URL Embedding

背景

广告landing page理解，是一项非常重要的基础技术，对于click quality预估、乱买词挖掘等核心工作具有重要意义。KR团队基于页面解析和NLP技术，对页面内容进行分析，输出核心term权重、页面topic分布等信息，为相关工作奠定了基础。同时，为了应对页面解析失效（图片、视频等）以及term赋权不准确等问题，我们也进行了很多基于用户行为的landing page摘要生成技术研究，作为重要补充。

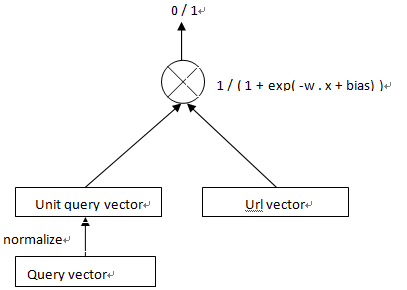
此前的页面摘要生成方法较为直接，例如将历史点击到当前url最高的top 10 term作为页面摘要，再基于topic model，word2vec等技术得到页面topic分布或vector表示。实际应用中，经常也能取得较好的效果。

Url embedding的考虑是，基于相关行为数据，统一建模并优化求解，直接得到页面embedd ing结果，避开截断选择term、结果合成等问题，同时得到理论上更优的解。

最终产出url embedding算法工具，如何收集数据，如何选取word表示方法由用户决定。这里使用的word表示方法，是从邵毅那里获取的word2vec词表。

URL to Vector

Url to vector的模型结构如下（图中x表示query vector，w表示url vector）：



输入训练数据的query vector，与url vector经过sigmoid变换后拟合数据label，增加正则项抑制过拟合。模型最终输出url to vector结果（每个url使用固定维数的向量表示）。

使用已有的query vector计算方法，而不进行联合优化的原因是：

1）Word to vector / topics的技术较成熟，直接使用更保险，联合优化可能得到不可靠的结果

2）联合优化求解更复杂，直接使用已有结果求解简单，易于并行化

训练数据可以根据实际情况选择，例如使用click quality的标注数据，或者点击与否作为正负样本，也可以考虑更加复杂的数据选取方法。

URL to Model / Distribution

前面所述的url to vector方法较为简单，一个非常明显的问题是，对于多个topic的页面，无法得到较好的效果。例如，一个页面即包含电子产品，又包含化妆品内容，word to vector无法做到同时拟合这两种语义，其本质仍然是一个线性模型。

基于此，url to model / distribution旨在将每个url表示为一个分布，或者说一个模型，其分布/模型能够表示多个topic信息。

首先考虑传统非线性模型，由于单个url的训练数据可能极少（例如<10个样本），同时query vector维度较大（通常几百维），tree model，nerual network首先被排除。Kernel-based model在这种情况下仍然可能work，因此首先尝试使用svm。针对支持向量过多的问题（导致训练和预估代价过大），通过两种方案尝试解决：

1）数据抽样

2）使用SVR减少支持向量

其它降低sv数量的简化svm方法未尝试。

实验中，方案1将训练数据抽样至<=1000，测试几个url，初步结论是建模效果弱于线性url to vector；方案2将支持向量数量限制在可接受的范围内（例如小于100），测试url中，建模效果同样差于线性方法。

至此，非线性discriminative model在这种应用场景下（高维少样本）基本被排除，接下来考虑generative model，或者说将url表示为一个分布。基于问题的特点，考虑使用gaussian mixture model（GMM）建模（实际中根据问题特点做了一些改动）。GMM与RBF kernel的SVM较为相似，主要是可变参数和求解方法上有差异。Url GMM建模主要步骤如下：

1）样本聚类

高维空间直接基于EM求解GMM，担心陷入较差的局部极值，因此首先对样本进行聚类（spherical k-means）。初始类别数量人工指定，迭代过程中允许减少类别数量。

2）GMM优化

聚类中心固定，模型参数仅包括控制类别幅值的alpha和控制kernel宽度的gamma，优化目标函数如下：

http://wiki.baidu.com/download/attachments/86030409/0f3d1c298fadedba154c2dc9b216209b.png?version=1&modificationDate=1429931109779&api=v2

其中

http://wiki.baidu.com/download/attachments/86030409/b8f25d17658d71ab855c0dc6e2766a97.png?version=1&modificationDate=1429931109780&api=v2

模型基于SGD优化求解。采用启发式初始化聚类方法，模型参数随机初始化，基本上可以得到较为稳定的局部最优解。接下来测试GMM模型效果。

这时问题来了，调研非线性模型的原因是，url to vector不能较好的拟合多topic页面，但是找了很久找不到多topic的页面。。。

针对大多数单topic页面，由于非线性模型无法得到全局最优解，同时相对更容易过拟合，GMM效果通常比url to vector差（测试了几个url，平均auc要低3%左右）。

但是假设应该是合理的，需要找到样本验证。于是我手贱的把两个url数据合成一份进行训练、测试，模拟多topic页面场景。。。，果然，GMM模型auc比url to vector高3%。。。

总结

Anyway，尽管测试的url不够多，初步判断线性的url to vector方法基本足够了，非线性GMM模型的意义仅停留在理论研究，或者特殊的多topic页面应用场景。由于每个url单独建模优化，算法很容易并行化。Url to vector的hadoop实现已经完成，能够支持一般应用。后续有空的时候，我会进一步整理好将其工具化，如果有必要，GMM版本也可以整理一下。