Title: Learning Deep Structured Semantic models for web Search using clickthrough Data

Author: Po-Sen Huang

Source: Microsoft research and university of Illinois at Urbana-Champaign

Abstract:

Latent Semantic model(LSA)， 这是隐含语义模型。它用语义将query映射到相关的文章，常用于在用关键词无法获取相关文章的情况。Dssm maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data. 为了适应web search的大数据量还采用了word hashing方法。

Introduction

我们query的关键词和文章中的关键词很可能因为表达的不同，语言的不同导致不能很好的匹配。为了解决这个问题提出了隐含语义模型。隐含语义模型通过将相似文章的不同item聚类到同一个簇从而解决web documents和query之间的语言不一致性。从LSA延伸出来的模型比如probabilistic LSA和Latent Dirichlet Allocation(LDA)，这些方法的训练用的是无监督的方法。

现在有两种拓展隐含语义模型的方法：1. 用clickthrough data进行语义建模。比如Gao提出的Bi-Lingual Topic model(BLTMs)和linear Discriminative Projection models(DPMs). BLTMs是一个通用的模型，它要求所有的topics都要有相同的分布，并且要求每个topic要有相似的词分配。DPM通过S2Net算法进行学习，follows the pairwise learning-to-rank paradigm outlined in. BLTMs模型的缺点是它通过最大化log似然criterion，这是一种次优的优化算法；而DPM包含了大量的矩阵乘法，这对于包含大量词汇的数据是不适用的。

2. Salakhutdinov利用了深度自编码器。他们认为可以通过深度学习找到documents和query之间的hierarchical semantic structure。但是这种方法还是用的无监督的方法。而且还同样面临矩阵乘法。

Dssm模型的步骤：1.首先进行query和documents在common semantic space中的非线性映射。2.计算在common semantic space中每个给定的query中的文章的余弦相似度。使用clickthroughdata分别训练神经网络模型，然后在戈丁query的情况下最大化clicked documents的条件概率分布。DSSM是一种有监督的算法。而且在DSSM模型中还提出了word hashing算法。这种算法是基于n-gram向量将高维特征下降为低维特征，而且不会有很大的损失。

