0、各位老师好，我是来自计日1401班的刘智强，我的毕设题目是基于排序学习的生物医学文献检索

1、我将从以下课题来源、研究内容及研究方案和进度安排四个方面对毕设任务进行介绍：

2、首先是课题来源

3、随着互联网的普及，信息出现爆发式增长，信息检索成为了一个重要研究课题

在排序学习之前，主要有两种传统的信息检索排序模型：基于查询与文档相关性的模型如概率模型和基于重要性的模型如PageRank算法

但传统模型一般都设计人工调参，过程繁琐且易出现过拟合；随着现在可用于排序的特征越来越多，传统方法难以处理这些多维复杂的排序特征；再者，现在基于单个模型得到的排序结果难以满足用户对准确度的需求

而机器学习可以对这些问题进行改善，因此研究人员便尝试利用机器学习技术解决排序问题，便并由此产生了排序学习

4、第二部分是研究内容

5、这是排序学习的模型框架，其中包含数据集、训练系统，排序系统和性能评估。据此可将研究内容分为以下几个部分。

1获取数据集，因为需要应用在生物医学领域，所以可以从一些生物医学数据库入手。

2数据预处理，将数据集对通过特征函数转化成特征向量，便于后续的训练。

3训练模型，利用排序学习的各种方法对训练集进行训练，可以借鉴一些排序学习工具包。稍后会再对排序学习算法进行介绍

4模型评估，利用评价指标对模型性能进行评价，发现问题，并尝试对其进行改进以

6、随着机器学习方法的发展，排序学习也出现了越来越多的算法，根据样本空间和损失函数的定义方法不同将这些方法分pointwise、pairwise和listwise三类方法。

pointwise方法将每个查询文档对看做样本点，对单个文档进行预测打分，将排序问题转化为回归、分类、有序回归等机器学习方法。

pairewise方法则将同一查询对应的一对文档看做一个样本点，对这对文档的偏序关系进行预测。一般都转化为二分类问题。蓝色框标出的三种方法分别利用了机器学习中的支持向量机，集成学习中的adaboot还有神经网络这几种模型。

Listwise方法将某个查询对应的所有文档看做样本点，更接近问题本质较前两种方法复杂度更高，但效果上优于前两类方法。

7、第三部分是研究方案

8、对应研究内容可设计以下研究方案：

获取数据集方面：

TREC是文本检索领域最具权威性的评测会议之一

2017年新增了Precision Medicine track

其中的数据集来源于MEDLINE数据库，前面提到过的PubMed生物医学搜索引擎的数据库之一就是MEDLINE。因此满足我们对数据集的要求

数据预处理方面：

需要编写预处理程序，构造特征向量，

在特征构造的过程中可以将现有的一些模型，如BM25和PageRank的输出用作特征

训练模型方面：

RankLib是一个开源的排序学习方法库

实现了RankNet等 8 种流行的方法

学习RankLib方法库的使用，可以快速熟悉各种模型

最后是模型评估方面：

常用的评估指标包括MAP、NDCG、MRR等

9、最后一部分是进度安排

10、

11、以上就是全部内容，感谢各位老师批评指正

1、为什么没有现状介绍？

一是因为排序学习的主要算法还是集中在前面介绍的三类方法中，现在只是对其中的一些方法进行优化，加之时间有限，所以就没有重复介绍；二是现在一些新的结合深度学习的排序学习方法还在探究中，加之自己对深度学习了解不够，就没有对其进行介绍，但这其实也在我的毕设任务考虑之中，将现有的经典模型掌握之后便计划学习深度学习算法，结合排序学习，尝试对现有模型进行改善；关于现状，近几年，排序学习在学术界和工业界都取得不少成果。SIGIR等国际顶级会议将排序学习作为一个主要的Session或Track，很多知名的搜索引擎公司、推荐系统和大型电子商务平台也依赖排序学习算法为用户提供精准的搜索和推荐结果。排序学习在未来几年仍旧是非常热门的研究领域。