基于卷积神经网络的搜索广告点击率预测

李思琴，林　磊，孙承杰

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘　要：广告点击率的预测是搜索广告进行投放的基础。目前已有的工作大多数使用线性模型或基于推荐方法的模型解决点击率预测问题，但这些方法没有对特征之间的关系进行深入的探索，无法完全体现广告点击预测中各个特征之间的关系。本文提出了基于卷积神经网络的搜索广告点击率预测的方法，阐述了卷积神经网络在特征的学习上模拟人的思维过程，并进一步分析了不同特征在广告点击率预测中的作用，在 KDD Cup 2012 中 Track 2 数据集上的实验结果验证了本文提出的方法能够提高搜索广告点击率的预测效果，其 AUC 值达到 0.792 5。

关键词：卷积神经网络；点击率预测；搜索广告

中图分类号： TP391.41　　**文献标识码：** A　　**文章编号：**2095－2163(2015)05－0022－05

Click－Through Rate Prediction for Search Advertising based on Convolution Neural Network

LI Siqin, LIN Lei, SUN Chengjie

(**School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China**)

Abstract: Click－Through Rate (CTR) prediction is the foundation of search advertising. Nowadays, lots of researches have been explored to predict CTR, and most of those researches either rely on liner model or employ method of recommendation system. However, the relations between different features in CTR predication have not been fully explored in previous works, and the relations between different features also cannot be fully embodied. In this paper, CTR prediction for search advertising based on convolution neural network is proposed, and process of convolution neural network simulating the process of human thought on feature learning is explained. Furthermore, the performance of different features have been analyzed in the task of predicting CTR. Experiments are conducted on the dataset of KDD Cup 2012 Track2 and the proposed method achieves 0.792 5 in AUC, demonstrating the effectiveness of the proposed approach.

Key words: Convolution Neural Network; Click－Through Rate Prediction; Search Advertising

0　引　言

随着 Web 搜索技术的成熟，搜索广告已经成为互联网行业的主要收入来源之一，其根据用户输入的查询词，在搜索的结果页面呈现出相应的广告信息。广告媒介的收益通过每次点击费用(CostPerClick，*CPC*)与广告点击率(Click－Through Rate，*CTR*)预测共同影响而得到，即 $CPC * CTR$。由于用户点击广告的概率随着广告位的排放顺序呈递减趋势，因此对 *CTR* 进行准确高效的预测，并将 *CTR* 高的广告投放在搜索结果页面靠前的位置，不仅能增加广告媒介的收益，还能提高用户对搜索结果的满意程度。

广告点击率预测是广告算法中最核心的技术，近年来被学术界广泛关注。部分学者使用基于推荐方法的模型来解决 *CTR* 预测问题。霍晓骏等人[1]采用协同过滤算法，为页面找到与其相似的其他邻居页面，实现 *CTR* 的预测，以此作为基础进行广告推荐，但当相似页面的数量增加时，该方法的结果质量会严重下滑。Kanagal 等人[2]提出了一种聚焦矩阵分解模型，针对用户对具体的产品的喜好以及相关产品的信息进行学习，解决因用户－产品交互活动少而造成的数据稀疏问题。在文献[2]的基础上，Shan 等人[3]提出了一种立方矩阵分解模型，通过对用户、广告和网页三者之间关系的立方矩阵进行分解，利用拟合矩阵的值来预测 *CTR*，虽然立方矩阵分解模型增加了一维交互关系，但所刻画的交互关系仍然十分局限，不能在 *CTR* 预测中充分挖掘广告所有特征之间的联系。

作为典型的预测问题，很多研究中通过将 *CTR* 预测问题看作分类或者回归问题来解决，其中最常见的是应用线性模型来预测 *CTR*。Chapelle 等人[4]使用动态贝叶斯网络，通过对用户产生的点击过程建立模型，考虑级联位置的信息模拟出特定位置与相近位置的相关性，以判断该位置上的广告是否满足用户搜索要求。Chakrabarti 等人[5]利用点击反馈的相关性，通过在网页和广告词等特征上使用逻辑回归模型提高广告检索和预测的效果。Wu 等人[6]基于融合的思想，将不同线性模型的实验效果相结合，来提高搜索广告 *CTR* 预测的结果。真实的场景中 *CTR* 的预测并非简单的线性问题，因