Related work

这个工作主要是基于两种工作：1. 通过监督学习的方式利用click through data 2.利用深度学习

2,1 使用click through data 进行隐含语义模型构建

最典型的方法就是LSA，这种方法通过SVD将高维特征空间映射到低维特征空间，然后通过余弦相似度度量

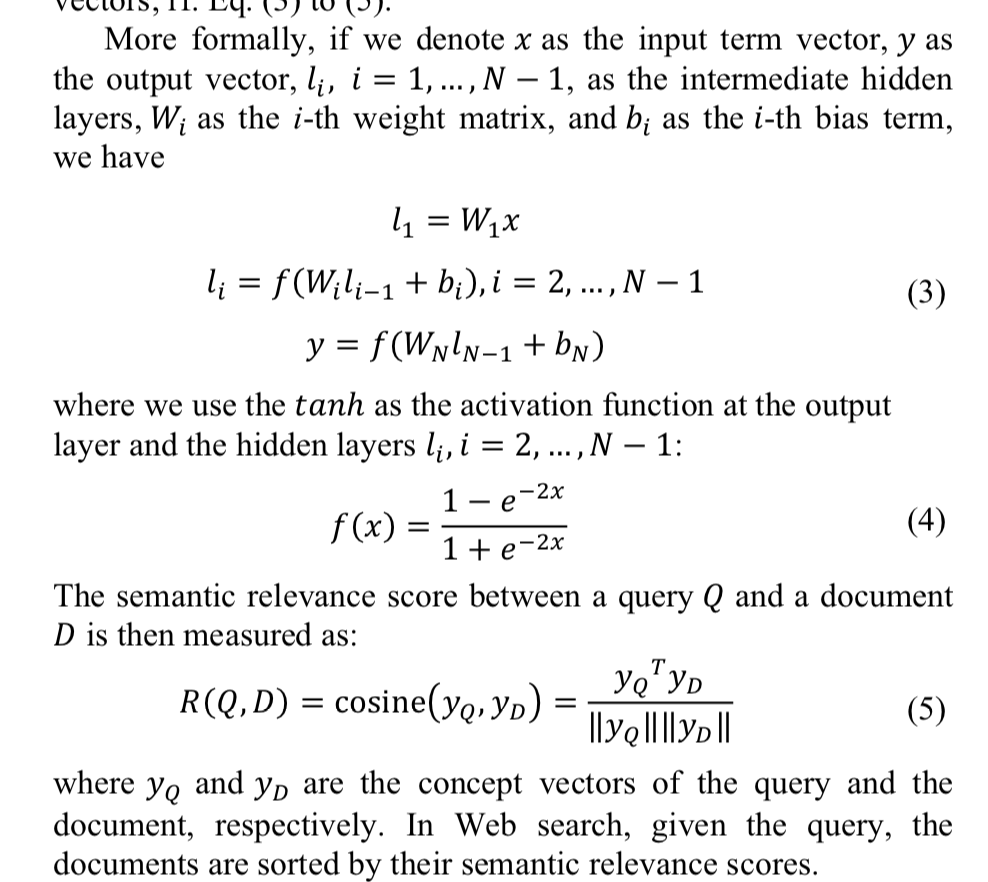
2.2 deep learning

这些深度模型用两种方法进行学习：1.