

# 多路召回策略



采用不同的策略、特征或简单模型分别召回一部分候选集，然后将候选集混合在一起供后续排序模型使用。

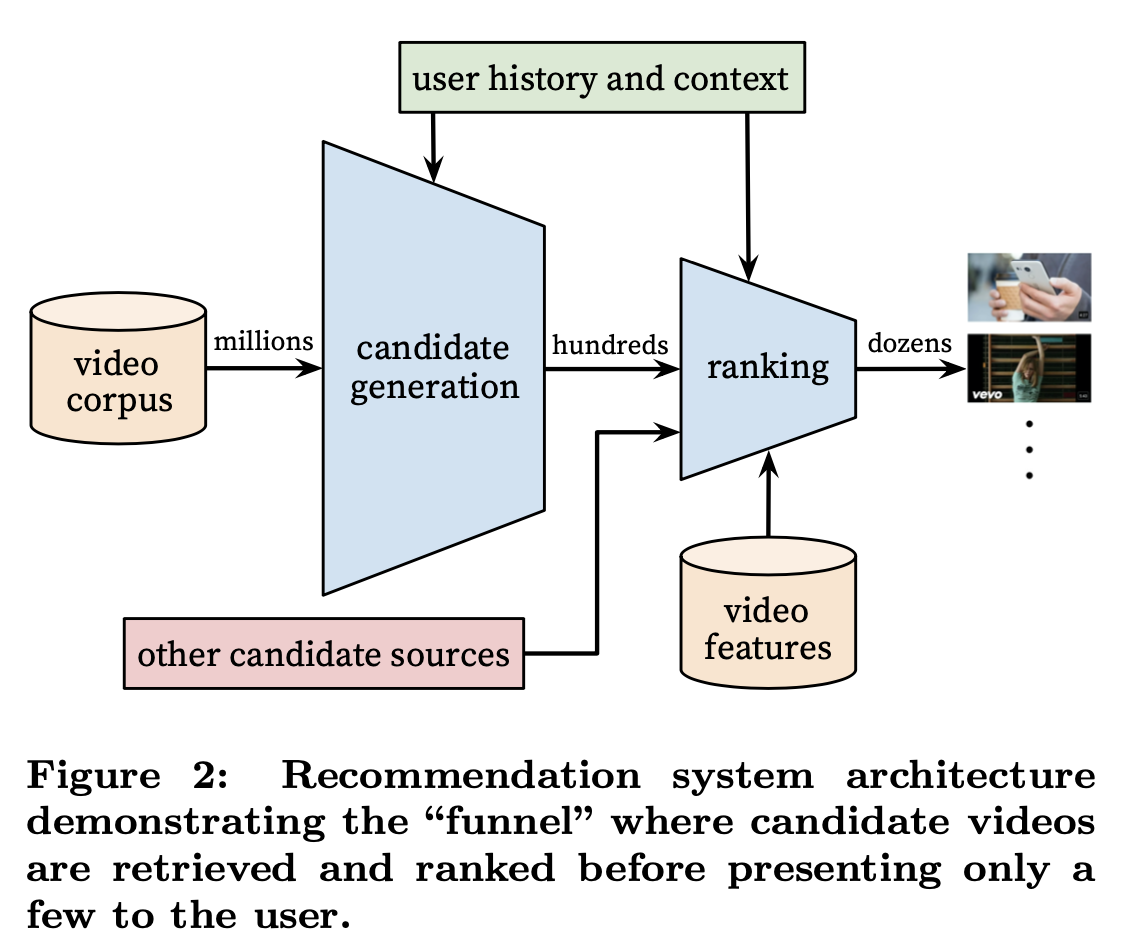
* **优点**：各简单策略能保证候选集的快速召回；召回策略的选择与业务强相关，从不同的角度设计的策略能使得召回率接近理想的状态
* **缺点**：从召回策略的选择到候选集大小参数到调整都需要人工参与；策略之间的信息也是割裂的，无法综合考虑不同策略对同一个物品的影响