收稿日期：2015－05－26
基金项目：国家自然科学基金(61300114，61272383)。
作者简介：李思琴(1990－)，女，广西桂林人，硕士研究生，主要研究方向：自然语言处理、计算广告学；
林　磊(1970－)，男，黑龙江哈尔滨人，博士，副教授，主要研究方向：自然语言处理、计算广告学；
孙承杰(1980－)，男，黑龙江哈尔滨人，博士，副教授，主要研究方向：计算广告学。

此,一些学者开始使用非线性模型来解决 *CTR* 的预测。Dave 等人[7]在搜索广告点击信息以及广告商账户信息上提取语义特征,使用基于投票思想的梯度提升决策树模型,提高了 *CTR* 预测的效果。Zhang 等人[8]利用神经网络模型对影响搜索广告点击率的因素进行的探索,从特征因素方面提高 *CTR* 预测的结果,但是资源单一,数据交互的关系没有获得良好的利用。

本文对基于卷积神经网络(Convolution Neural Network, CNN)的 *CTR* 预测进行研究,通过卷积与亚采样操作的结合,能更好地学习出数据特征之间的关系,不仅解决了线性模型无法模拟真实广告数据场景的问题,也解决了浅层学习模型无法深入挖掘特征间相互关系的问题,并且较之于传统的神经网络,CNN 能更好地理解特征之间的关系。在真实的数据集上的实验验证了本文的方法能够提高搜索广告中 *CTR* 预测的 *AUC* 值。

1　卷积神经网络模型

1.1　卷积神经网络基本模型

卷积神经网络作为人工神经网络之一,目前已成为深度学习领域中研究的热点,权值共享以及局部窗口滑动的特点使之能更好地模拟出生物神经网络。卷积神经网络在结构上有两个重要的组成部分:卷积层和亚采样层,如图 1 所示。

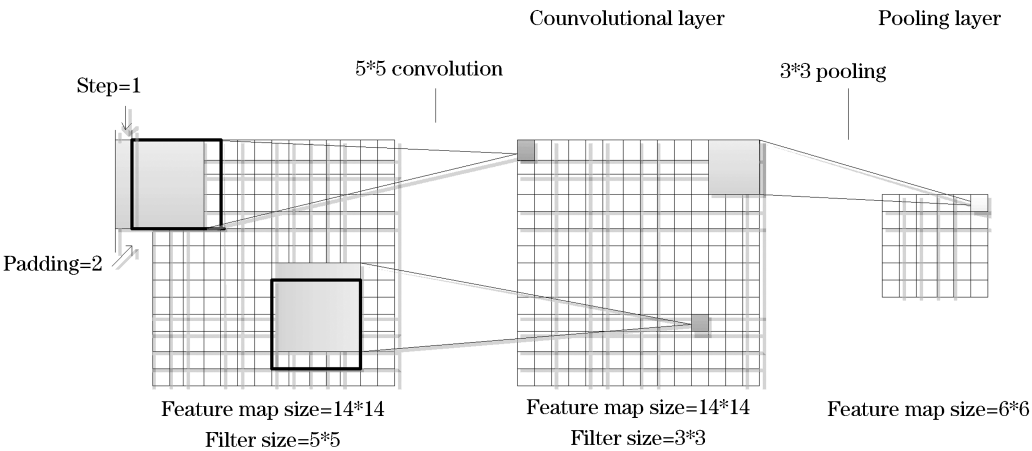

图 1 卷积层与亚采样层结构

Fig. 1 Convolution layer and sub – sampling layer structure

在卷积层中,原始特征通过卷积核进行卷积得到输出的特征,使用不同的卷积核就可以得到一系列不同的输出特征。对卷积层的计算,有如下公式:

$$x_j^l = f\left(\sum_{i \in p_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

