大连理工大学本科毕业论文开题报告

基于排序学习的生物医学文献检索

Learning to Rank Based Biomedical Information Retrieval System

学 院（系）：电子信息与电气工程学部

专　　业：计算机科学与技术（日语强化）

学 生 姓 名：刘智强

学　　号：201487020

指 导 教 师：杨志豪

开题报告日期：2019.03.21

大连理工大学

Dalian University of Technology

说　　明

一、开题报告应包括下列主要内容：

1. 课题来源及研究的目的和意义；
2. 主要设计要求；
3. 国内外在该方向的研究现状及分析；
4. 主要研究内容；
5. 研究方案及预期达到的目标；
6. 为完成课题已具备和所需的条件和经费；
7. 主要参考文献。

二、开期报告由指导教师填写意见、签字后，统一交所在院系保存，以备检查。

指导教师评语：_____

_____

_____

_____

_____

_____

_____

指导教师签字：_____　　检查日期：_____

1　课题来源及研究目的和意义

1.1 课题来源

人们通过信息来了解世界，获取知识，其重要性不言而喻。而如何获取所需信息一直是人们面临的一个难题。在信息匮乏的年代，人们面临的是“巧妇难为无米之炊”的困境；伴随着互联网的迅猛发展，信息的传递不再受物理因素所约束，信息出现爆发式增长，每时每刻都会有大量的来自世界各地的信息汇入互联网中。可获取所需信息这一难题依然伴随着我们，只是这次是因为信息量过于庞大，且质量参差不齐。如何准确地从海量的数据信息中获取用户所需要的信息成为重要的研究课题，即信息检索技术，其本质即信息的排序问题。

在排序学习之前，主有两种传统的信息检索排序模型：基于相关性的排序模型如向量空间模型、概率模型和语言模型，基于重要性的排序模型如 PageRank 算法、HITS 算法。但这些模型算法有两个突出的问题：一是这些模型中一般都涉及到一些需要人工调整的参数，这些参数难以调整且容易出现过拟合；二是随着可供排序的特征不断增加，传统模型不再适合处理如此多维复杂的排序特征；另外随着研究的不断深入，更多的排序模型被提出，由单一模型生成的排序结果难以满足用户对排序结果准确率的需求，因此需要一种新的可以将不同排序方法进行融合利用的排序模型。而机器学习具有可自动调参、易于融合多个模型的特点。于是许多研究人员便尝试使用机器学习的一些方法来解决排序问题，排序学习便并由此产生。

1.2 研究目的和意义

排序学习在信息检索、自然语言处理、文本挖掘等领域有着十分广泛的应用，典型应用有文献检索、专家系统、定义查询、协同过滤、问答系统、关键词提取、文档摘要还有机器翻译[1]。在医学文献检索领域，排序学习同样起着重要的作用。PubMed[2]是世界上搜索生物医学文献最多的搜索引擎之一，其中包含 2800 多万篇文章摘要，并且还在快速增长。一个普通的搜索就可能返回成百数千篇文档，生物医学文献的急速增加使得医学研究人员，临床医生，医疗服务人员和普通大众难以找到需要的生物医学信息。现代医学发现很多时候是由一些小的灵感发展壮大形成，而这些灵感通常又都是在对相关领域里众多医学研究进行信息整合过程中碰撞而出。因此，按研究人员需求准确地检索出相关领域的信息将为医学研究提供极大的便利。可以说，一个准确有效的生物医学检索系统能为医学研究与人类健康提供有力的支撑和保障。

2 主要设计指标

在研究中主要通过将模型生成的排序列表与实际排序列表进行对比来对模型性能进行评价。具体的评价指标通常包括 MAP 和 NDCG。

2.1 MAP(Mean Average Precision)

AP 指对于单个查询的平均准确率，等于每篇相关文档的准确率的平均值。而 MAP 就是对所有查询的平均准确率求平均值。

2.2 NDCG(Normalized Discounted Cumulative Gain)

对于某个查询返回的所有文档，我们根据相关度对其进行打分。CG 指所有文档的分数和，可以看出 CG 与文档的返回顺序无关，分数高低只代表结果页面总体质量高低而无法衡量排序效果，因此引入位置因素。根据位置前后对分数进行折扣，位置越后折扣力度越大，再对折扣后的分数进行求和，便可得到 DCG[3]。但不同模型返回的结果数量不一，通过 DCG 仍无法比较两个模型的效果，因此我们便对 DCG 进行标准归一化处理，即用返回结果的 DCG 除以完美结果的 DCG，得到一个小于等于 1 的数，这就是 NDCG。

