復旦大學

本科毕业论文

论文题目：大规模生物医学文本自动索引研究

院　　系：计算机科学技术学院

专　　业：计算机科学与技术

姓　　名：刘 坷

学　　号：10300240063

指导教师：朱山风

2015 年 6 月 5 日

摘要

医学主题词(MeSH)被美国国家医学图书馆(NLM)用来索引在生物医学文献数据库（MEDLINE）中几乎所有的文档（citations），极大地促进了生物医学信息检索和文本挖掘的应用。为了减少手动标注的时间和经济成本，NLM 开发了一个软件程序包——医疗文本索引器(MTI),这个软件程序包采用最近邻算法(KNN)，通过模式匹配和索引规则来协助（MeSH)的标签。通过 MeSH 分类器进行预测，也可将其他类型的信息用于自动主题词(MeSH)标签。但是，现有的方法不能有效地整合多个证据（evidence）来进行 MeSH 标签。

我们提出了一个新的框架——MeSHLabeler，通过“Learning to rank”集多种方法，实现准确标注（MeSH)。特征包括（MeSH)分类器的预测结果，KNN，模式匹配，医学文本索引（MTI）以及不同的（MeSH term）间的关联等等。每个（MeSH)的分类器是独立训练的，因此不同分类器的预测分数是无法比较的。为了解决这个问题，我们开发了一个有效的预测分数标准化过程，来提高预测准确性。

2014 年，在欧洲信息检索评测平台举办的会议上主办的大规模生物医学文献语义索引挑战赛(BioASQ)中,MeSHLabeler 赢得了任务 2A 第一名,在由 BioASQ 提供的 9,040 份文档中，取得了 Micro F-measure 0.6248 的成绩，这比由医疗文本索引器（MTI）获得的 0.5724 的精度高出约 9.15％。

关键词：算法，医学主题词，大规模多标签分类，MeSHLabeler

Abstract:

Medical Subject Headings (MeSH) are used by National Library of Medicine (NLM) to index almost all citations in MEDLINE, which greatly facilitates the applications of biomedical information retrieval and text mining. To reduce the time and financial cost of manual annotation, NLM has developed a software package,

Medical Text Indexer (MTI), for assisting MeSH annotation, which uses k-nearest neighbors (KNN), pattern matching and indexing rules. Other types of information, such as prediction by MeSH classifiers (trained separately), can also be used for automatic MeSH annotation. However, existing methods cannot effectively integrate

multiple evidence for MeSH annotation.

We propose a novel framework, MeSHLabeler, to integrate multiple evidence for accurate MeSH annotation by using "learning to rank". Evidence includes numerous predictions from MeSH classifiers, KNN, pattern matching, MTI and the correlation

between different MeSH terms, etc. Each MeSH classifier is trained independently, and thus prediction scores from different classifiers are incomparable. To address this issue, we have developed an effective score normalization procedure to improve the prediction accuracy.

MeSHLabeler won the first place in Task 2A of 2014 BioASQ challenge, achieving the Micro F-measure of 0.6248 for 9,040 citations provided by the BioASQ challenge. Note that this accuracy is around 9.15% higher than 0.5724, obtained by MTI.

Key words: algorithm, Medical Subject Headings (MeSH), large scale multi-Label classification, MeSHLabeler

目　录

第一章　绪论

1.1 研究背景

阅读科学文献是广大生物医学研究人员跟踪科学进展，获取和更新知识的一个重要途径。随着生物医学研究领域的突飞猛进，新的科学发现层出不穷，与之相关的论文数量也迅速增长。作为最大的生物医学文献数据库，MEDLINE 覆盖了全世界 5600 多种学术期刊。MEDLINE 文档(citation)由医学主题词 MeSH(Medical Subject Headings)来标注。美国国立医学图书馆 NLM（National Library of Medicine）收录的文件、书籍和音像也通过 MeSH 做分类标注。MeSH 采用分级标签的组织构架，每年都做少量的更新，截止 2014 年底最新版的 MeSH 包含了 27,455 个主题词（main headings）[1]。在 MEDLINE 中，平均每篇文档由 13 个主题词来描述其内容。MeSH 的主要用途是文献的索引，MeSH 还被广泛用于生物医学信息检索和生物医学文本挖掘等任务。因此 MeSH 的精确标注对于广大研究人员意义重大，不仅能够帮助研究人员快速找到所需文档，而且为研究人员进一步挖掘生物医学文献内容、提出新的科学假设、发现新的知识奠定坚实基础。

由于 MEDLINE 每年增长约 70 万文章，为新的文档标注 MeSH 成为一个越来越困难的任务。目前，美国国立医学图书馆（NLM）雇用专人手工标注 MeSH，平均每篇文档标注约花费 9.4 美元(Mork et al.,2013)。随着 MEDLINE 每年添加文档数目的迅速增长，标注所需的财务和时间成本越来越高。而且由于标注人员知识、背景等不同，人工标注带有相当强的主观性和不一致性，因而检索的结果往往不能十分精准，难以满足专业科研人士对高精度检索的需要。因此，我们越来越需要一种高效而又准确的方法来对新增文献进行标注。

1.2 国内外研究现状

为文档（citation）标注 MeSH 关键词的问题，可以看作是文本分类（Text Classification）问题，每个 MeSH 是一个类（Category）。对于一个 citation，考虑每个 MeSH 是否可以分到该类，即是否可以将该 MeSH 关键词分给该 citation

1 http://www.nlm.nih.gov/pubs/factsheets/mesh.html

作为其标签（Label）。

因为每篇文档具有多个 label，即多个 MeSH 关键词，所以，这是一个多标签分类问题(Multi-Label)。又因为 MeSH 本身是具有其特定的结构和层次化的，所以这是一个层次化多标签分类问题 HMC（Hierarchical Multi-Label)。这种问题在很多研究中都会有涉及，比如文本筛选(document filtering)，文本分类（text categorization)（ Lewis et al, 2004)，网页挖掘（web mining)（Liu et al, 2005）和社会性书签系统中的标签推荐（tag reco-mmendation in social bookmarking system)（Katakis et al, 2008 ）等。多标签分类问题的处理，已有很多方法被提出，包括 margin-based methods，结构化支持向量机（structural support vector machines)（Tsochantaridis et al, 2004)，参数混合模型(parametric mixture models)(Ueda et al, 2002),K 近邻(k-nearest neighbors)（Zhang et al, 2007)，最大熵模型（maximum entropy models)（Ghamrawi et al, 2005; Zhu et al, 2005)，和集成方法（ensemble methods)（Tesic et al, 2007)。这些方法大多数利用了多标签之间的关系，用作集合推断（collective inference)。也有一些方法尝试找到一个多标签共享的低维子空间（low-dimensional subspace shared among multiple labels)（Ji, et al, 2008)。

