规模数据(十亿级别的训练集合，数百万潜在的预测因子)。考虑到准确性和扩展性，采用了一个基于尖峰和平板回归(spike and slab prior)的内置筛选过程，删除那些影响预测准确性的因子，保证准确性。

2.2.2 基于相似项的预估模型

Richardson 等[19]提出了一种方法利用新广告和已知点击率广告的相同或者相似项(Term)去预测新广告的点击率。根据新广告与旧广告的相似项在线地根据新广告数据评估新广告的点击率，采用聚类的方法，通过广告内容的相似度来预估点击率。

2.2.3 基于时间空间的预估模型

Agrawal 等[20] 在 2009 年提出了时空模型(spatio-temporal predicting models)预估点击率，通过动态伽马泊松模型(dynamic gamma-poisson model)计算一段时间内固定位置的文档点击率；通过动态线性回归模型(dynamic linear regressions)结合相关位置的文档信息，有效地提高每一位置的点击数，文中的各个模型通过基于特殊用户和重复曝光性特征的首次点击概率(probability of click on first article exposure)的指数级数来调整用户的疲劳度，并且该模型支持个性化的推荐。

2.2.4 基于先验的实时点击预估模型

Fang 等[21]提出了一种针对具有极其稀疏和瞬时性特征的广告数据实时点击的预估模型。鉴于好的 ID 特征数据具有极其稀疏和瞬时性特征，这使传统的机器学习处理起来很困难。提出了基于先验的实时点击预估模型(prior-based real-time estimator model，PRE)，该模型可以直接使用上述的特征数据，首先从之前学习的先验模型计算不同维的经验点击率数据，然后构造最小方差无偏估计量(minimum variance unbiased estimator)来作为点击率数据的加权和，最后使用权值参数的另一个数集来放宽独立性假设这个条件，独立性假设这个条件影响每一维的数据。PRE 模型最大的好处就是它自身具有实时性，只需要一些参数进行离线学习，PRE 模型经过一段时间训练就可以得出相对稳定的结果，并且简单。与此同时，所有的在线计算都在封闭中进行的，而且证明很有效果。为了进一步提

高估计效果,还使用了若干模型的融合技术去更好的结合 LR 模型和 PRE 模型。最后通过实验得出,PRE 模型可以提高点击率预估模型的准确性和排名能力,特别是结合最新的数据,该模型的时效性超过一般的机器学习模型。

2.2.5　因子分解机模型

Rendle 等[22]提出因子分解机模型(factorization machine models),过去因子分解模型虽然是预测效果很好的模型,但是只针对特定的数据集,并且需要用不同的方法去处理不同的数据集,例如有平行因子分析法(parallel factor analysis)、因子分解个性化马尔可夫链(factorizing personalized markov chains)等因子分解的方法。因子分解机模型是结合支持向量机和因子分解模型的优点,支持向量机无法对稀疏数据进行预估,因子分解机不断的事实化参数对参数变量进行建模,所以因子分解机仍然适用于稀疏数据的预估,这也是与支持向量机相比最大的优点。

2.3　各种模型的对比和总结

前面介绍了各种广告点击率预估模型,针对不同广告数据来源采用不同的预估模型,不同的预估模型有它的优缺点,适用的场合也不尽相同,各模型具体的比较如表 1 所示。

由于每个模型都有优缺点,为了克服一些缺点,新的算法不断地被提出,例如平衡采样逻辑回归算法[23]采用平衡采样,由于删除了大量的负样本集,能缩短了训练时间,能在不牺牲点击率预估效果提升系统的性能,解决了训练时间的问题;基于联合概率矩阵分解的上下文广告推荐算法[24],该算法适用于广告数据稀疏和大规模数据的情况,解决了过去了一些模型无法预估稀疏广告数据和大规模广告数据的缺点。

表 1　广告点击率预估模型的优点和缺点

模型类别	优　点	缺　点	适用场合
逻辑回归模型	采用最大似然估计,在极端情况下结果可靠性较高。	不适用于稀疏的广告数据。	
支持向量机模型	可以处理非线性问题、高维的数据问题,避免神经网络结构选择和局部极小点问题。	对缺失数据比较敏感,影响模型的准确性。无法对稀疏广告数据进行预估。	
基于贝叶斯网络模型	所需估计的参数很小,对缺失数据不太敏感。数据属性相关性较小时,性能最为良好。	必须知道先验概率且属性之间是相互独立的,对新广告点击率无法预估。	历史广告数据丰富的场合。
基于神经网络模型	分类的准确性高,并行分布处理能力强,能充分逼近复杂的非线性关系,充分利用。	需要大量的参数,无法了解之间的学习过程,学习时间长。	
决策树模型	在相对短的时间内能够对大型数据源做出可行且效果良好的结果,一般无需参数准备过程。	存在过度拟合的问题忽略数据集中属性之间的相关性。	
层次聚类模型	针对稀疏广告数据进行预估,可以得到很好的效果。	广告数据要具有层次性,聚类时的时空复杂性高;聚类的簇效率低、误差大。	
时间空间模型	该模型支持个性化推荐。适用于各类广告数据。		稀疏和新广告数据的场合。
基于先验的实时点击预估模型	自身具有实时性,训练得出的结果相对稳定简单。	该模型只是处于初步研究,还有很多深入的研究。	
因子分解模型	预测质量很高。	只适用于某一类或者特殊的指定数据类型,对参数的调节非常敏感。	