这里,$f(\sim)$是 *sigmoid* 函数,$f(u) = \dfrac{1}{1+e^{-u}}$,$u = \sum_{i \in p_j} x_i^{l-1} * k_{ij}^l + b_j^l$;$P_j$代表输入特征上选定的窗口,即在卷积过程中当前卷积核在计算时所对应在输入特征上的位置;x_i^{l-1} 和 x_j^l 分别是第 $l-1$ 层输入特征和第 l 层输出特征上相应的值;k_{ij}^l是卷积核的权重值;b_j^l 是特征的偏置,每一层对应一个。

卷积过程,一个卷积核通过滑动会重复作用在整个输入特征上,构建出新的特征。同一个卷积核进行卷积时,共享相同的参数,包括同样的权重和偏置,这也使要学习的卷积神经网络参数数量大大降低了。而当使用不同的卷积核进行卷积时,可以得到相应的不同的输出特征,这些输出特征组合到一起,构成卷积层的输出。

在亚采样层,前一个卷积层的输出将作为该层的输入特征,首先设定大小的窗口,然后通过滑动,用窗口区域中最大(或平均)的特征值来表示该窗口中的特征值,最后组合这些特征值得到降维后的特征。亚采样过程可表示如下:

$$x_j^l = f(pool(x_j^{l-1}) + b_j^l) \tag{2}$$

这里,类似于卷积层,$x_i^{(l-1)}$ 和 x_j^l 分别是第 $l-1$ 层输入特征和第 l 层输出特征上相应的值,b_j^l 是特征的偏置;*pool*$(\sim)$表示取最大值 Max(x)或者平均值 Avg(x)的函数。

典型的卷积神经网络通常由 $n(n>=1)$个卷积层和亚采样层以及最末尾的 $m(m>=1)$全连接层组合而成。一个亚采样层跟随在一个卷积层后出现,通过这若干卷积层和亚采样层后得到的特征,将经过全连接层与输出层相连。全连接层公式如下:

$$x^l = f(u^l), u^l = K^l x^{l-1} + b^l \tag{3}$$

这里,$f(\sim)$是 *sigmoid* 函数,K^l 是计算第 $l-1$ 层到第 l 层时的权重值。

1.2　基于卷积神经网络的 CTR 预测模型

研究中使用卷积神经网络对搜索广告的 *CTR* 进行预测,网络结构如图 2 所示。

实验中一共设置了两个卷积层、两个亚采样层以及一个全连接层。首先从历史日志中提取相应的特征构建出输入(Feature_Input),设置好卷积的窗口大小后根据公式(1)对输入特征进行卷积,每一次卷积是对窗口内所有值的组合,因此卷积过程相当于特征融合过程。对卷积后得到的特征,设置亚采样的窗口并根据公式(2)进行最大值 – 采样,选取出窗口中的最有表达能力的特征值(最大特征值)表示整个窗口的特征,因此亚采样过程相当于特征的萃取过程。整个卷积和亚采样过程的结合,模拟出了人对事物的理解和总结的过程。最后将特征经过一层全连接后连接到输出,得到最终的预测结果。

在一次特定的卷积(或亚采样)全过程中即训练的一次迭代过程中,权值并不会随着窗口的滑动而改变,即在计算中,所有窗口滑过的特征享受同样的权值。这也是 CNN 区别于其他神经网络的特点——权值共享。如此即使得 CNN 更方便训练,更能多角度地对特征进行学习。

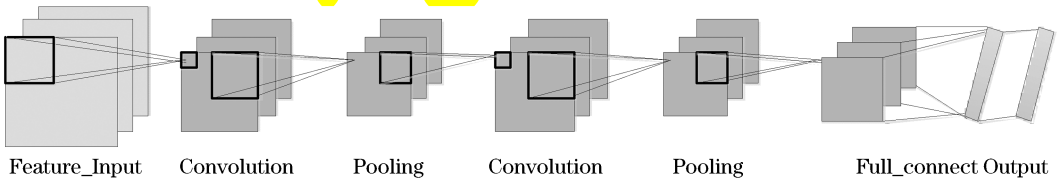

图 2 卷积神经网络在搜索广告点击率预估中的应用

Fig. 2 Convolution neural network in search ad click rate through prediction

2 特征构建

本文所采用的实验数据集为 KDD Cup 2012 中 Track 2 提供的数据集。该数据由腾讯公司下的搜索品牌搜搜(SOSO)搜索引擎提供,因为涉及公司商业信息,数据经过哈希处理。实验数据集中,每条记录包含 12 个属性,各属性详解如表 1 所示。