3 排序学习研究的历史和现状

通常我们将使用了机器学习技术解决排序问题的方法都统称为排序学习方法，近些年，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为三类：pointwise 方法、pairwise 方法和 listwise 方法。[4]

pointwise 方法样本空间中的一个样本是单个文档和对应查询构成的特征向量，损失函数评估单个文档的预测得分。2002 在 NIPS 会议上提出基于有序回归的 Pranking[5]算法，2006 年在 COLT 会议上又提出基于回归的 Subset Ranking[6]算法，接着 2007 年基于分类的 McRank[7]算法在 NIPS 会议上发表。以上方法都属于 pointwise 方法，只是根据使用的机器学习方法不同，将其进一步分为基于回归、基于分类和基于有序回归的算法。

Pairwis 方法样本空间中的一个样本是同一查询对应的两个文档和对应查询构成的一对特征向量，损失函数评价预测的偏序关系与实际偏序关系的差异程度。2000 年提出基于 SVM 的 RankingSVM[8]算法，2003 年在 JMLR 会议上提出基于 AdaBoost 的 RankBoost[9]算法，2005 年在 ICML 会议上提出基于神经网络的 RankNet[10]算法，2006 年在 SIGIR

会议上提出 IR SVM[11]算法，该算法针对 Ranking SVM 算法的一些不足进行改进。2007 年在 SIGIR 会议上提出基于提升树的 GBRank[12]算法。

Listwise 方法较前两种方法复杂度更高，但效果上优于前两种方法。Listwise 方法样本空间中的一个样本由一个查询及其对应的所有文档构成的特征向量列表。而跟据损失函数不同的构造方式可将 listwise 方法分为两类，一类为直接优化评价指标的算法和的算法。2007 年，在 LR4IR 会议上提出 SoftRank[13]算法，在 SIGIR 会议上提出 SVM-MAP[14]算法和 AdaRank[15]算法，这三种算法都是直接优化评价指标的方法。在 2007 年和 2008 年的 ICML 会议上分别提出 ListNet[16]和 ListMLE[17]两种直接定义 listwise 损失函数的算法。

总的来说，Pointwise 算法都将排序转化为回归、分类、有序回归。Pairwise 算法几乎都转化为二元分类问题，这两类方法的优点是可以直接应用 ML 中许多现有的理论和工具，但同时也忽略了检索问题的特殊性。例如，大多数 IR 评估指标都是查询级别的并且是位置相关的，但在这两类算法都没有使用查询信息和位置信息。而 Listwise 算法将排序视为新问题并为其定义特定算法，可以在训练模型中更好地利用查询信息与位置信息。但同时 listwise 算法通常会比 pointwise 算法和 pairewise 算法更加复杂，另外还需要一个可以对 listwise 算法行为进行解释的理论基础。[18]

近几年，排序学习在学术界与工业界在学术界和工业界都取得不少成果。SIGIR、WWW、WSDM、CIKM 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。但该研究领域也仍然存在许多有待探讨的问题，在未来几年仍旧是非常热门的研究领域。

