基于DNN+GBDT的Query类目预测融合模型

吴嘉伟怀风 2018-02-01 9559浏览量

简介： 用户搜索意图的理解在搜索排序体系下有着重要的作用。在搜索引擎中，分析用户的搜索Query和哪些文档类目的意图更相关，被称为Query的类目预测。本文通过集合Query的语义和行为等特征，计算得到与Query最相关的类目，并在线上得到了相关性□的体验的提升。

1. Query类目预测与分析

用户搜索意图的理解在搜索排序体系下有着重要的作用。要理解用户的意图，一部分可以通过用户的关键词文本来理解语义上的意图；另一方面可以通过用户的行为积累来获得用户的潜在需求意图，即基于用户行为的意图预测。

文本意图中的类目意图预测是淘宝搜索相关性中的重要组成部分。商品在关键词索引召回之后，在第一轮海选粗排阶段通过类目相关性，可以优先选择更相关类目的商品进入第二轮精排中。一方面保证排序的效率，使得排序在类目相关的商品集合上进行；另一方面从最上层保证类目的相关性，保证用户的体验效果。

Query类目预测主要目标是，分析用户的搜索Query和哪些类目的意图更相关。Query类目预测不同于商品类目预测和一般的文本分类问题，主要是因为Query带有的描述信息比较少，而且往往意图比较分散，也就是对于一个用户搜索的Query会有多个可能相关的类目，对预测意图的难度会比较大。例如上图用户搜索“电脑显示器”，其中转接线就属于与用户意图不相关的类目商品。

1.1 线上模型实现

目前线上的版本模型主要包括点击模型和先验模型两部分。

1.1.1 点击模型

用户对于搜索的query的点击商品行为，很大程度上反应出类目的相关，即点击越多的类目，越有可能和query的意图在类目维度更相关。

所以点击模型主要依赖于搜索词的历史行为，也就是Query在各个召回商品的类目下的历史一段时间的点击行为做了统计，并且根据各个类目的点击分配比例，通过分数阈值的规则划分类目相关与否的档位。

1.1.2 先验模型（新）

先验模型主要为了解决点击模型带来的马太效应问题，相似类目的点击行为差异和新发类目的商品冷启动问题。

目前依赖的方法主要是是通过Query在类目下的召回结果数以及在类目下商品的占比等因素，相当于商品体量的一个估计。两个模型之后进行融合，通过规则的方式确定相关和不相关的档位阈值。

1.1.3 在线实现

每天通过搜索日志中选择一段时间内的中高频Query，通过上述统计规则方法得到Query的类目预测结果，并存储为离线词典。（在线访问时，主要依赖于这份基础数据词典。对于不在词典中的偏长尾的Query，则通过线上丢词的长尾类目预测逻辑进行识别，将在另一篇文章中介绍~）

1.2 存在的问题与分析

基于这种融合模型的方法，点击模型带来更准确的类目相关度量，而先验模型对于行为稀疏的类目可以得到召回上的提升，使得线上的相关性效果得到一定的保障。但是在应用中还是存在着以下几个问题：

首先是点击模型，统计历史的类目点击数据对于曝光较多的类目会占优势，对于类目商品体量不均的情况，会使得点击分数更偏向体量大的类目；而且如果存在类目新发商品或新增类目的情况，即使点击数据在近日内有增量，但是也很难在统计值上追上已有类目，所以需要一个根据增量预估整体的方法。其次，在先验模型中的一个重要假设，就是召回的商品越多，越可能是相关的类目，其实并没有考虑语义上是否能保证相关，例如：“茶几”这个Query可以召回“纸巾盒”类目的很多商品，但其实文本语义上这两个意图并不相关。所以先验模型中对于召回的类目，要做语义上的判断和区分。最后，通过人工经验的规则方式将两者融合，得到最终的类目判别结果。而随着商品分布的和用户行为的变化，固定的阈值方法难以满足线上的准确率要求，需要有更合理的综合特征的方式。

基于以上的问题，我们提出了基于DNN+GBDT的类目预测融合模型，对其中的问题进行优化，并得到了准确率的效果提升。在第二部分中会详细介绍每个部分。

2. 基于GBDT融合模型的类目预测

2.1 基于反馈的模型

2.1.1 点击模型（新）

类似于排序中ctr预估的方法，通过Query和类目下的商品的历史点击行为，预测Query和类目的未来的点击行为。主要使用Query历史7天、15天的曝光、点击等统计特征。