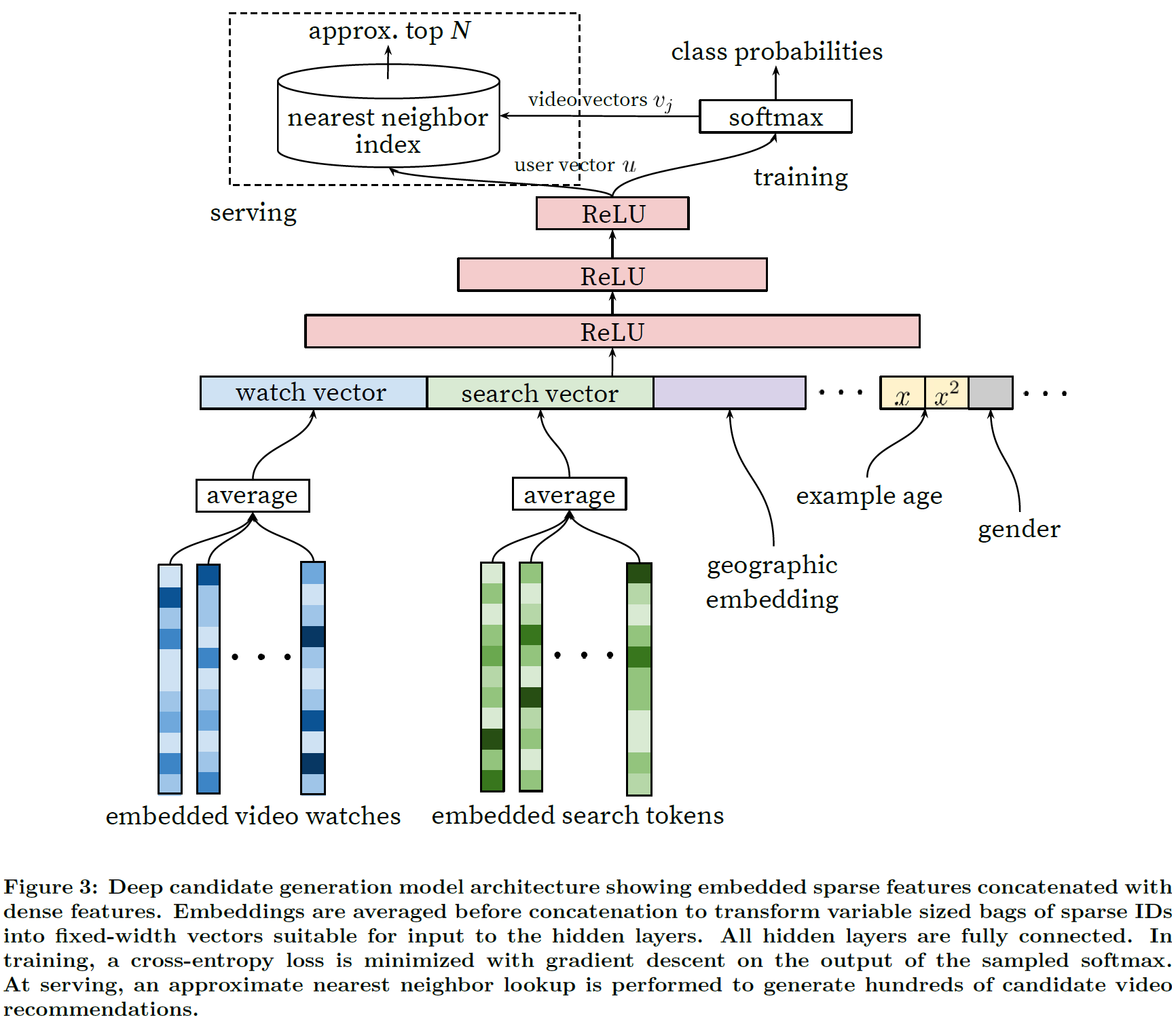
# 基于Embedding的召回方法



## YouTube DNN

### **Deep Neural Networks for YouTube Recommendations (Google 2016)**





总的来说，YouTube推荐系统的候选集生成模型（candidate generation model）是一个标准的利用Embedding预训练特征的深度神经网络模型。

#### 输入层

底层输入包括用户的历史观看视频Embedding和搜索词Embedding，以及用户的地理属性特征Embedding、性别、年龄等。

* 为了生成视频Embedding和搜索词Embedding，YouTube利用用户的观看序列和搜索序列，采用Word2Vec方法对视频和搜索词作做Embedding，再作为候选集生成模型的输入
* 除了进行预训练Embedding，还可以直接在深度学习网络中增加Embedding层，与上层DNN一起进行端到端训练，但Embedding层的存在往往会拖慢整个神经网络的收敛速度
  + Embedding层输入向量往往维度很大，导致整个Embedding层的参数数量巨大，大部分训练时间和计算开销都被Embedding层占据
  + 由于输入向量过于稀疏，在随机梯度下降过程中，只有与非零特征相连的Embedding层权重才会被更新，进一步降低Embedding层的收敛速度

#### 输出层

**模型训练（training）**

YouTube把选择候选视频集看作给用户推荐下一次观看视频（next watch），是一个多分类问题，模型最终的输出是一个在所有候选视频上的概率分布，采用Softmax函数作为输出层。