1.2.1 国内研究现状

清华大学 Minglie Huang 博士和美国国立生物信息技术中心(NCBI)Zhiyong Lu 博士等将医学主题词自动标注问题视为一个排序学习问题（Learning to Rank)，提出 L2R 方法（Huang et al, 2011; Mao et al, 2013;2014)。对于每个目标文档，L2R 将相似文档的医学主题词作为候选，这个问题就是一个对候选医学主题词进行排序学习的问题。为了区分各个医学主题词的重要性，L2R 使用了多种特征,其中最重要的特征包括基于 KNN 的得分、检索得分、MTI 的结果等。它的一个明显缺点是没有考虑医学主题词的全局分类器信息。在 BioASQ2014 挑战中，尽管 NCBI 进一步将全局分类器的分类结果融入，但是他们没有考虑具体预测分数，因而遗漏了其中蕴含的重要信息。除此之外，他们在预测 MeSH 数目方面采用比较粗糙的启发式方法。

1.2.2 国外研究现状

为了有效和高效地完成主题词索引任务，美国国立图书馆从 2004 年起开发

MTI(Medical Text Indexer)软件协助 MEDLINE 文献的标注任务(Aronson et al,2004;Mork et al,2013，2014)。 MTI 包括两个重要组成部分，MetaMap Indexing (MMI)和 PubMed Related Citation （PRC)。MMI 使用 MetaMap 将标题和摘要中的出现的关键词映射到 MeSH(Aronson and Lang,2004)。PRC 借助是美国国家医学图书馆（NLM）所属的国家生物技术信息中心（National Center for Biotechnology Information,NCBI）开发的 PubMed Related Articles(PRA)方法（一种 KNN 算法）获取相似文档的 MeSH(Lin and Wilbur,2007)。MTI 进一步把这两种方法获取的 MeSH 打分进行线性组合，然后经过一系列处理得到最后的预测结果。从这里我们可以看出 MTI 主要使用了 KNN 和模式匹配，并未考虑基于整个 MEDLINE 文档集学习各个 MeSH 的全局分类器。

在过去两年的欧洲信息检索评测平台(CLEF2013、CLEF2014)会议上主办的大规模生物医学文献语义索引竞赛中(BioASQ2013（Partalas et al, 2013)、BioASQ2014（Balikas et al, 2014)),为参赛者提供了一些新算法，除了上述中国团队的成果外，国外参赛者的主要方法有：

1.KNN （最近邻算法）

欧洲生物信息研究所的 Trieschnigg 等人比较了多种方法，发现通过查找与目标文档最相似的 MEDLINE 文献（如前 10 个最相似的文献），将这些相似文献的 MeSH 进行加权打分的方法简单有效（Trieschnigg et al, 2009)。它的主要缺点是只考虑了和目标文档相似的少量文献，而没有考虑基于整个 MEDLINE 文档集学习各个医学主题词的全局分类器。

2.MetaLabeler

希腊亚里士多德大学的 Grigorios Tsoumakas 博士等人参加 BioASQ2013 比赛使用 MetaLabeler 算法最终取得第一名(Tsoumakas et al, 2013)。MetaLabeler 算法将整个预测任务分成三步。首先，对每个标签（MeSH)，都独立训练一个分类器（例如使用 SVM)。给定一个目标文档，运行所有分类器，然后基于预测分数将所有标签(MeSH)排序。其次，训练一个分类器预测每个目标文档的标签(MeSH)个数。最后，对于一个目标文档，根据预测的标签（MeSH）个数 K，将预测分数最高的前 K 个标签（MeSH）作为结果返回。从这里我们可以看到 MetaLabeler 主要从每个 MeSH 的全局分类器出发，并没有考虑每个目标文档的局部信息（如 KNN)。

1.2.3 当前国内外研究存在的主要问题

虽然现有的研究成果提出了多种医学主题词自动标注算法，这些算法在文献的特征表示、预测方法和信息融合上有一些共同的缺陷，导致预测性能不高。具体表现在以下几点上：

第一、文献的特征表示方法单一。目前所有方法都采用传统的词袋（bag of words）模型，不考虑单词之间顺序和语义联系，因而忽略了文档中蕴涵的重要信息。除此之外，这些方法一般将标题和摘要合并在一起处理，没有考虑两者的不同角色。

第二、预测方法没有充分考虑多标签问题的特点。MetaLabeler 算法对所有 MeSH 分别独立训练分类器，这些分类器的预测分数严格来说不能直接比较，另外它也没有考虑这些 MeSH 之间的联系。常用的 KNN 算法直接使用 MeSH 加权打分的启发式方法，也没有考虑多标签问题的特点。

第三、融合的信息和方法较为单一，同时缺乏有效的融合算法。一方面，目前这些算法只融合了少量的信息和预测方法；另一方面，如何有效融合不同类型的信息和预测方法仍然是一个有待解决的难题。

1.3 研究内容

1.3.1 研究要解决的问题

许多研究已经解决了 MeSH 自动索引的挑战。现在的问题是，每个 MeSH 可以看作是一个标签(Label)，每篇文献被索引了多个 MeSH，因此，生物医学文献 MeSH 的自动索引实际上是一个大规模多标签的分类问题(Zhang and Zhou, 2014)。解决这个问题存在三个方面的难题：

第一，MeSH 的数目庞大，而且各个 MeSH 出现频率差异巨大。在 27,000 多个 MeSH 中，最频繁的主题词——Humans,出现超过 8 百万次；而排名 25000 的 Pandanaceae 只出现了 31 次。这就意味着为大部分 MeSH 训练分类器面临严重的正负样本不平衡的问题。除此之外，庞大数目的 MeSH 也为设计新算法要同时考虑两个或多个 MeSH 的相关性带来挑战。

表 1 显示六个主题词，它们的频次在所有 MEDLINE 的 12,504,999 条文档和摘要条目中排名为第 1、第 100、第 1000、第 10000、第 20000 和第 25000。这是我们通过实验得到的结果。

表 1　MeSH 中出现频率排名为第 1、第 100、第 1000、第 10000、第 20000 和 25000 的主题词的统计结果

Rank	Counts	MeSH(ID)
1	8,152,852	Humans(6801)
100	129,816	Risk Assessment(18570)
1,000	23,178	Soil(12987)
10,000	1532	Transplantation Tolerance(23001)
20,000	199	Hypnosis, Anesthetic(6991)
25,000	31	Pandanaceae(31673)

第二，每篇文挡被标注的 MeSH 数目变化巨大。有的文献有多达 30 个 MeSH，而有的文档只标注了 5 个左右的 MeSH。

第三，通常全文不可自动在 MeSH 中索引，或重要的 MeSH 概念可能存在于全文中。

有效集成不同方法和信息是提高预测性能的基本思路，在各种机器学习任务中广泛采用。在生物医学文献自动标注问题中，不同特征表示和不同预测算法基于不同原理设计开发，因此能够互为补充。除此之外，文献的发表时间和期刊等信息对于标注的 MeSH 也有直接的影响。例如有的期刊集中发表某一领域的论文，因而某些 MeSH 的出现频率就特别高。MeSH 每年都会有少量更新，这也会导致某些 MeSH 的出现和时间段紧密相关。如何设计有效策略融合不同特征表示、不同预测算法和各种信息，提高预测精度是本研究需要解决的第三个关键问题。