考虑到前面提到的第二个问题曝光带来的不公平点击差异，所以在回归的目标上加上了该类目下展现pv的正则项，使得本身展现高的类目会在未来的展现比例得到一定的弱化。

2.1.2 价格&成交模型

类目之间的商品的价格差异，在区分类目意图的差异上有着重要作用。

例如，用户搜索“手机”，自然召回的商品有主件“手机”类目和配件“手机保护套”类目，我们可以通过query下的成交价以及各个类目的商品价格加入到特征中。

所以我们建立了一个基于成交和价格的回归模型，同样利用历史的统计信息作为特征主要包括

a. query在类目下的成交价格

b. 类目下商品的价格

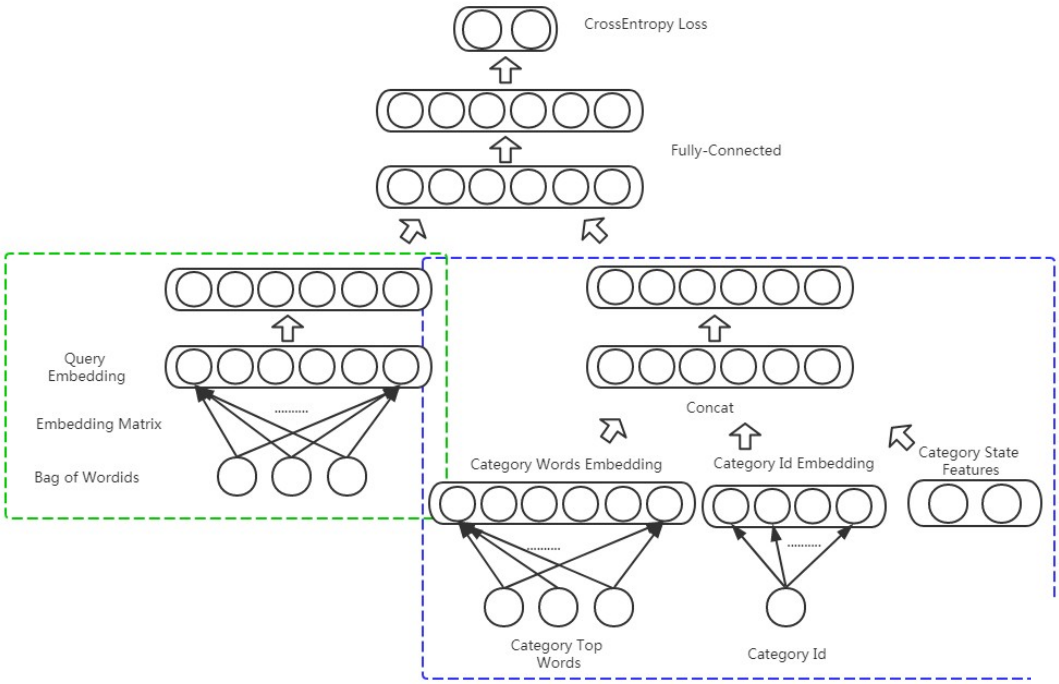
c. 类目的成交金额

d.query的成交均价等

其中的类目下商品价格，为了排除掉过高过低的商品价格，我们在模型中使用在一段时间窗口内，该类目下成交商品价格平均值。

2.2 基于DNN的先验模型（Deep Prior Model）

首先，对于Query能召回商品的类目的集合可以作为相关类目的一个较小候选集合；然后，对于这部分类目，需要计算词构成的语义的匹配程度的分数。在计算语义的匹配模型中，我们采用了多层神经网络的方法，通过词的embedding方式表示Query和类目，然后通过Query和类目的采样做目标，来训练这个网络。网络结构图如下：



2.2.1 特征说明

a. Query Embedding：用每个词id的embedding做组合，其中加入了每个词的统计权重和意图权重。词的统计权重主要为词在Query日志中的搜索pv；意图权重主要为词的标签信息设置的人工权重，例如品牌词、品类词、型号词可能在意图中有更高的权重。最终词的权重表示为 log(pv) * tag_weight。

b. Category Words Embedding：每个类目下的词，计算tf*idf，这里的idf计算将每个类目作为一个doc，即在越多类目中出现的词，约不重要。选取每个类目下的最高权重的词来表示类目，并进行带权的embedding计算（全连接）。

c. Cate Id Embedding：类目id直接做embedding

d. Cate State Feature：主要为类目下商品的平均成交价等连续值统计特征。

2.2.2 样本&采样

样本为Query和类目的pair对，主要来源于以下几个部分：

a. 当前类目与该类目路径名称（其中的词作为Query）为正样本；当前类目和与该类目具有相同父类目（只向上一层）的叶子类目的名称为负样本。

b. 对于有行为的类目，每个类目下随机抽取有行为商品，对每个商品，获取其ctr高的Query作为正样本；同时，这部分Query可以召回的非该行业或一级类目的商品类目为负样本。

c. 对于没有行为的类目和商品，对商品标题中的词做随机采样，作为正样本。此时选择组成Query的词数目有限制，选择过短的随机Query会带来很大的歧义。同时和前面的方法相同，生成的Query召回的非该行业或一级类目的商品类目为负样本。

d. 基于少量人工规则的采样，基于类目之间的互斥关系，Query的正样本类目对应的互斥类目为负样本。

2.2.3 总结&思考

1. 为什么使用点击的数据?