用生成模型将文章的term vector representation 映射成低维的semantic concept vector。 2. 调整模型参数来最小化交叉熵。Semantic hashing模型有两个问题：1.模型的参数是通过对documents term vector进行重构来实现的，而不是通过对于给定query区分相关和不相关文章。2.文章的term vector只包含频率最高的2000个单词。

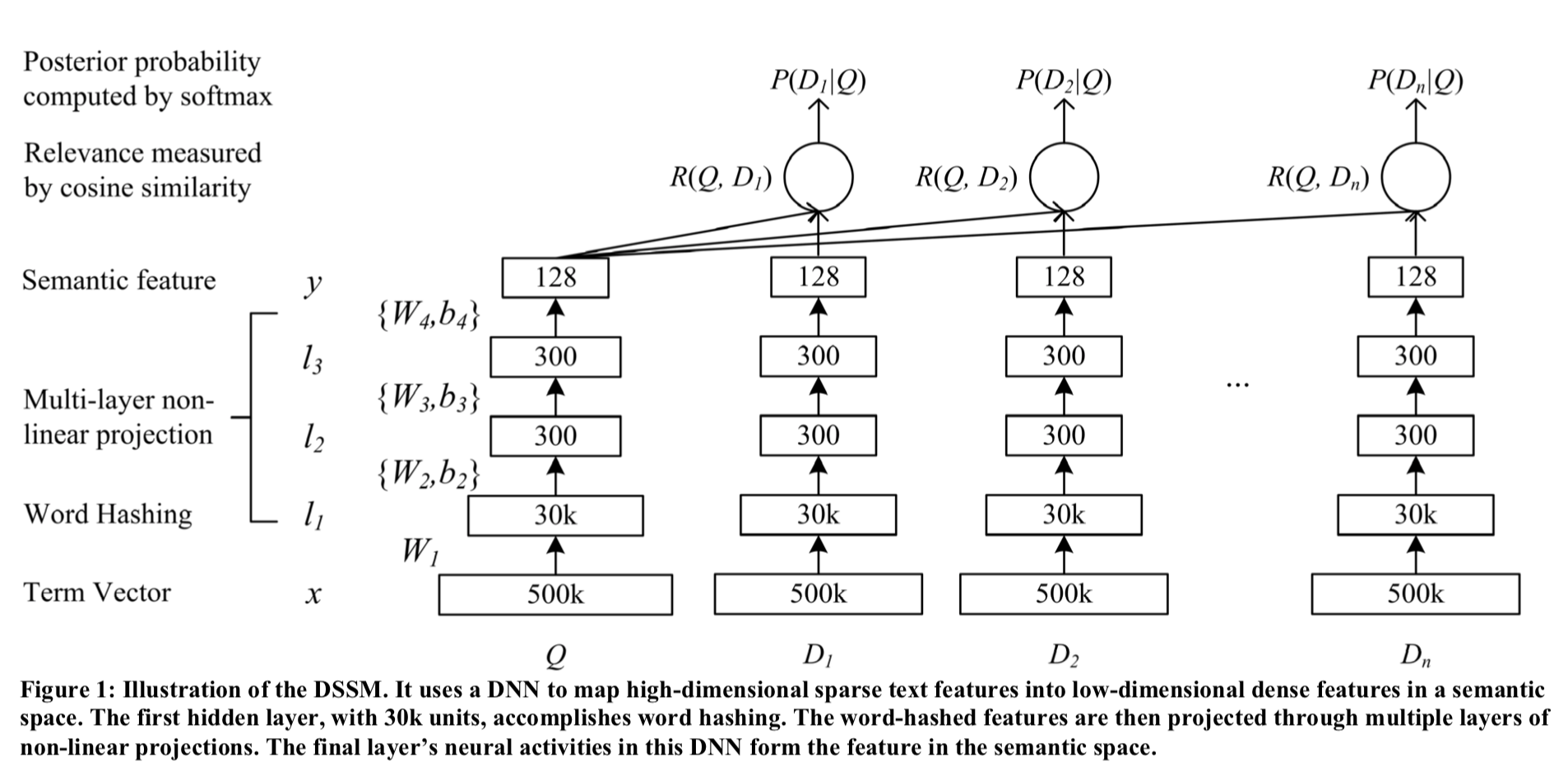
Deep Structured semantic models for web search