1.3.2 技术路线

基于排序学习融合文档的不同表示、预测算法和各种信息对于每个目标文档，我们可以先获取一些候选 MeSH，生物医学文献的自动标注问题就可以看做是一个对候选 MeSH 的排序问题。本研究拟进一步基于排序学习融合不同文档特征表示和预测算法，以及各种有用的信息（如目标文档发表时间、期刊等），从而提高生物医学文献 MeSH 标注精度。和目前基于排序学习的 L2R 方法不同，本研究将融合更多基于各种文档表示的高性能预测算法，同时：（1）设计生成更完备更

精确候选 MeSH 的算法；（2）为候选 MeSH 设计更有区分能力的特征；（3）设计更精确的预测目标文档 MeSH 个数的算法。如后文图 2 所示，本项目拟采用在很大信息检索任务中取得优秀排序性能的 LambdaMart 作为排序模型。

（1）生成候选 MeSH

基于不同预测方法如 KNN，MetaLabeler，LMEL 等获得打分高的 MeSH，我们将设计算法生成候选 MeSH，使得在 Top K（例如 K=60）中含有尽可能多正确的 MeSH。拟考虑的算法包括投票、加权打分、更复杂的学习模型等。

（2）MeSH 特征编码

本研究拟设计多种有较强区分能力的特征来表示每个候选 MeSH。这些特征包括：

1）基于特定文档表示的不同预测方法的预测分数

基于不同文档表示的不同预测方法因为设计原理各有千秋，所以在不同文献的不同 MeSH 预测上表现有高有低。每个方法的预测分数都是从一个特定角度刻画候选 MeSH 的重要性。

2）经过标准化后的预测分数

因为设计原理不同，不同方法的预测分数有时并不能直接比较。例如 MetaLabeler 算法为每个 MeSH 独立训练分类器，每个 MeSH 对应分类器的预测分数严格来说并不能直接比较。本研究拟考虑多种方法对预测分数进行标准化。这些方法包括（1）用 0,1 代表预测结果；（2）用排序代表结果；（3）将其转化成可比较分数。和 MetaLabeler 中预测分数标准化策略一致，本研究拟考虑借助大规模 MEDLINE 文献作为测试集，统计预测分数的分布，从而计算以预测分数作为阈值时对应的准确率（Precision）作为标准化后的分数。

3）该 MeSH 或者其同义词是否出现在标题、摘要或者文档中。

该特征反映模式匹配的结果，分别对应三个二进制值。

4）该 MeSH 与其他高预测分数主题词之间的相关性。

基于整个 MEDLINE，统计 MeSH 之间的共现关系，计算两两之间的条件概率。对于目标 MeSH，通过打分最高的前 K 个 MeSH 来进行加权投票，计算公式见后文公式（2）。如果一个候选 MeSH 和许多预测分数高的 MeSH 相关性很高，它是真实标注的可能性也非常大。

5）该 MeSH 在本期刊的出现概率。

（3）MeSH 个数预测

本研究拟使用支持向量回归模型（SVR）预测目标文档 MeSH 个数。拟使用的特征包括：

1）基于 KNN 的前 K（比如 K=20）个相似文档的 MeSH 平均个数。

2）同一期刊同一年内发表论文的 MeSH 平均个数

3）同一期刊和目标文档发表时间最近论文的 MeSH 平均个数。

1.3.3 实验手段

整个 MEDLINE 生物医学文献数据集可以直接从美国国立生物技术信息中心网站下载。除此之外，BioASQ 组织者也提供了 BioASQ2013 和 BioASQ2014 两次比赛的标准测试集。本研究拟采用这些数据来评估各种预测算法的性能。本项目拟使用的评估标准为信息检索和多标签学习中常用的 MiF（Micro F-Measure），MaF(Macro F-Measure) 和 EBF（Example Based F-Measure），其中 MiF 和 MaF 基于标签（Label-based）评估算法性能，而 EBF 基于每个文档(Example-based)来评价算法性能。

1.4 研究意义

本研究基于排序学习（learning to rank）融合多种文献特征表示、预测算法和各种信息，从而提高医学文献自动标注精度。给定一个测试文档，在获得候选 MeSH 之后，如何为候选 MeSH 设计有高度区分能力的特征是特征是取得优秀性能的关键。本研究收集候选 MeSH 全方位的信息，如不同预测算法的预测分数和标准化分数、模式匹配和 MTI 的结果、候选医学主题词和其他重要 MeSH 的依赖关系、候选主题词在目标期刊中的出现概率等，从而设计出有效特征，进一步提高预测性能。

MeSHLabeler 的性能优势在 2014 年 BioASQ 挑战中得到了证明。在本文中，我们进一步通过从 MEDLINE 下载的 12,504,999 篇文档和 2014 年 BioASQ 挑战赛 51,724 篇文档检查 MeSHLabeler 的卓越表现。从该系列的实验中，MeSHLabeler 取得了 Micro F-measuer 0.6248 评分，这比由 MTI 得到 0.5724 的评分高出约 9.15%。

本研究通过建立生物医学文献自动标注医学主题词的新模型，从而提高预测精度，大大降低人工标注成本，同时帮助生物医学研究人员迅速找到相关文档，缓解日益严重的信息过载问题。

第二章 MESHLABELER 的原理

2.1 概述：MeSHLabeler = MeSHRanker + MeSHNumber

问题设定如下：给定任意一个 MEDLINE 文档及其标题和摘要，要求从超过27000 个可能的 MHs 中选取一定数量的 MHs 做为该文档的 label。图 1 （a） 给出了 MeSHLabeler 的工作流程。该流程包含了 MeSHRanker 和 MeSHNumber 两个组成部分。对于每一个输入的 文档，MeSHRanker 返回一个有序的候选 MHs 列表，MeSHNumber 则预测应当从候选列表中选出的 MHs 数量。

2.2 预备知识

本文的问题是多标签分类（multi-label classification）。MetaLabeler 是一个解决该问题的简单方法，本文将该方法作为本次研究的基准(Baseline)。在涉及到 MetaLabeler 之前，我们需要回顾一个二分类(binary classification)问题：每一个 MH 是一个二项类，每一个文档是一个实例（instance），而一个文档与任意一个 MH 的相关程度，则可通过二分类来预测。在 MetaLabeler 和 MeSHLabeler 中，将使用逻辑回归（LogReg）和 K-近邻算法（KNN）两种最为广泛接受的分类方法。请注意，这两种方法从机器学习的角度来看是互补的：逻辑回归使用整个数据库来训练模型，尝试得到全局信息（global evidence），而 K-近邻算法则从相似的实例入手，尝试得到局部信息（local evidence）。接下来，本文从描述这两种方法以及 MetaLabeler 开始。

2.2.1 逻辑回归（LogReg）

Logistic regression （逻辑回归）是当前业界比较常用的机器学习方法，用于估计某种事物的可能性。比如某用户购买某商品的可能性，某病人患有某种疾病的可能性，以及某广告被用户点击的可能性等。