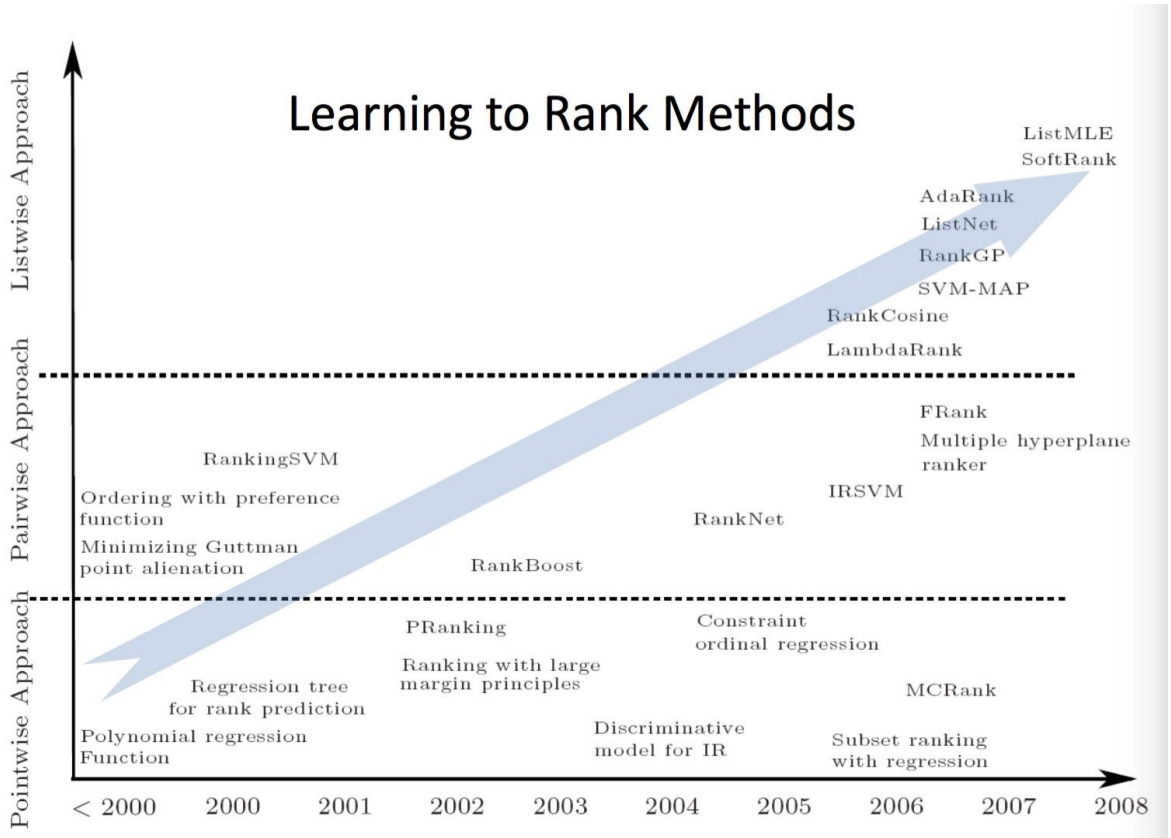

图 1：Learning to Rank Methods

4　主要研究内容

图 2 为排序学习的模型框架，其中包含训练集与测试集，排序学习模型，其中包含数据集、训练系统，排序系统和性能评估。以此为根据可将研究内容主要包括以下几个部分。

1、获取训练数据集。排序学习属于监督学习，具有训练集与测试集，所以首先要做的便是为实验获取数据集。这次实验用到的数据集主要来源于 MEDLINE，第 1 节有提到过 PubMed 搜索生物医学文献最多的搜索引擎之一，PubMed 的数据库来源之一就是 MEDLINE，因此从 MEDLINE 中获取实验数据集是可行并符合生物医学文献搜索这一研究课题的。2、特征工程，对获取的数据集进行预处理，将每一个查询文档对通过特征函数转化成特征向量，便于后续的训练，通常使用的特征可以是查询词在文档中出现的频率，文档之间的关系，也可以是一个现有的检索模型的输出，如 BM25 和 PageRank 的得分等。3、利用排序学习的各种模型对训练集进行训练。4、分析实验结果，利用评价指标对模型性能进行评价，找出影响算法性能的原因，并尝试对其进行改进以获得更