表 1 实验数据介绍

Tab. 1 The description of dataset

属性名	属性描述	属性类别
Click	点击次数	广告 – 网页属性
Impression	展示次数	广告 – 网页属性
DisplayURL	网址	广告属性
AdID	广告	广告属性
AdvertiserID	广告商	广告属性
Depth	广告展示个数	网页属性
Position	广告展示位置	广告 – 网页属性
QueryID	用户查询	用户属性
KeywordID	广告关键字	广告属性
TitleID	广告标题	广告属性
DescriptionID	广告描述	广告属性
UserID	用户	用户属性

研究按照实际含义将这 12 个属性构造了四大类特征:历史点击率特征、相似度特征、位置特征和高影响力特征。

2.1 基于卷积神经网络的 *CTR* 预测模型

历史点击率特征是不同类别 ID 在历史数据中的点击率,虽然比较简单但十分有效,因为历史点击率在一定程度上代表了类别 ID 对某个广告感兴趣程度的高低,当一个 ID 对某个广告的历史点击率高时,意味着其对这个广告更感兴趣,后续点击的概率也更大。

历史点击率(*pseudo – CTR*)是点击数(*#click*)与展示数(*#impression*)之比,在统计计算过程发现在很多情况下有些类别信息没有点击实例,因此研究采用了平滑方法解决零值问题,根据公式(4)来计算平均点击率。计算公式如下:

$$pseudo - CTR = \frac{\#click + \alpha \times \beta}{\#impression + \beta} \tag{4}$$

公式中的 α 和 β 是调节参数,根据公式(4)计算出 AdID、AdvertiserID、QueryID、KeywordID、TitleID、DescriptiomID、UserID 的历史点击率。

2.2 相似度特征

相似度特征用来刻画属性两两之间的相似程度,用户搜索的内容与被投放的广告属性相似度高时,广告被点击的概率更大。例如当搜索内容 Query 与广告关键字属性 Keyword 相似度高时,意味着网页投放的广告与用户期望搜索的广告结果相似度高,更符合用户点击广告的动作。

通过对 Query、Keyword、Title、Description 的属性描述文件构造出相关的 *TF – IDF* 向量,Query 为用户搜索内容,Keyword、Title、Description 是广告的相关属性,数据集提供的属性信息都是经过哈希后的数字形式,但是属性之间的相对含义不变,然后计算相互之间的余弦相似度作为特征。

2.3 位置特征

该特征描述的是指定广告在搜索结果页面中的位置信息。用户搜索时需求的多样化要求在对广告进行排序和投放时,在结果页面靠前的位置中尽可能地投放满足用户需求的广告,从而最大化用户的满意度、提高用户点击的兴趣[9]。因此,研究即用当前预测广告的相对位置 *Pos* 来刻画该广告在结果页面中排序靠前的程度,其定义如下:

$$Pos = \frac{total\_ads - ad\_position}{total\_ads} \tag{5}$$

这里,*total_ads* 指页面投放的广告总数,*ad_position* 指当前所预测广告的位置。

2.4 位置特征

在预测模型中,ID 属性信息通常采用 one – hot 形式的特征编码方式,在将不同的属性经过 one – hot 编码后的特征向量组合在一起,这样方式简单直观,却使得特征的维度巨大并且非常稀疏。然而在这庞大且稀疏的特征中,绝大部分维度上的特征值对整个模型的预测结果贡献非常小甚至为零,只有少数维度上的特征值对预测结果有较高的影响力。因此研究采用 L1 范数正则化的方式,在逻辑回归模型的代价函数中加入 L1 范数[10],使得模型学习得到的结果满足稀疏化,在学习参数中按大小顺序取出前 N 维权重较大的,将这 N 维权重对应位置上的特征值构建新的特征,称为高影响力特征,考虑到实验硬件,取 $N = 180$。