本文使用了一般最优化方法来估计逻辑回归(LogReg)的参数。整个 MEDLINE 记录集合被用来训练逻辑回归的参数，用以得到全局信息。

2.2.2 K-近邻（KNN）

作为KNN算法的输入，我们需要能找任意一篇指定文档的相似文档，以及他们之间的相似度。本文使用NCBI efetch (http://www.ncbi.nlm.nih.gov/books/NBK25499/) 来检索相似的文档。这些相似的文档集合则由与KNN算法很类似的PubMed Related articles (PRA) (Lin and Wilbur, 2007) 方法生成。对于给定文档 A，被检索出的类似文档的 MHs 集合，可以在预测文档 A 的 MHs 时作为非常有效的候选者。因此，在给定目标文档的情况下，我们用如下公式计算MH的得分：

$$\frac{\sum_{i=1}^{K_{NN}}(S_i \cdot B_i)}{\sum_{i=1}^{K_{NN}} S_i} \qquad (1)$$

其中，K_{NN}是与目标文档最为相似的文档数量，S_i是第 i 个文档与目标的相似性得分，B_i是一个用来表示候选MH是否在第 i 个文档中出现的二进制变量。

2.2.3 MetaLabeler

MetaLabeler 是一个直接有效的解决多标签分类问题的方法(Tang et al.，2009)，它基于一个分类器和一个回归器（classifiers)。

1. 分类器A
   本文为每一个label（在本文中即为MH）独立地训练一个二分类器（默认方法为支持向量机，但可以替换成其他方法)。给定一个实例（即本文问题中的文档)，使用所有的已训练的本类型分类器，预测 label 的得分，并据此对label排序。
2. 回归器B
   我们训练了一个用于预测实例label数量的回归器。给定实例后，该回归器将预测该实例的label数量。
3. 最终预测
   对于给定实例，我们通过分类器A得到一个有序label列表，通过分类器B得到我们需要的label的数量。之后从有序label列表中选出需要数量的得分最高的label。

另外需要注意的是，任何二元分类方法都可在分类器A中使用，而逻辑回归（LogReg)在效率方面就是一个很好的选择。此后，如果逻辑回归用于分类器A，

我们将 MetaLabeler 称之为 MLogReg。而如果用了得分标准化（将在后文描述）使得 MHs 在 MLogReg 中可比，我们则称之为 MLogRegN。

MetaLabeler 的主要流程被合并进入了 MeSHLabeler，因此本文可以作为与 MeSHLabeler 性能比较的基准（Baseline）。此外，MeSHLabeler 有很多其他重要特性，例如不同 label 间的预测得分归一化（normalization）、label 间的从属关系、排序学习（Learning to rank）。它们有效地优化了 MeSHLabeler 相对于 Baseline（MetaLabeler）的预测性能。

2.3 MeSHRanker

图 1 （b）给出了 MeSHRanker 的过程。本文将据此解释 MeSHRaner。

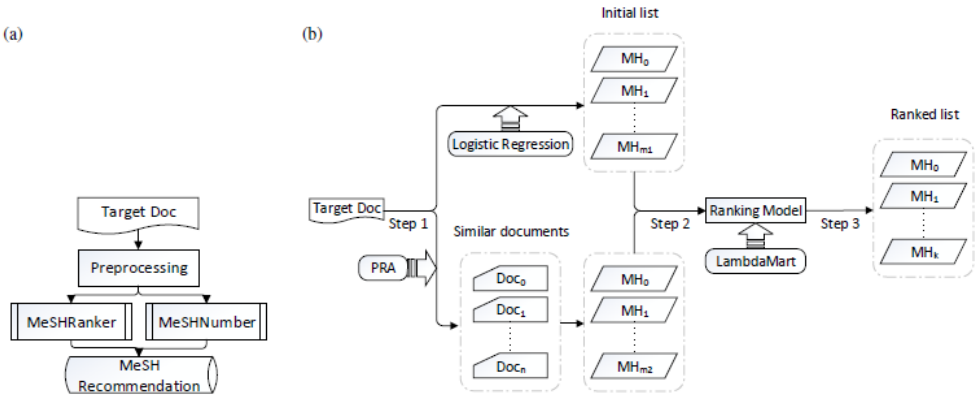


图 1 MeSHLabeler （a）和 MeshRanker（b）的工作流程

2.3.1 步骤 1：生成候选 MeSH

本文研究的问题中有超过 27000 个 MeSH，因此在这个多标签分类问题中也有非常多的类（labels）。为了在之后的步骤中减少无关 MHs 带来的额外计算负担，首先运用如下的方式为目标文档生成候选 MHs，从而使研究集中在有限的 MHs 上：通过逻辑回归预测并排序获得一个有序 MHs 列表L_{LogReg}，再通过满足公式(1)的 K-近邻方法得到另一个列表L_{KNN}，最后，我们将两个表中满足以下两个要求中至少一个的 MH 合并起来得到的候选 MHs：

- 在有序表L_{LogReg}的前N_{LogReg}个中出现
- 在有序表L_{KNN}的前N_{KNN}个中出现

2.3.2 步骤 2：为步骤三中排序 MHs 生成下列 7 个特征

1. MetaLabeler 与 LogReg（MLogReg）

原始的 MetaLabeler 使用了一个支持向量机（SVM），而这里我们除了选择使用 LogReg 外，保持其他部分和 MetaLabeler 完全一样。我们注意到 MLogReg 使用到了所有的文档,这意味着它得到的是全局信息。

事实上，对于每个 MH，本文使用了最新发布的一百万个文档来训练模型。对于不常见的 MHs，我们会使用放宽文献的时间要求，采样足够多的正样本，以避免过拟合问题。

2. K-近邻（KNN）

本文使用了公式 （1）来求 KNN。

3. 经过得分归一化的 MLogReg——MLogRegN

本文将 MLogReg 得到的初始得分使用如下方法归一化：首先，将所有文档的得分降序排列。因为每个文档只可能是正样本或负样本，对于每一个文档，我们可以将比该文档得分更高的正样本数目除以所有正样本数，得到该 Citaton 预测结果的精确率（precision）。直接地说，给定原始得分 x，得到的归一化之后的得分是：用 x 作为阈值，分类器的精确率(precision)。这意味着对于一些文档，任意得分均能被转换为范围从 0 到 1 的精确率。而对于每个 MH，通过精确率归一化后的得分代表了其预测结果正确的可能性，而且相互之间可比。我们对每一个 MH 均执行上述转换，并用得到的精确率作为所有 MHs 归一化后的得分。

4. MeSH 依赖性

对于一个候选 MH，用$\widehat{MH}$表示，其 MeSH 依赖性的得分计算方法如下：

$$\sum_{i=1}^{K_{MeSHdepend}} f_{\mathrm{MLogRegN}}(MH_i)\cdot \mathrm{P}\left(\widehat{MH} \mid MH_i\right) \qquad (2)$$

其中，MH_i是候选列表按 MLogRegN 得分排序后的第 i 个 MH，$f_{\mathrm{MLogRegN}}(MH_i)$则是 MLogRegN 对$MH_i$给出的得分。$\mathrm{P}\left(\widehat{MH} \mid MH_i\right)$是给定$MH_i$时$\widehat{MH}$的条件概率。获得$\mathrm{P}\left(\widehat{MH} \mid MH_i\right)$的方法下：

$$P\left(\widehat{MH} \mid MH_i\right)=\frac{\left|N\left(\widehat{MH},\ MH_i\right)\right|}{|N(MH_i)|}$$