3 广告点击率模型的评估方法

3.1 KL 距离

KL 距离[25](KL-divergence)又叫相对熵,它是两个概率分布的距离,这里的距离不是真实的距离,相对熵衡量的是相同事件空间里的两个概率分布的差异情况,其意义就是概率分布 $P(x)$事件空间,如果使用 $Q(x)$概率(也可以叫做真实的概率情况)去编码,其基本事件的平均编码长度增加了多少比特。其计算公式如下:

$$D(P||Q)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)} \qquad (2)$$

式(2)中 $D(P||Q)$ 就是 KL 距离。$P(x)$ 信息熵的含义是平均每个基本事件至少需要多少比特编码。根据信息熵的知识可知,不存在其他比按照自身概率分布更好的编码方式。所以相对熵是大于等于 0 的。预估的点击率概率分布是 $Q(x)$,真实的点击率概率分布是 $P(x)$,由此可以得出 KL 距离越小,越接近真实的概率分布,所以模型预估的点击率越准确,效果越好。

3.2 ROC 曲线下面积法[26]

ROC(receiver operating characteristic)曲线分析,它是医疗分析领域的一种新的分类模型性能的评估方法,其中 ROC 的混淆矩阵主要用于比较分类结果和实例的真实信息,矩阵的每一行代表实例的预测类型,每一列代表实例的真实类别,在 ROC 坐标中,横坐标表示假正率,纵坐标表示真正率,真正率表示正例分到正的概率,假正率表示负例错误的分到正的概率。

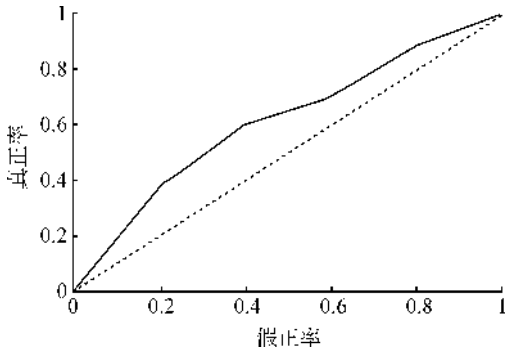

图 6 ROC 曲线

图 6 所示,曲线下面积 AUC 就是处于 ROC 曲线下方的那部分面积的大小。AUC 的值在[0.5,1]区间内,值较大表示性能较优。首先根据模型预测的每个广告的点击率的不同,按高低依次确定投放广告的位置,预测值大的放在前面,然后根据真实的点击所反馈的信息,假正率等于 1 减去真正率,预设一个阈值,根据阈值将实例分成正类和负类,根据分类结果来绘制 ROC 曲线,其中 Y 轴方向代表被点击率,X 轴代表未被点击率,由此可以得知 ROC 曲线下面积就越大,AUC 就越大,预估的广告点击率就越准确。

4 结 语

广告点击率是计算广告的重要内容,也是提高广告收益的主要手段之一,本文首先介绍了一些相关知识,然后重点介绍了多种广告点击率的预估模型,基于历史广告数据的预估模型,例如逻辑回归模型、贝叶斯模型、决策树模型等,还有针对历史广告数据不足的预估模型,例如分层聚类模型,对未投放的新广告的预估模型,例如 Term CTR 模型,还有适用于所有广告数据的模型,例如时间空间模型和层次模型,最后介绍了点击率预估模型的常用的评估方法。

互联网的快速发展,广告点击率的预估模型也在不断的改变,传统上 Google、百度等搜索公司是以逻辑回归模型作为预估模型,百度意识到 LR 严重限制了模型学习与抽象特征的能力[27],百度尝试将 DNN(deep neural network)深度学习应用到搜索广告,并于 2013 年服务于百度搜索广告系统。但 DNN 在搜索广告的应用远远不够,结合海量的广告点击数据,如何充分发挥分布式分析计算的最大能力去实现广告点击率预估,如何提高广告点击率预估的准确性以及更好地实现个性化广告精准推荐,是未来的发展方向。

参考文献:

[1] 李 敏. 计算广告学将成为数字商业的奠基学科[J]. 程序员, 2014 (5): 109-109.

[2] 周傲英, 周敏奇, 宫学庆. 计算广告:以数据为核心的 Web 综合利用[J]. 计算机学报, 2011, 34(10): 1805-1891.

[3] 纪文迪, 王晓玲, 周傲英. 广告点击率估算技术综述[J]. 华东师范大学学报: 自然科学版, 2013 (3): 2-14.

[4] Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads [C]//Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 521-530.