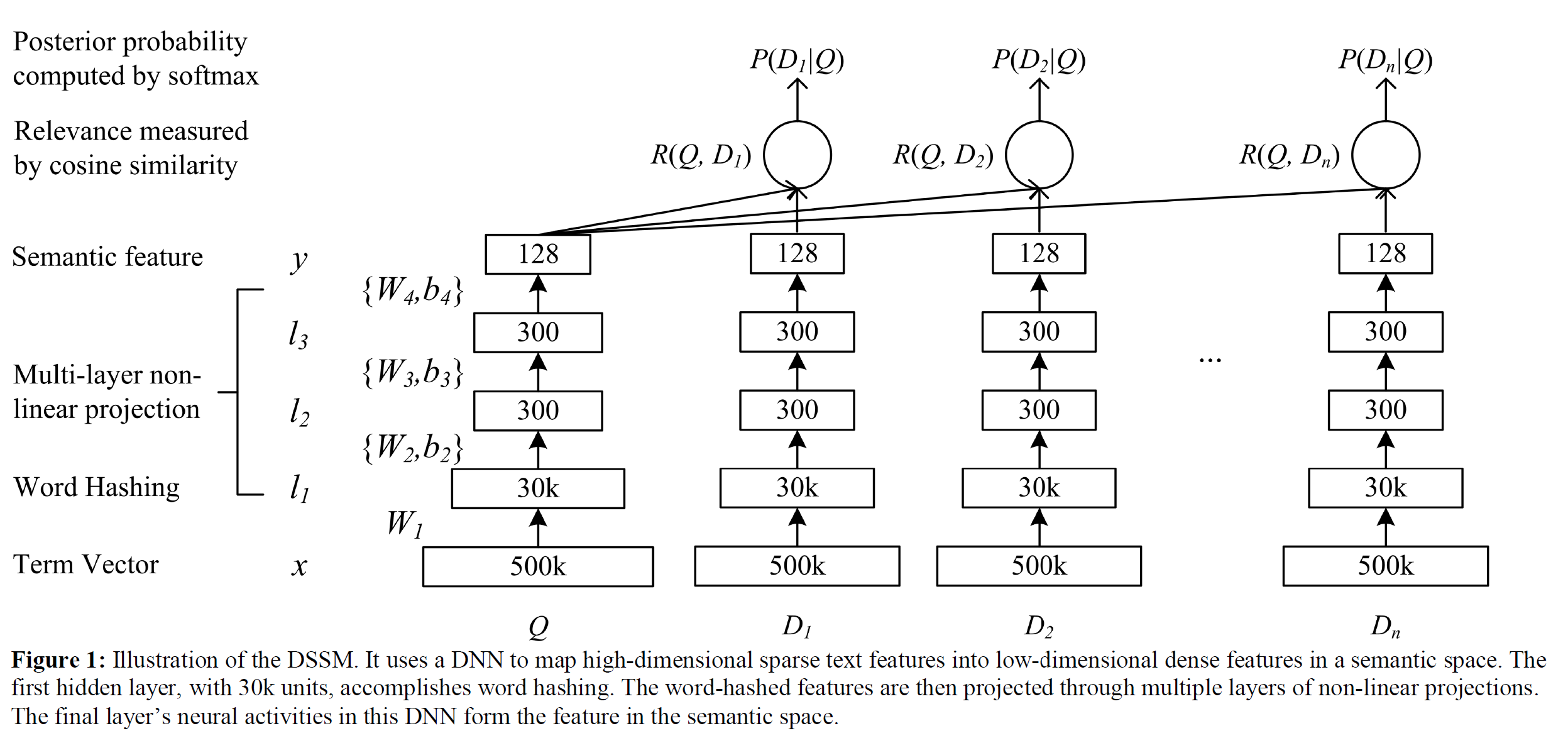
**模型服务（serving）**

在模型训练完成后，逐个输入所有用户的特征向量到模型中，得到所有用户的Embedding向量导入线上Embedding数据库。在预测某用户的视频候选集时，先得到该用户对应的Embedding向量，再在视频Embedding向量空间中利用局部敏感哈希（Locality Sensitive Hashing, LSH）等最近邻搜索算法搜索该用户Embedding向量的Top近邻，即可得到个候选视频

* 视频Embedding向量 最终输出Softmax层的参数假设是一个维的矩阵，其中指的是最后一层ReLU层的维度，指的是类别数目，即YouTube所有视频的总数，那么这个维的列向量就是视频Embedding
* 用户Embedding向量 输入特征向量都是用户相关的特征，在某用户的特征向量所谓模型输入时，最后一层ReLU层的输出向量就可以作为该用户的Embedding

