## 学习排序算法

。但是，此时仅仅使用了文本语义相似度这一个特征，在数据集中还有一些没有使用的特征。综合使用其他的特征可以提升结果的准确率，

### 排序算法

正如第二章中所提到了，单文档方法仅考虑了单个文档与查询的绝对相关度，忽略了文档间的顺序关系；文档对方法考虑了任意两个文档之间的相对前后关系，相比单文档方法的效果更好；文档列表方法需要考虑每次对查询候选结果列表，当文档数量较大时，需要考虑的数量较大，相对而言没有文档对方法的效率高。综合以上原因，Tr-WELR方法模型中使用了文档对方法，并改进了当前成熟的IR SVM算法。

IR SVM的基本思想和文档对方法处理思想一致，都是将排序问题转化为分类问题。IR SVM模型将两个样本和表示成一个训练样本，然后使用SVM模型训练和预测分类。

选择排序完成后，使用较为成熟的IR SVM，进行选择排序。IR SVM是对 rank SVM 应用于 Information retrieve 时进行的改进算法，rank SVM是将排序问题转化为分类问题，应用。。。。。 SVM进行分类从而排序的算法。IR SVM相对于rank SVM 进行了一下的改进：，解决了一些ranSVM 未考虑到的问题。

Rank SVM 算法介绍：

文档对的方法， 对文档对的向量进行处理，

根据学习的依据（已知排序结果的资料），确定 权重向量，

#### Ranking SVM

Ranking SVM算法的思想是将排序问题转化为pairwise的分类问题，然后使用SVM分类模型进行学习并求解[43]。在排序应用中，假定有两组查询对应的文档集合，每组查询结果集合中有三个等级，分别是等级1、等级2和等级3。举例说明，在第一组查询结果中有三个对象、和，分别属于等级1、等级2和等级3，如图13所示。图中的向量是第二章公式(2.2)中中的权重向量。但是，此时对于每一个查询得到的候选文档集合，都有一个各自的权重向量，在应用中非常麻烦。因此，对同一组查询结果集合中的不同等级的对象的特征向量进行组合，形成新的特征向量，即把对象映射到另一个向量空间，然后根据另一个向量空间的向量进行排序。比如将上述的三个对象重新组合为：、和，并且给这些对象重新打标签，如将、和标记为负相关，相应的、和标记为正相关，如图14所示，即可将以上的排序问题转化为二值分类问题，在该例子中就可以通过训练线性SVM分类器对上述新的向量空间的向量进行分类，进而可以计算得到同一组查询结果集中的不同等级的向量的前后词序。

训练数据表示为，其中为，每一个训练实例由两个特征向量表示，标签由表示，并且的取值为+1和-1，分别表示正相关和负相关。Ranking SVM的最优化问题可以形式化的表示为公式(3.12)。

(3.12)

*s.t.(使得)*

其中为松弛变量，是训练实例的数量，2-范数：https://img-blog.csdn.net/20130624111745156?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvbGVmdF9sYQ==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center，Euclid范数（欧几里得范数，常用计算**向量**长度），即**向量**元素绝对值的平方和再开方，matlab调用函数norm(x, 2)是第二范式，把松弛变量代入公式(3.12)中，可以得到：

(3.13)

其中，加和的第一项是合页损失函数, 第二项为正则项，防止过拟合。

#### IR SVM

IR SVM是Ranking SVM在信息检索领域的一个改进[36]。Ranking SVM在信息检索领域的不足之处包括两个方面：

（1）Ranking SVM是将排序问题转化为分类问题，在学习过程中使用了0-1分类损失函数。在信息检索任务中，最终结果列表中排在前面的对检索效果的影响更大，而Ranking SVM对每两个文档的相对顺序都一视同仁。举例说明，有三个等级，其正确的排序顺序应该为等级1、等级2、等级3，则等级3>等级2和等级3>等级1都是错误的相对顺序，但是二者对Ranking SVM的训练过程造成的影响是相同的，显然，这与实际的排序过程有一定的误差。

（2）另一方面，Ranking SVM同等对待不同查询下的结果对。举例说明：每个文



图13 排序问题示意图



图14 将排序问题转化为分类问题示意图

档使用相应等级来表示，两个查询的结果如表4所示。对于查询一，在转换向量空间时可以有2个“等级3-等级2”的文档对，4个“等级2-等级1”的文档对，2个“等级3-等级1”的文档对，共8个正相关文档对可供训练；对于查询二，则有两个“等级3-等级2”的文档对，八个“等级2等级1”的文档对，四个“等级3-等级1”的文档对，共有14个正相关文档对可供训练。虽然两个查询得到的结果中等级结构相同，但是由于数量不同，查询二对Ranking SVM模型的影响会比查询一大，因此最后的结果会有偏差。同时，这也与信息检索任务中所要求的“每个查询的重要性等同”[23]是不相符的。

表4 排序列表举例

|  |  |
| --- | --- |
| **查询** | **结果** |
| 查询一 | 五个文档：等级3、等级2、等级2、等级1、等级1 |
| 查询二 | 七个文档：等级3、等级2、等级2、等级1、等级1、等级1、等级1 |

由于Ranking SVM以上两个方面的不足之处，IR SVM将二值分类问题改进为代价敏感的分类问题，即对来自不同查询的文档对，或者不同等级的文档对设置不同的损失权重。由于在排序结果列表前面的结果比后面的结果的重要程度高，在计算时会对在前边出错的文档对加大损失函数的权重，即出错的文档在列表中顺序越靠前，所需要付出的代价越高；IR SVM对上述Ranking SVM第二点不足的改进是，对排序出错的文档，如果某条查询的结果列表文档数量较少，则加大损失函数的权重，反之，如果结果列表中文档数量较多，则减轻权重，从而减少因为查询本身对查询效果的影响。IR SVM的最优化问题可以表示成如公式(3.14)所示。

(3.14)

其中表示第*i*个文档对的等级，表示等级下的权重值，表示第*i*个文档对对应的查询文档，表示查询对应的相关文档对的参数。和作为惩罚因子，分别用来弥补上述Ranking SVM两个方面的不足。参数值的确定是使用一个启发式算法：选定一个评价指标，对每个查询，找到能够使得评价指标最优的排序序列，然后随机交换任意两文档的位置，查看评价指标的降低值，重复该过程，最后对降低值取平均值作为分类下对查询文档对所加的权重。参数值是对查询项对应的相关文档的数量取倒数，当相关文档数目较少时，该值相对较大，进而弥补了在文档数较少时Ranking SVM优化函数对其重视程度低的问题。