其中，$N(MH_i)$是MH_i在整个 MEDLINE 中出现的次数。$N\left(\widehat{MH},\ MH_i\right)$则是在整个 MEDLINE 中$\widehat{MH}$和$MH_i$同时出现的次数。

直观地说，MeSH 依赖性特征表明：如果$\widehat{MH}$与得分很高的 MHs 有高度关联，

则说明$\widehat{MH}$更可能是一个正确的标注。

5. 模式匹配

我们可以检测目标文档的文本（标题与摘要）中是否直接包含每个 MH。该过程如下：1)对于每个 MH，能在 MeSH THEAURUS 中检索其 entry term 和同义词。2)对于一个文档，扫描它的标题和摘要，如果对应的 MeSH entry term 或同义词被找到了，则给其分配 1，否则分配 0。因此我们能够给 entry term 和同义词生成共 2 个二进制特征。这个过程的计算量很小，因为其只需对目标文档做一次字符串匹配。我们也注意到模式匹配可以有如下三种模式：只检测标题、只检测摘要、摘要和标题均检测。

6. MeSH 频率

我们能够计算一个文档所属期刊中$\widehat{MH}$的出现概率如下：

$$\frac{\left|N(\widehat{MH})\right|}{N_J}$$

其中N_J是期刊 J 中文档的数量，$N(\widehat{MH})$是$\widehat{MH}$在期刊 J 中全部文档中的出现次数。

7. MTI

我们可以使用 MTI 推荐的 MHs。MTI 整合了 KNN、模式识别以及 indexing rules 作为特征。我们用了 MTI 的两个选项:MTI default(MTIDEF)和 MTI FirstLine Index(MTIFL)。MTIDEF 试图在精确率和召回率上取得平衡，而 MTIFL 则推荐更少的 MHs，从而使得精确率更高。本文则以 MTIDEF 和 MTIFL 中是否出现生成两个二值特征。

2.3.3 步骤 3：通过 learning to rank 排序 MeSH

本文使用“learning to rank”来排序 MHs。这是一个广泛应用在信息检索中，根据查询信息，将文档排序的方法。在 MeSHRanker 中，若转化对应到为原始搜索引擎中的 LTR 问题，相当于将 MHs 作为文档，而等待预测的文档作为询问。这意味着候选 MHs 将按照与需要标注的文档的相关程度排序。Learning to rank 有很多实现方法，本文使用了 Lambda MART(Burges, 2010)，该方法已经在若干实际问题中成功地应用。我们重申，我们的想法是将多个、尽可能独立的、不同的分类器整合到“learning to rank”框架中来。

2.4 MeSHNumber

MeSHNumber 预测应当从 MeSHRanker 输出的 MHs 列表中取得分最高 MHs 的个数。这一部分的关键点在于，使用多种、不同且多样化的特征来达到对每个文档的 MHs 个数的高可预测性。我们首先生成如下六种不同类型的特征，并使用支持向量回归（SVR）来预测。

1. 同一期刊中的文档
   我们首先查询与目标文档发表在同年份且同期刊的文档及其 MHs 数量。接下来计算 MHs 数量的平均值和标准差，作为特征使用。类似的，我们找出在同期刊中发表时间距离目标文档最近的 5 个文档以及 MHs 数量。它们 MHs 的均值和标准差也被用作特征。
2. PubMed 相关文章
   使用 PRA 计算出的与目标文档最相似的文档集合，计算其 MHs 的均值和标准差作为特征。
3. MeshRanker 的原始分类器输出
   选择在使用逻辑回归预测目标文档的 MHs 时得到的若干最高得分，直接作为特征。这些得分由 MeSHRanker 的第一步得到。
4. Learning to rank 的输出
   选择 learning to rank 预测目标文档的 MHs 时得到的若干最高得分，直接做为特征。这些得分可由 MeSHRankder 的第三步得到。
5. MetaLabeler 的输出
   训练 MetaLabeler 来预测目标文档的 MHs 数量（使用支持向量回归选项），并使用预测结果作为特征。
6. MTI 的输出
   我们将 MTIDEF 和 MTIFL 预测目标文档的 MHs 数量结果作为两个特征加入。

第三章 实验

3.1 数据

我们从 NLM 中下载了 22,376,811 个 2014 年 BioASQ 挑战赛开始之前的的 MEDLINE/ PubMed 文档，过滤掉了没有摘要的文档后，得到 12,504,999 个过滤后文档，并将这些文档存储在本地作为一个训练集。通过 BioTokenizer (Jiang and Zhai,2007)符号化这些文档，形成一个含有 3,712,632 个 tokens 的字典。在这个过程中，本文使用单字组（unigram）和两字组（bigram）特征代表每个文档，并且在整个数据中，只考虑那些出现六个或更多次的特征(Tsoumakas et al.,2013)。这是因为不常见的单字/两字的特征含有的信息不够丰富，保留它们会使计算代价变得很高。我们获得了 111,034 单字组和 1,867,013 两字组的特征，这样每个文档由 1,978,047 元素的非常稀疏向量表示。此外，我们使用了 TF-IDF 得分方案，以确定每个单字/两字组特征的权重。这些特征将用作逻辑回归。

然后，再从 BioASQ 挑战赛的测试数据集下载了 51,724 个文档，随机选择 32,684 个文档进行 MeSHRanker 的步骤 3 训练，10,000 个文档进行 MeSHNumber 训练，以及 9,040 个文档进行 MeSHLabeler 性能检验。这 51,724 个文档，每个文档平均有 10.3 个句子和 162.6 个单词。为了得到公平的比较结果，文中提到的所有方法均在这 9040 个文档上进行了测试。

3.2 实现

我们使用一个开源工具——RankLib[2]，来实现 Lambda MART (Burges, 2010)。逻辑回归和 SVM(Fan et al., 2008)均通过 LIBLINEAR 来实现。 SVR (Chang and Lin, 2011)则通过 LIBSVM 实现。

3.3 性能评价测量

我们使用三种不同的衡量标准，它们全部基于 F-measure，这个测量方法常用于信息检索。F-measure 用来计算使用精确率和召回率，对于每一种 F-measure,

[2] 7 http://sourceforge.net/p/lemur/wiki/RankLib/

我们也附上相应的 Precision 和 Recall，所以总计会有 9 个指标。

3.3.1 标注

分别设 K 表示所有 MeSH 的数量，N 为实例的数量。使得 y_i 和 $\hat{y_i} \in \{0,1\}^K$ 为真并分别为每一个文档 i 的真实 label 和预测 label。

3.3.2 三种类型的 F-measure

- F-measure：EBF

EBF 作为可被计算的标准精确率（EBP）和召回率（EBR）的调和平均值，是标准的 F-measure。其计算方法如下：

$$\text{EBF} = \frac{1}{N}\sum_{i=1}^{N} EBF_i \qquad (3)$$

其中

$$EBF_i = \frac{2 \cdot EBP_{\ i} \cdot EBR_{\ i}}{EBP_i \cdot EBR_i}$$

其中

$$EBP_i = \frac{\sum_{k=1}^{K} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \cdot \hat{y}_i^k} \qquad EBR_i = \frac{\sum_{k=1}^{K} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \cdot \hat{y}_i^k}$$