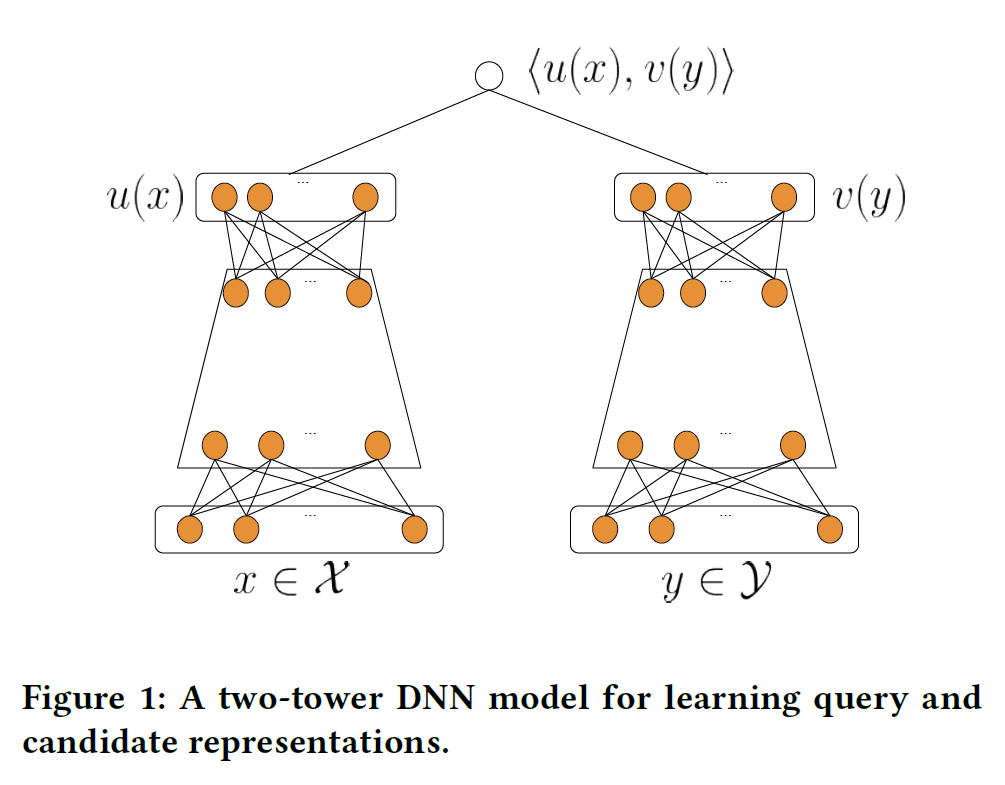
## 双塔结构DSSM

### Learning Deep Structured Semantic Models for Web Search using Clickthrough Data (Microsoft 2013)



DSSM（Deep Structured Semantic Model）是最早提出双塔结构的模型，由于用到了点击数据，相比LDA等隐语义模型，DSSM是有监督等学习。原文中单词query和文档doc等匹配和doc等匹配对应到推荐系统中是user和item的匹配。

### Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations (Google 2019)



地图的截图

描述已自动生成

YouTube双塔结构召回模型：左塔输出是包含user信息的embedding向量，右塔输出是item embedding向量，均使用DNN来学习，两塔联合训练，确保user embedding和item embedding在同一向量空间，内积才有意义。

## FM召回模型

### FM模型——隐向量特征交叉

FM模型在LR模型的基础上引入任意两两特征的交叉项，对每个特征学习一个大小为的一位向量，特征和交叉的权重系数通过特征对应向量和的内积来表示，本质上是对特征进行embedding化表示。

### 如何利用FM模型做统一召回模型

**Step1** 对于每个用户，查询离线训练好的FM模型对应的用户特征embedding向量，然后将用户特征embedding累加，得到用户embedding向量；类似地，查询物品特征的embedding向量并累加，可得到物品embedding 向量，存入在线数据库。

**Step2** 当用户登陆或刷新页面时，可根据用户ID取出其对应的embedding向量，与物品embedding向量做内积运算，按得分由高到低取Top作为召回结果。

# 统一名单召回策略

* 类比推荐系统，普惠可看成唯一“用户”，提交客户名单看成“推荐产品”的过程
  + 策略输入部分可看成“召回层”，主要负责名单的初步筛选，目标是尽可能圈中更多放款客户
  + 智能诊断部分可看成“排序层”，利用意向/资质模型打分提高名单预约率/征信通过率/转化率等指标
* **假设与放款客户在某些特征上相似的候选客户未来更有可能放款**，可在全量客户中寻找与放款客户相似的Top客户作为候选集。即每个标杆可带来个候选客户

1. 计算所有客户的Embedding向量
2. 对于每个标杆客户，利用最近邻搜索算法寻找与之相似的Top非标杆客户加入候选集，最后对候选集进行去重