好的排序模型。5、将最近比较热门的机器学习算法与排序学习算法相结合[19]，例如将深度学习与排序学习结合，测试新的机器学习算法在信息检索任务中的表现。

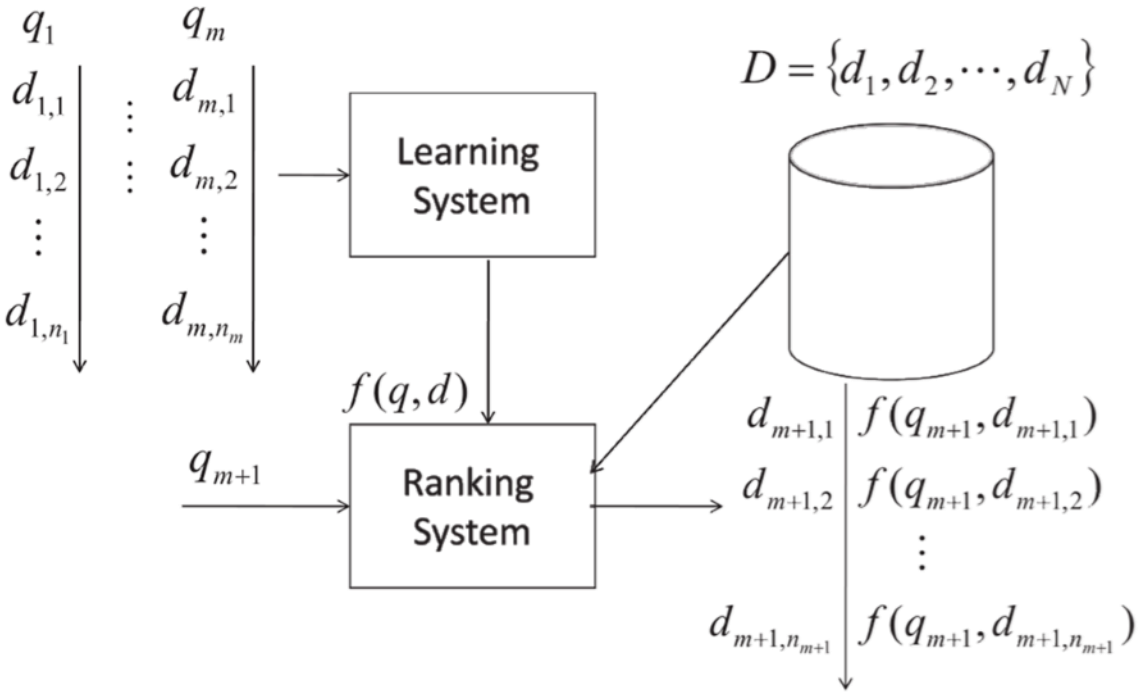

图 2：Learning to Rank Approach Architecture

5　研究方案及预期达到的目标

5.1　研究方案

根据研究内容可设计以下研究方案：1、通过 TREC 下载数据集。TREC（Text REtrieval Conference ）是文本检索领域人气最旺、最权威的评测会议，由美国国防部和美国国家技术标准局（NIST）联合主办。Precision Medicine 是 TREC 2017 年的新任务，该任务就包含 MEDLINE 医学科研文献的摘要部分，可以直接用来进行检索。计划用 TREC 2017 的数据集当作训练集，TREC 2018 的数据集当作测试集。2、利用 python 编写预处理程序，对原始数据进行特征构造。3、RankLib 是一个基于 Java 语言的排序学习方法库的开源实现，当前已经实现了 8 种流行的方法，包括 MART、 RankNet、 RankBoost、 AdaRank、 Coordinate Ascent、 LambdaMART、ListNet 和 Random Forests。它实现了许多信息检索的评价标准，同时还提供了多种执行方式去实施评价。利用 RankLib 加载不同算法对数据进行训练。4、对测试集进行预测并对模型进行评估。

5.2 预期达到的目标

在经典模型的复现方面，期望在各个模型上都能获得较为稳定良好的结果，且各个模型的优劣也符合调研结果，以此证实前期调研的可信性；在此之上，结合现在的一些深度学习算法，期望能获得更好的结果。

6 为完成课题已具备和所需的条件

经过前期学习对排序学习的各种算法有一定程度的了解，但还需更深层次的理解；已经获取研究所需数据集，但还需进一步对数据进行处理；对 Python 语言比较熟悉，但还需学习 RankLib 工具包的使用方法；为了结合深度学习算法，需要对深度学习进行更系统性的学习。

参 考 文 献

[1] Li H. Learning to rank for information retrieval and natural language processing[J]. Synthesis Lectures on Human Language Technologies, 2011, 4(1): 1-113.

[2] Fiorini N, Leaman R, Lipman D J, et al. How user intelligence is improving PubMed[J]. Nature biotechnology, 2018, 36(10): 937.

[3] Jarvelin K, Kekalainen J. IR evaluation methods for retrieving highly relevant documents[J]. international acm sigir conference on research and development in information retrieval, 2000, 51(2): 41-48.

[4] Li H. A short introduction to learning to rank[J]. IEICE TRANSACTIONS on Information and Systems, 2011, 94(10): 1854-1862.

[5] Crammer K, Singer Y. Pranking with Ranking[C]. neural information processing systems, 2001: 641-647.

[6] Cossock D, Zhang T. Subset ranking using regression[C]. conference on learning theory, 2006: 605-619.

[7] Li P, Wu Q, Burges C J, et al. McRank: Learning to Rank Using Multiple Classification and Gradient Boosting[C]. neural information processing systems, 2007: 897-904.

[8] Herbrich R. Large margin rank boundaries for ordinal regression[J]. Advances in large margin classifiers, 2000: 115-132.

[9] Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences[J]. Journal of machine learning research, 2003, 4(Nov): 933-969.

[10]Burges C J, Shaked T, Renshaw E, et al. Learning to rank using gradient descent[C]. international conference on machine learning, 2005: 89-96.

[11]Cao Y, Xu J, Liu T, et al. Adapting ranking SVM to document retrieval[C]. international acm sigir conference on research and development in information retrieval, 2006: 186-193.

[12]Zheng Z, Zha H, Zhang T, et al. A General Boosting Method and its Application to Learning Ranking Functions for Web Search[C]. neural information processing systems, 2007: 1697-1704.

[13]Taylor M, Guiver J, Robertson S, et al. Softrank: optimizing non-smooth rank metrics[C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 77-86.

[14]Yue Y, Finley T, Radlinski F, et al. A support vector method for optimizing average precision[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 271-278.

[15]Xu J, Li H. Adarank: a boosting algorithm for information retrieval[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 391-398.

[16]Cao Z, Qin T, Liu T Y, et al. Learning to rank: from pairwise approach to listwise approach[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 129-136.

[17]Xia F, Liu T Y, Wang J, et al. Listwise approach to learning to rank: theory and algorithm[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1192-1199.

[18]Liu T Y. Learning to rank for information retrieval[J]. Foundations and Trends® in Information Retrieval, 2009, 3(3): 225-331.

[19]Li H, Lu Z. Deep Learning for Information Retrieval[C]. international acm sigir conference on research and development in information retrieval, 2016: 1203-1206.