我们注意到，对所有的实例，都可以分别对EBP_i和EBR_i求和，计算出 EBP 和 EBR。

- Macro F-measure：MaF

MaF 是 Macro 平均调和平均精确率（MaP）和 Macro 平均召回率（MaR）计算方法如下：

$$\text{MaF} = \frac{2 \cdot MaP \cdot MaR}{MaP + MaR} \qquad (4)$$

先分别计算出每个标签（MH）的精度，然后将所有标签的精度求平均，就得到了 Macro 平均精度和召回率，如计算方法如下：

$$\text{MaP} = \frac{1}{K}\sum_{k=1}^{K} R^k \qquad MaR = \frac{1}{K}\sum_{k=1}^{K} R^k$$

其中

$$P^k = \frac{\sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{i=1}^{N} \hat{y}_i^k} \qquad R^k = \frac{\sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{i=1}^{N} \hat{y}_i^k}$$

- Micro F-measure：MiF

MiF 是 Micro 平均精度（MiP）和 Micro 平均召回率（MiR）的调和平均值，具体计算如下：

$$\mathrm{MiF} = \frac{2 \cdot \mathrm{MiP} \cdot \mathrm{MiR}}{\mathrm{Mip} + MiR}, \qquad (5)$$

其中

$$\mathrm{MiP} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} \hat{y}_i^k} \qquad \mathrm{MiR} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} \hat{y}_i^k}$$

根据这些定义，我们可以看到，Micro F-measure 受频繁使用标签的影响更大，而 Macro F- measure 对待所有的标签（包括不常见的）反应一样。正因为如此，BioASQ 挑战赛的系统运用 MicroF-measure 进行评估，MiF 也是我们系统中评估的重点。

3.4 参数设置

在 MeSHRanker 的第 1 步中，对于N_{LogReg}和N_{KNN}，我们需要选择合适的参数来尽量减少计算负担和噪声，以获得足够好的性能。在初步实验时，我们分别为N_{LogReg}和N_{KNN}选择 40 和 50。实际上，初步实验表明，如果这些参数被设置为 30 或更多时，性能几乎能达到饱和。

对于N_{KNN}，我们使用 25。在初步实验中，我们发现 25 的数量已经足以找到与目标文档的类似的文档。 $K_{MeSHDepend}$设定为 80，这对找到重要的（即使不常见的）MHs 也足够大。

为了使 MeSHNumber 在总数为 229 个特征(4+2+200+20+1+2)中捕捉到足够的信息，我们分别设置M_{PRA}、M_{LogReg}、M_{LTR}为 10,200,20。

3.5 性能结果

我们首先通过 MetaLabeler 检查得分标准化的影响，并与现有的索引方法如 KNN、模式匹配和 MTI 进行比较。其次，在考查 MeSHRanker 的性能时，通过加入第 2.3.2 节所示的各个特征增量，检验整合不同类型特征的影响。在第一步和第

二步实验中，用 MetaLabeler 预测了 MHs 的数量。最后，我们探讨了通过组合 MeSHNumber 与 MeSHRanker 后的 MeSHLabeler 性能。

3.5.1 得分标准化的效果

MHs 出现的次数变化很大，从而导致分类器的预测性能差别大。图 2 显示了由逻辑回归得到的用于预测 4 个 MHs 的 4 个精确率-召回率曲线：Humans（最常见的 MH）、cell survival、prosthesis failure、follicular fluid。

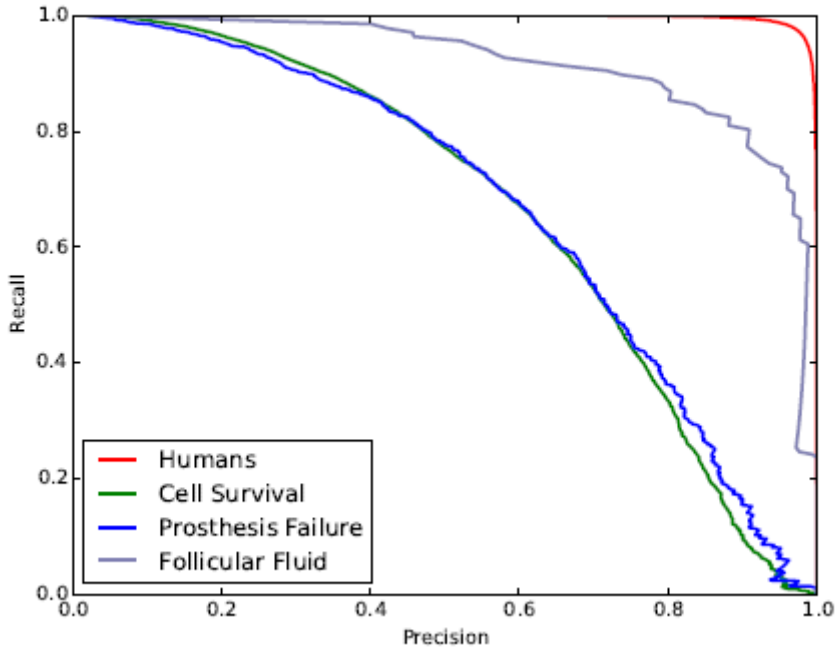


图 2：逻辑回归得到四个 MeSH(Humans, Cell Survival, Prosthesis Failure, Follicular Fluid)的精确率/召回率曲线