3.1 DNN for computing semantic features

用DNN进行web documents ranking主要进行两个步骤：1.将term vector映射到他们相关的语义内容向量 2.通过余弦相似度计算query和documents之间的相似度评分。



一种常用的item vector是词袋模型但是这种模型的缺点是当词量很大的时候它的维度很高。为了解决这个问题使用了word hashing的方法。Word hashing是DNN的第一层。

以下是DSSM模型的概念图：  


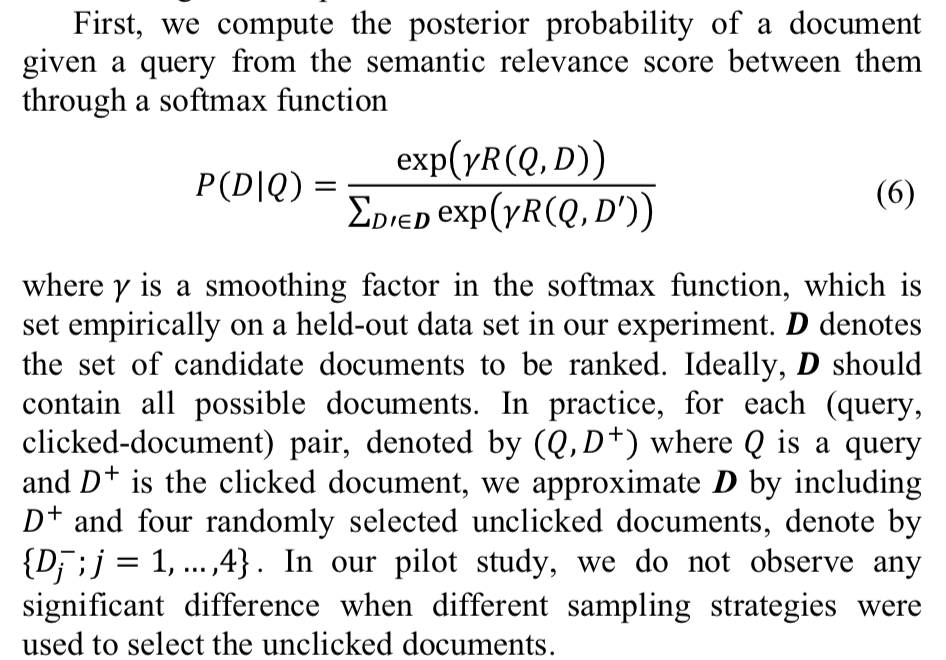
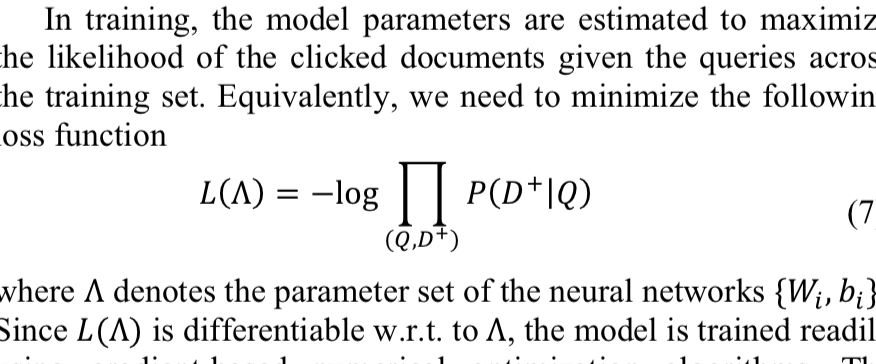
3．2 word hashing

Word hashing是为了降低词袋模型的维度，它是基于n-gram模型，同时又为我们的模型特别设计的。首先需要对每个词加入开始和结束标记，然后对于每个词将其分解成n-gram模型，最后用一个n-gram向量来表示一个词。

这种方法有一个问题就是冲突，两个不同的词可能有一样的n-gram vector representation。

虽然英语单词是无限的，但是n-gram letters是有限的。基于n-gram的word hashing 对于未知的单词是很鲁棒的。基于word hashing的n-gram可以看作是一个固定的线性变化，它能够将输入层中的term vector映射成n-gram vector。因为n-gram vector的维度比词袋模型的小的多，这就让DNN能够高效的运行。

3．3 learning the DSSM

1.首先计算给定一个query时，documents的后验概率：  
˛

3.4 implementation details

用了一个三层的网络。第一层是word hashing层，会将特征矩阵降到30k，第二层是隐含层有300个隐含层节点，第三层是输出层有128个节点。Word hashing是基于固定的映射矩阵，相似性度量是基于输出层的128维特征。在训练阶段采用mini-batch SGD方法进行训练。

Experiments

4,1 data sets and evaluation methodology

The experiments only the title fields of the web documents are used for ranking.

4.2 results

就是一些模型的结果比较

Conclusion:

We present and evaluate a series of new latent semantic models, notably those with deep architecture which we call the DSSM. The main contribution lies in our significant extension of the previous latent semantic models in three key aspects. First, we make use if the clickthrough data to optimize the parameters of all versions of the models by directly targeting the goal of documents ranking. Second, inspired by the deep learning framework recently shown to be highly successful in speech recognition. They extend the linear semantic models to their nonlinear counterparts using multiple hidden-representation layers. The deep architectures adopted have further enhanced the modeling capacity so that more sophisticated semantic structures on queries and documents can be captures and represented. Third, we use a letter n-gram based word hashing techniques that proves instrumental in scaling up the training of the deep models so that very large vocabularied can be used in realistic web search. In our experiments, we show that the new techniques pertaining to each of the above three aspects lead to significant performance improvement on the document ranking task, A combination of all three sets of new techniques has led to a new state-of-the-art semantic model that beats all the previously developed competing models with a significant margin.