3 实验结果与结论分析

3.1 数据准备

实验目标是通过给定的信息预测搜索网页的广告点击率,由于数据量过大并且正负样本不平衡,实验中从训练集随机采样 10% 作为本文实验中模型训练的训练集,既缩小了样本空间,同时随机采样也保持了原始数据的分布信息。实验中随机抽取部分样本作为验证集用于参数的调节。本文所用测试集为 KDD Cup 2012 中 track 2 的全部测试数据,因

此本文的结果与 KDD Cup 2012 中 track 2 比赛的结果具有可比性。数据的统计信息如表 2 所示。

表 2　实验数据统计信息

Tab. 2 Statistics of dataset

数据集	样本数	点击数	展示数
KDD Cup 2012 track2 训练集	149 639 105	8 217 633	235 582 879
测试集	20 297 594	418 403	13 303 612
本文所用训练集	15 000 000	876 389	23 652 694
验证集	500 000	25 649	609 694

使用 *AUC*(Area Under Curve)[11]作为点击率预测的评价标准,*AUC* 值等于以 *TPR*(True Positive Rate)为纵坐标、以 *FPR*(False Positive Rate)为横坐标所画曲线下的面积值,其中 *TPR* 与 *FPR* 的计算定义如下:

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

这里,*TP*、*TN* 分别表示结果中预测对的正样本数和负样本数,*FP*、*FN* 分别表示结果中预测错的正样本数和负样本数。对于广告点击率预测问题,较大的 *AUC* 值代表了较好的性能。

3.2　实验设置和结果分析

实验的操作系统为 Ubuntu 12.04 LTS OS,卷积神经网络在 4G RAM 的 NVIDIA GeForce GT 610 GPU 条件下运行。过程中选用了 Dense Gaussian 对卷积层、亚采样层的边和节点进行初始化,用常数初始化输出层,学习卷积神经网络各边权值时的优化函数使用梯度下降算法,其中学习率为 0.01、动量项为 0.9,训练步数为 100,设置公式(4)中参数 $\alpha = 0.05, \beta = 75$。实验时使用逻辑回归模型(LR)、支持向量回归模型(SVR)和深度神经网络(DNN)作为对比方法,所有方法都使用相同的特征,其中 DNN 的层数以及每层的节点数与卷积神经网络相同。

具体地,首先探究了卷积神经网络中节点的设置,因为在 CNN 中后续层的节点数根据第一个卷积层和每层卷积(或亚采样)滑动窗口的大小计算得到,并以第一个卷积层节点的设置为实验变量,同时控制 DNN 中每层的节点数均与 CNN 相同,在验证集上的实验结果如图 3 所示。

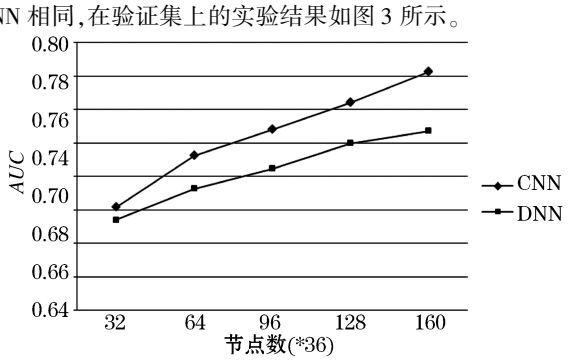

图 3 *AUC* 随第一个卷积层中节点变化的趋势图

Fig. 3 Convolution neural network in search ad click rate through prediction

从图 3 可以看出,随着节点的增加,*AUC* 的值也在不断增长,在一定范围内,节点数越多,实验的结果越好。但随着节点数的增大,模型的训练时间也在延长,对设备的开销需求也在升高,综合上述因素,最终将第一层的节点数设为 9 216。