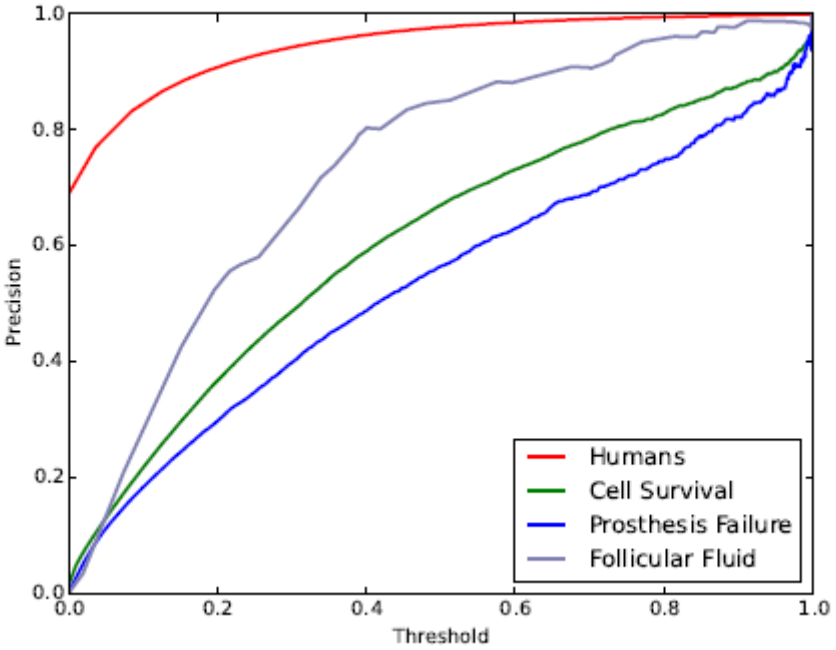


图 3：图 2 中四个 MeSH 在不同阈值（threshold）下逻辑回归的精确率

图 2 显示，当我们比较精度—召回率曲线（AUPR）下方的区域，最频繁使用的 MHs——Humans，显然达到了 AUPR 最高。同样的 4 个 MHs，图 3 显示出通过改变初始预测得分的阈值所得到的精确率。这个数值也明确显示了 4 个 MHs 之间的偏差，而且，对于相同的初得分，精确率的差距非常大。例如，分类为“Humans”的 MHs，在初始得分为 0.6 时，获得了大于 0.9 高精确率，而分类 prosthesis failure 的 MHs 在同一得分时。只有约 0.6 的精确率。这一结果意味着如果我们直接使用初始的预测得分，不常见的 MHs 比频繁使用的 MHs 更有可能被选中。得分标准化更侧重于频繁出现的 MHs，这意味着 MiF（及 MiP 和 MiR）等等标准化将有效的改善 Micro F-measure。事实上，这也是 BioASQ 挑战的主要评价指标。

我们通过 MetaLabeler 考查运用得分标准化的效果，也就是说将 MLogRegN 的性能与其它现有的典型方法相比较。表 2 给出了 MLogRegN（和 MLogReg）以及那些现有的几种方法，包括模式匹配，KNN 和 MTI 的性能比较。该表显示，在 MiF 这个项目上，MLogRegN 达到的 0.5754 的最高值，之后是 MTIDEF（0.5724），MTIFL（0.5624），MLogReg（0.5595），KNN（0.5213）和模式匹配。特别是，在所有 MiP、MiR 和 MiF 项上，MLogRegN 优于 MLogReg，而这在所有 MaP、MaR 和 MaF 项上则是反向的，这就验证了我们的预期。例如，MLogRegN 在 MiF 上得分达到 0.5754，在 EBF 上得分达到 0.5628，而 MLogReg 相应的在 MiF 上的得分为 0.5595，在 EBF 上得分是 0.5502。另一方面，MLogReg 在 MaF 方面的得分达到 0.4612，而

MLogRegN 在 MaF 上只获得了 0.4335。有趣的是，MTIDEF 和 MTIFL 在 MaF 上得到了最高分，分别达到了 0.5247 和 0.5038，这意味着它们可能对不常见的 MHs 能够有效好的作用。我们还可以看到，MTIFL 相当专注于改善精确率，同时，通过精确率和召回率之间的平衡，实现 MTIDEF 比 MTIFL 能获得更好的 F-measure。例如，MTIFL 的 EBP 得分最高，达到了 0.6192；而 MTIDEF 也实现了 EBF 0.5645 的最高得分。在本实验所有的测试方法中，KNN 体现出的性能比较平均，如 MiF 为 0.5213，EBF 为 0.5095 和 MaF 为 0.3927。三种模式匹配方法中，只使用标题获得的精确率最高，只使用摘要比仅使用标题实现的召回率更高，于是导致同时使用标题和摘要所实现的三种类型的 F-measure 的值最高。

总体而言，MLogRegN 取得了最高的 MiF，这表明并入了得分标准化的有效性。另一从这个结果重要的发现是，所有这些方法都显示出其独特的优点，这取决于不同类型的评价，表明这些方法是可以互补的，这一分析为后面将这些方法的思路或特征整合在一起提供了良好的基础和理由。

表 2：MLogRegN 与现有的典型方法的性能比较

Method	MiP	MiR	MiF	EBP	EBR	EBF	MaP	MaR	MaF
MLogReg:MetaLabeler with LogReg	0.5576	0.5614	0.5595	0.5555	0.5772	0.5502	0.4600	0.4623	0.4612
MLogRegN:MLogReg with score normalization	0.5734	**0.5774**	**0.5754**	0.5702	0.5884	0.5628	0.4508	0.4175	0.4335
KNN	0.5196	0.5231	0.5213	0.5176	0.5314	0.5095	0.4142	0.3733	0.3927
Pattern matching using titles only	0.5151	0.1273	0.2041	0.5112	0.1426	0.2101	0.3444	0.1997	0.2528
Pattern matching using abstracts only	0.2315	0.2990	0.2609	0.2445	0.3117	0.2582	0.3607	0.3956	0.3773
Pattern matching using both titles and abstracts	0.2363	0.3139	0.2696	0.2498	0.3291	0.2681	0.3739	0.4153	0.3935
MTIFL	**0.6142**	0.5217	0.5642	**0.6192**	0.5386	0.5549	**0.5159**	0.4923	0.5038
MTIDEF	0.5740	0.5707	0.5724	0.5785	0.5909	**0.5645**	0.5128	**0.5372**	**0.5247**

3.5.2 MeSHRanker 的性能

我们通过连续添加不同类型的证据（特征），一步一步地检验 MeSHRanker 的性能，并以表 2 中的实验取得最佳性能的 MLogRegN 为基准进行比较。在这里，所有比较方法采用的 MHs 的数量都来自于 MetaLabeler（回归器 B）的预测。我们从只用两种类型的特征开始，即 MLogReg 和 KNN，来检查 MeSHRanker 的性能时。然后再按照如下顺序加入其它类型的特征到已经包含 MLogReg 和 KNN 特征的 MeSHRanker 中测试性能：MLogRegN，MeSH dependency，pattern matching，MeSH

frequency 和最后的 MTI。表 3 为 MLogRegN 和使用不同特征的 MeSHRanker 的性能测试结果。该表显示，加入了 MLogReg 和 KNN 的 MeSHRanker 得到了 0.5743 的 MiF 分值，比基准的 0.5754 的 MiF 略低。将 MLogRegN 并入到特征中，MeSHRanker 的性能得到极大地改善，MiF 值为 0.5899，比基准要优秀了，与 MLogRegN 相比已经有了较大的提高。加入 MeSH dependency 后，性能提升影响也很显著， MiF 从 0.5899 提升到 0.5957，EBF 从 0.5802 提升到 0.5861，MaF 从 0.4602 提升到 0.4938。MaF 大幅度改善还表明纳入 MeSH dependency 有可能协助发现不常见的 MHs，其中必定蕴含有常见的 MHs。加入 pattern matching 的特征也很有益，这意味着模式匹配可能带来其他特征的补充信息。特别是，MiF 从 0.5957 改善到 0.6056，EBF 从 0.5861 改善到 0.5955，MaF 从 0.4938 改善到 0.5205，也揭示了模式匹配发现不常见 MHs 的实力。有趣的是，通过增加 MeSH frequency 后性能的变化幅度非常小。这可能是因为 MeSH frequency 提供的信息已经由其它类型的特征获得，如 KNN 和 MeSH dependencies。最后，在 MeSHRanker 中加入 MTI 的特征，性能得到了很大的提高，在所有的措施中导致了最高性能，例如，MiF 达 0.6166，EBF 达 0.6082 和 MaF 达 0.5389。总体而言，