[5] Joachims T. Optimizing search engines using clickthrough data[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining. ACM, 2002: 133-142.
[6] Chapelle O, Zhang Y. A dynamic bayesian network click model for web search ranking[C]//Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 1-10.
[7] Guo F, Liu C, Kannan A, et al. Click chain model in web search[C]//Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 11-20.
[8] 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]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 13-20.
[9] Dupret G E, Piwowarski B. A user browsing model to predict search engine click data from past observations [C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 331-338.
[10] Zhang Y, Dai H, Xu C, et al. Sequential click prediction for sponsored search with recurrent neural networks[J]. AAAI, 2014:1369-1375.
[11] Auli M, Galley M, Quirk C, et al. Joint language and translation modeling with recurrent neural networks [C]//EMNLP. 2013: 1044-1054.
[12] Dave K, Varma V. Predicting the click-through rate for rare/new ads [R]. Centre for Search and Information Extraction Lab International Institute of Information Technology Hyderabad - 500 032, India, 2010.
[13] Rofimov I, Kornetova A, Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search [C]//Proceedings of the 6th International Workshop on Data Mining for Online Advertising and Internet Economy. ACM, 2012: 2.
[14] Zhang W V, Jones R. Comparing click logs and editorial labels for training query rewriting[C]//WWW 2007 Workshop on Query Log Analysis: Social and Technological Challenges. 2007.
[15] Wang X, Lin S, Kong D, et al. Click-through prediction for sponsored search advertising with hybrid models[C]//KDD Workshop. 2012.
[16] Regelson M, Fain D. Predicting click-through rate using keyword clusters[C]//Proceedings of the Second Workshop on Sponsored Search Auctions. 2006, 9623.
[17] Agarwal D, Broder A Z, Chakrabarti D, et al. Estimating rates of rare events at multiple resolutions [C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 16-25.
[18] Agarwal D, Agrawal R, Khanna R, et al. Estimating rates of rare events with multiple hierarchies through scalable log-linear models[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 213-222.
[19] Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads [C]//Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 521-530.
[20] Agarwal D, Chen B C, Elango P. Spatio-temporal models for estimating click-through rate [C]// Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 21-30.
[21] Fang Y, Liu J. A novel prior-based real-time click through rate prediction model[J]. International Journal of Machine Learning and Cybernetics, 2014, 5(6): 887-895.
[22] Rendle S. Factorization machines[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 995-1000.
[23] 施梦圜, 顾津吉. 基于平衡采样的轻量级广告点击率预估方法[J]. 计算机应用研究, 2014, 31(1): 33-36.
[24] 涂丹丹, 舒承椿, 余海燕. 基于联合概率矩阵分解的上下文广告推荐算法[J]. 软件学报, 2013, 24(3): 454-464.
[25] Kullback S, Leibler R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951,22(1): 79-86.
[26] 刘　唐. 基于多类别特征的在线广告点击率预测研究: 以腾讯搜搜为例[D]. 北京:北京邮电大学,2012.
[27] 余　凯, 贾　磊, 陈雨强. 深度学习: 推进人工智能的梦想[J]. 程序员, 2013 (6): 22-27.

(下转第 871 页)

Study on Acquisition of Printing and Dyeing MES Production Data with RFID-Based Electronic Material Vehicle

RUAN Deng-feng[1], *ZHOU Yan-jiang*[1], *XU Guang-ming*[2]

(1. School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2. Hangzhou Kaiyuan Computer Thechnology Co., Ltd., Hangzhou 310013, China)

Abstract: Most of current Manufacturing Execution Systems (MES) adopted by printing and dyeing enterprises apply barcode technology as the identificaiotn of production task to indirectly acquire process data. It is hard to ensure instantaneity and accuracy of production information. aiming at this problem, a method of acquiring real-time process date in printing and dyeing MES by using RFID (Radio Frequency Identification) technology through material vehicle in shop floor was proposed. The design of the electronic material vehicle based on RFID technology, the technologies of anti-collision recognition, location and radio frequency card read-write were studied additionally. The above method can achieve synchronization of material processing information and production management information in informatization architecture of two databases and three layers of two data base and three layers through MES-based real-time information acquisition in production process.

Key words: printing and dyeing; RFID; MES; production data acquisition

(责任编辑: 陈和榜)

(上接第 857 页)

Overview of Advertisement Click-through Rate Estimating Techniques

CHEN Qiao-hong, *YU Shi-min*, *JIA Yu-bo*

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The prediction of advertisement click-through rate is an important research content in the field of computational advertising. Accurate prediction of advertisement click-through rate can improve real advertisement click-through rate and increase income. Logistic regression model, support vector machine (SVM) model, the Bayesian model and neural network model are applicable to enriching historical advertisement click-through data, the models without historical advertisement click-through data and sparse advertisement click-through data, similar term prediction model and factorization machine etc. Time-space model and hierarchical model apply to all the above situations. According to the characteristics of different advertising data, different models can get good prediction effect.

Key words: cadvertisement click-through rate; prediction model; neural network; factorization machine

(责任编辑: 陈和榜)