CNN 与各对比实验的实验结果如表 3 所示,可以看出 CNN 的效果最佳,此外在表中还列出了 KDD Cup 2012 track 2 比赛中第一名的结果。DNN 的 *AUC* 值优于 LR 和 SVR,验证了深度学习模型比浅层学习模型更适合解决 *CTR* 预估问题,同时 CNN 的结果高于 DNN,说明 CNN 中卷积层的特征融合和亚采样层的特征萃取过程是有效的。本文中 CNN 目前的实验结果略低于 KDD Cup 2012 track 2 中第一名的结果,原因是比赛队伍使用了多模型融合并提取了庞大的(千万维)输入特征。

表 3　*CTR* 预测的结果

Tab. 3 The result of *CTR* prediction

方法	*LR*	*SVR*	*DNN*	*CNN*	*KDD*
AUC	0.741 2	0.735 0	0.754 7	0.792 5	0.806 9

进一步地,实验探索了每一类特征对搜索广告点击率预测的贡献。在所有特征的情况下,去掉某一类特征来进行预测,实验结果如表 4 所示。实验结果表明,去掉任意一类特征都将使得实验效果有所下降。其中去掉历史点击率特征效果下降得最明显,说明用户是否点击广告,与其之前的点击行为非常相关。而去掉位置特征时,效果下降得最为不明显,因为在实验使用的数据集中,每个页面最多仅呈现三个广告,页面中的广告数少时,位置对用户点击的影响小。

表 4　各类特征的贡献

Tab. 4 Contribution of each feature class

特征	去掉历史点击率特征	去掉相似度特征	去掉位置特征	去掉高影响力特征
AUC	0.642 8	0.772 0	0.771 4	0.788 3

4　结束语

对搜索广告点击率的有效预测不但能够更好地提高在线广告投放的性能,增加广告商的收益,还能增强用户的体验。研究使用卷积神经网络 CNN 对搜索广告点击率进行预测,对特征因素的分析之后,在真实数据的环境下对搜索广告点击率进行预测的实验本文的方法的效果相对于其他方法有明显的提高。本文的主要贡献有:(1)本文提出了基于卷积神经网络的搜索广告点击率预测的方法。(2)针对高维特征,提出了一种特征选择策略,可以在计算能力受限的情况下使用 CNN 模型来解决广告点击预测问题,并取得较好效果。在未来的工作中,一方面要继续研究更有效的特征来提高对点击率的预测效果,另一方面也将尝试对 CNN 模型的内部细节进行改进,使之更适合人们的预测场景。

参考文献:

[1]霍晓骏,贺樑,杨燕. 一种无位置偏见的广告协同推荐算法[J]. 计算机工程, 2014, 40(12): 39 - 44.

(下转第 28 页)

adodb 包的路径

```
$conn = ADONewConnection('mssqlnative');
$conn->Connect($dbhost,$dbuser,$dbpass,$dbname);
? >
```

其中,*ADONewConnection*( ) 函数功能是连接数据库,'mssqlnative'是连接 SQL Server 的数据库驱动程序名称。*Connect*( ) 函数功能是实现与数据库的连接,$dbhost, $dbuser, $dbpass 和 $dbname 分别为 SQL Server 数据库的服务器 IP 地址,用户名,密码和数据库名。

在以上两种 PHP 访问 SQL Server 方法中,支持 UTF－8 字符集的关键是在连接 SQL Server 时,设置 SQL Server 的字符集 CharacterSet 为 UTF－8。

5　PHP 与 SQL Server 之间编码的转换

在 PHP 访问 SQL Server 过程中,当连接 SQL Server 的字符集设为 UTF－8 编码时,SQL 语句需要使用 UTF－8 编码,返回的数据集的字段名需要使用 GB－2312 编码。因此,在 PHP 网页设计过程中,需要应用 *iconv* ( ) 函数进行编码转换[4]。

在网页设计过程中,要注意 PHP 文件的存储格式。当转换文件格式后,要调整文件中 SQL 语句和数据集字段名的编码。

例如,在维吾尔语考试阅卷系统中,应用 adodb 方式访问 SQL Server,查询"试题 1"表,返回"题号"和"答案"字段值。

(1)当网页文件格式为 UTF－8 时,查询语句为:

```
$Strsql = "SELECT * FROM 试题 1";
$rs = $conn->Execute($Strsql);
```