点击的样本在准确性上有较高的保证，而且从数据的观察发现，每个类目下的高点击的Query并不一定是和类目相关，但是Ctr高的query往往与类目的关联性很大。而且由于我们丰富了类目的表示，即类目用词来表示的同时也增加了一定的泛化性，也就是说对于词表达比较相似的类目，即使自身行为并不丰富，也可以通过与其表达相似的类目来学习得到和Query的关联性。

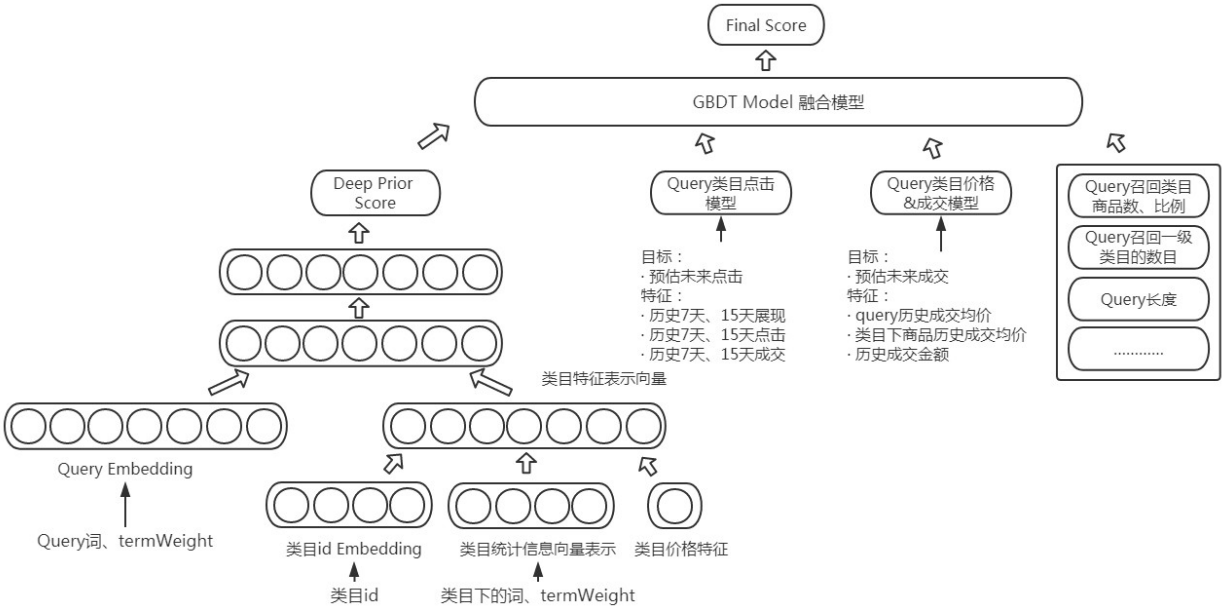
选择类目下各个商品的Query与选择类目整体的Query，前者可以带来更丰富的特征，后者可以在统计的丰富上保证准确性。

2.为什么不使用多分类目标的模型?

最开始设计方案时其实调研过多分类的方案，最后还是选择二分类的框架主要原因有两个：首先 Query的类目预测不同于商品，往往会存在多个合适的类目预测结果，所以如果采用Softmax loss的多分类，则会由于loss本身的限制，难以使得多个类目同时为正，就算是整体的平均loss最低，对于多label的样本中的loss也是比较高；另外，最重要的一点是先验模型的计算是可以预先知道候选类目的，也就是通过召回限制，可以得到Query的候选类目在一个小的集合范围中，相比多分类的全label空间，要缩小了很多倍（15000->50），而且如果
征而不是label，可以得到很多类目的统计以及描述类特征。

2.3基于GBDT的Ensemble模型

前面所述分别从行为反馈和商品分布得到不同维度的相关性描述，在线上版本的模型中对于各个特征采用分数的规则来决策得到最终的相关类目。我们改进了原有通过人工经验的方式设置相关性分档的规则，通过人工标注的样本，将多个特征通过GBDT模型做最后的决策分档模型。整体图如下：



目前在GBDT的特征包括：

a. 先验模型分数(Deep Prior Score)

b. 点击分数、价格&成交分数，在Query下所有类目的归一化、比例等。

c. Query召回类目下的商品数、占Query的商品数比例、占类目下的总商品比例。

d. Query分词长度、Query召回的类目总数（描述宽泛性的Query）

e. 类目所属的一级类目、行业等

3. 模型效果与分析

3.1 数据效果 & 页面效果

新老版本效果对比，Query搜索“电脑显示器”，类目预测的结果对比（上图为老版本，下图为新版本）

类目预测统计版 类目预测_DNN融合模型版 类目预测多分类 类目预测统计版_树结构 类目价值

类目id	类目ratio分数	类目档位	类目全路径
110502	0.66685	2	电脑硬件/显示器/电脑周边 >> 显示器&支架 >> 显示器
50009808	0.25883	2	大家电 >> 大家电配件 >> 电视机配件 >> 电视机架
50019650	0.012149	2	影音电器 >> 影音家电配件 >> VGA线
50018888	0.008979	2	3C数码配件 >> 家电影音周边配件 >> VGA切换器
50025922	0.005661	2	收纳整理 >> 家庭收纳用具 >> 收纳盒 >> 桌面收纳盒
125320002	0.003564	2	汽车/用品/配件/改装 >> 汽车影音/车用电子/电器 >> 汽车电子防盗安防 >> 抬头显示/HUD
50008759	0.00126	2	电脑硬件/显示器/电脑周边 >> 显示器&支架 >> 组装液晶显示器


query 电脑显示器 搜索

类目预测统计版 | 类目预测_DNN融合模型版 | 类目预测多分类 | 类目预测统计版_树结构 | 类目价值

类目id	类目名称	finalScore	档位	先验分数	点击分数	成交分数	召回商品数	召回商品数目比例
110502	电脑硬件/显示器/电脑周边 >> 显示器&支架 >> 显示器	0.9378289	2	0.9579	130.59901	22.03039	18821	0.09648
50009808	大家电 >> 大家电配件 >> 电视机配件 >> 电视机架	0.5268703	1	0.7181	25.45536	3.3734	50314	0.25793
50008759	电脑硬件/显示器/电脑周边 >> 显示器&支架 >> 组装液晶显示器	0.33100647	1	0.3494	17.24643	1.61429	275	0.00141

在页面BTS的商品类目对比效果：


1号测试桶结果 共214615件宝贝，当前页宝贝类目分布情况：显示器(110502):32;VGA线(50019650):2;VGA切换器(50018888):1;桌面收纳盒(50025922):1;
当前页宝贝平均价格：￥867.79 当前页宝贝中位价格：￥729.00 当前页平均销量：1,616人付款

7号基准桶结果 共214615件宝贝，当前页宝贝类目分布情况 显示器(110502):36;
当前页宝贝平均价格：￥955.22 当前页宝贝中位价格：￥899.00 当前页平均销量：1,242人付款

基准桶中的配件不相关类目已经在测试桶中不出现。

3.2 指标效果

先验模型的训练中用人工标注的样本作为validation集合，通过网络参数以及样本的分布调整，模型在验证集合的auc可以达到0.8。

人工评测整体模型时会根据算法产生的类目档位相关（2档）、不相关（1档）进行准确率和召回率的判别，主要关注的是2档类目的准确率和召回率。样本抽取时为了更好的暴露各个行业中可能存在的问题，采用分别对每个行业下按照Query pv分布进行采样的方法。在评测样本中，新方法相比老的方法在2档准确率上有**10%左右的提升（68.51%->78.75%）**，**召回率略有下降**1.61%(60.98%->59.37%)。（因为评测采样的数据样本是每个行业的覆盖量在同样的量级，所以对于较难的长尾行业的数据也会放大，相比线上真实的流量比例，是偏难的一种评测方法）。目前整体的模型数据已经在主搜PC开始BTS。

从效果看对于覆盖率的提升空间还比较大，主要源于一些宽泛Query例如搜索品牌、宽泛描述的品类等问题，对于子品类类目的覆盖率需要在先验模型的采样优化中多加考虑；而且在一些长尾行业的类目上的预测准确率还是需要提升，因为行为比较稀疏，需要更多的商品分布特征。

后续的优化空间主要在先验模型的部分和整体模型的调优，现在只有基于DNN的网络，后续可以考虑更有局部特征能力的CNN和全局性表示的LSTM作为Query和类目的表示。目前由于受限于人工标注样本的数量限制，没有办法做到end to end的整体框架模型，这个也是后续值得优化的一个方向。