"题号"和"答案"字段表达式为:

```
$rs->fields[iconv("utf－8","gb2312",'题号')];$rs->fields[iconv("utf－8","gb2312",'答案')];
```

(2)当网页文件格式为 ANSI 时,查询语句为:

```
$Strsql = "SELECT * FROM 试题 1";$rs = $conn->Execute(iconv("gb2312","utf－8",$Strsql));
```

"题号"和"答案"字段表达式为:

```
$rs->fields['题号'];
$rs->fields['答案'];
```

6　结束语

本文是对[5]的补充和完善,在[5]中,由于尚未搞清 mssql、adodb 访问 SQL Server 方式支持 UTF－8 字符集的问题,在维吾尔语口试评卷系统的网页设计中,采用图像方式显示维吾尔文试题和答案。

在 PHP 访问 SQL Server 的 mssql、adodb 方式支持 UTF－8 字符集这一问题解决后,应用 PHP 与 SQL Server 处理维吾尔文数据的问题也相应地得到了解决。在维吾尔语口试评卷系统中,应用 mssql、adodb 方式访问 SQL Server 数据库中维吾尔文数据,取得了很好的结果。

参考文献:

[1]维尼拉· 木沙江,艾尔肯· 伊米尔. 维吾尔文 Unicode 在线处理技术与实现[J]. 新疆大学学报(自然科学版), 2004,21(3):332－334.

[2]Tali Smith. Install the SQL Server Driver for PHP [EB/OL]. [2009－11－15]. http://www.iis.net/learn/application－frameworks/install－and－configure－php－on－iis/install－the－sql－server－driver－for－php.

[3]潘凯华,刘中华,等编著. PHP 从入门到精通[M](第二版). 北京:清华大学出版社,2010.

[4]陈军红,王瑞敬. PHP 编程从基础到应用[M]. 北京:清华大学出版社,2014.

[5]贾志先. 维吾尔语口试阅卷系统开发中若干问题的研究[J]. 智能计算机与应用,2015,5(4):30－32.

(上接第 25 页)

[2]BHARGAV K, AHMED A, PANDEY S, et al. Focused matrix factorization for audience selection in display advertising[C]// Data Engineering (ICDE), 2013 IEEE 29th International Conference on, Brisbane , Australia: IEEE, 2013:386－397.

[3]SHAN Lili, LEI Lin, DI Shao, et al. CTR Prediction for DSP with Improved Cube Factorization Model from Historical Bidding Log [M]// C K Loo, et al(Eds.): Neural Information Processing, Switzerland:Springer,2014,8836:17－24.

[4]OLIVIER C, ZHANG Ya. A dynamic bayesian network click model for web search ranking[C]//Proceedings of the 18th international conference on World wide web, Madrid: ACM,2009:1－10.

[5]DEEPAYAN C, AGARWAL D, JOSIFOVSKI V. Contextual advertising by combining relevance with click feedback[C]//Proceedings of the 17th international conference on World Wide Web, Beijing: ACM,2008:417－426.

[6]WU Kuanwei, FERNG C S, HO C H, et al., A two－stage ensemble of diverse models for advertisement ranking in KDD Cup 2012 [J]. KDDCup, 2012.

[7]DAVE K S, VARMA V. Learning the click－through rate for rare/new ads from similar ads[C]// Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland:ACM,2010.

[8]ZHANG Ying, JANSEN B J, SPINK A. Identification of factors predicting clickthrough in Web searching using neural network analysis [J]. Journal of the American Society for Information Science and Technology, 2009, 60(3): 557－570.

[9]林古立. 互联网信息检索中的多样化排序研究及应用[D]. 广州:华南理工大学,2011.

[10]YUAN Guoxun, HO C H, LIN C J. An improved glmnet for l1－regularized logistic regression[J]. The Journal of Machine Learning Research, 2012, 13(1): 1999－2030.

[11]FAWCETT T. ROC graphs: Notes and practical considerations for researchers[J]. Machine learning, 2004, 31: 1－38.