Step	MiP	MiR	MiF	EBP	EBR	EBF	MaP	MaR	MaF
MLogRegN	0.5734	0.5774	0.5754	0.5702	0.5884	0.5628	0.4508	0.4175	0.4335
MeSHRanker (MLogReg+KNN)	0.5724	0.5763	0.5743	0.5708	0.5900	0.5637	0.4597	0.4396	0.4496
+MLogRegN	0.5878	0.5919	0.5899	0.5878	0.6072	0.5802	0.4741	0.4471	0.4602
+MeSH dependency	0.5937	0.5978	0.5957	0.5935	0.6134	0.5861	0.4889	0.4988	0.4938
+Pattern Matching	0.6036	0.6077	0.6056	0.6043	0.6246	0.5955	0.5162	0.5248	0.5205
+MeSH frequency	0.6038	0.6079	0.6059	0.6043	0.6246	0.5956	0.5166	0.5205	0.5185
+MTI	**0.6145**	**0.6187**	**0.6166**	**0.6159**	**0.6363**	**0.6082**	**0.5364**	**0.5414**	**0.5389**

表 3：MLogRegN 与不同类型的证据逐步增加的 MeSHRanker 的性能比较

我们可以看到整合多个不同类型的证据，使 MeSHRanker 的性能极大地提高。

3.5.3 MeSHLabeler 的性能

在上个实验中，运用了 MetaLabeler 对目标文档中的 MHs 数量进行了预测。相对应地，在这个实验中，我们运用 MeSHNumber 为 MHs 的排名列表，这一排名由包含所有特征的 MeSHRanker 来预测，故此我们称之为组合 MeSHLabeler。此外，我们将上个实验的最终结果命名为 MeSHRanker（显示在表 3 的最后一行）。

表 4 显示的是 MeSHLabeler 和 MeSHRanker，与 MTIDEF 的性能比较。我们注意到 MTIDEF 是由 NLM 提供的当前最新索引工具，我们的数据已经对其性能进行了评价并显示于表 2 的最后一行。在表 4 的所有三种类型指标中，MeSHLabeler 取得了较好的精度，同时 MeSHRanker 或得了较好的召回率。唯一的区别是用于 MeSHRanker 和 MeSHLabeler 预测的 MHs 数量。因此，更高的精确率意味着 MeSHLabeler 要返回较小数量的 MHs 以达到较高的精确率。在 F-measure 方面，MeSHLabeler 在 MiF 和 EBF 两项上都实现更高的性能，而 MeSHRanker 在 MaF 上取得更高的性能，这意味着不常见的 MHs 可能被 MeSHLabeler 忽略，这导致较低的 MaF。

最重要的是，由该结果带来两个基本发现：1）MeSHRanker 优于 MTIDEF 所有九个评价方法，和 2）在 MiF 上，与 MTIDEF 的 0.5724 相比，MeSHLabeler 达到 0.6248，这是一个 9.15%左右的改进（近 10%）。

表 4. MeSHRanker 和 MeSHLabeler 与由 NLM 提供的当前处于前沿的索引工具 MTIDEF 的性能比较

Step	MiP	MiR	MiF	EBP	EBR	EBF	MaP	MaR	MaF
MTIDEF	0.5740	0.5707	0.5724	0.5785	0.5909	0.5645	0.5128	0.5372	0.5247
MeSHRanker	0.6145	**0.6187**	0.6166	0.6159	**0.6363**	0.6082	0.5364	**0.5413**	**0.5389**
MeSHLabeler	**0.6566**	0.5959	**0.6248**	**0.6618**	0.6108	**0.6160**	**0.5450**	0.5172	0.5054

3.6 计算效率

我们仅用一台有四个 2.7GHz 的 Intel XeonE5-4650 CPU 和 128GB 的内存的服务器运行 MeSHLabeler。大多数的计算时间是花在对 27,000 多个 MHs 进行逻辑回归分类的训练上，这大约花了五天。所有其它部分的训练花了一天。然而，给出一个新文档，标注 MeSH 仅需要不到一秒钟，因此 MeSHLabeler 的标注效率之高是显而易见的。

第四章 讨论

MeSHLabeler 的基本思路是整合多种类型的证据提高 MeSH 索引的性能。在 MiF 和 EBF 方面，MeSHLabeler 非常有效，相对于由 NLM 提供的当前最尖端的索引工具——MTI，显示出近 10％改善。MeSHLabeler 的高性能来源于多样性和准确性证据的相互补充。 MeSHLabeler 用了五种类型的证据：全局证据，本地证据，模式匹配，主题词依赖性和索引规则（源自 MTI）。前两种类型的证据是通过机器学习获得的，这意味着它们可以从训练数据获得。全局证据是对所有实例进行预测模型训练，而本地证据只使用相似的实例给出用于测试的实例。模式匹配是一个字符串匹配技术，仅使用测试实例。与此相反，在最后两个类型的证据是与其他类型证据非常不同的。MeSH 依赖性从 MeSH 之间的相关性得到，这要求扫描整个 MEDLINE 数据库。不同的 MHs 组合的数据是巨大的，这使得直接将不同的 MeSH 信息组合很艰难，而且现成的方法都没有考虑 MeSH 依赖性。我们捕获 MeSH 依赖性策略是非常有效的，从而产生如我们的实验所显示出的高的性能改进。MTI 索引规则与其他的证据也有很大的不同，因为它们包含专家的知识，它们中的一部分要从数据学习是非常困难的。MeSHLabeler 集成了所有这些不同类型的多样证据，使 MeSHLabeler 能够实现高性能地标注 MeSH。

我们发现，三组评价指标（即 EBF，MaF 与 MiF）中，MiF 和 EBF 可能更多专注于常见的 MHs，而 MaF 更多地是平等地考虑所有 MHs。 MeSHLabeler 是以期改善 MiF 的性能的目的开发的，因为 MiF 是 BioASQ 挑战一个主要指标。例如，不同的 MeSH 分类器预测分数具有不同的精度（即使它们的值是一样的，如我们的实验中所述）。其它方法，如 MetaLabeler，必定会出现选择不常见的 MHs（因为高预测分数）多于的常见的 MHs 情况。为了克服这个问题，我们产生了在 MeSHLabeler 中将分数标准化的想法，结果使得 MiF 有大的改进。同样，MiF 和 MaF 之间的权衡也是一个值得讨论的话题。例如，MeSHNumber 主要是对少数的 MeSH 具有高精确率，且在 MiF 上取得了显著增加，但是在 MaF 上略有下降。

第五章 结论

本文介绍的 MeSHLabeler，在 2014 年 BioASQ 组织的大规模生物医药语义索引挑战赛中，获得了最佳的表现。我们的实验已经显示了 MeSHLabeler 在 MiF 项上，超过了由 NLM 提供的当前领先的解决方案，同时实现超过 MTI 9%的 MiF 增长。MeSHLabeler 最重要的思想是整合五种不同类型的证据：全局特征，局部特征，模式匹配， MeSH 依赖性和索引规则。此外，MeSHLabeler 有许多其他方法没有特性，如 MeSH 得分标准化，和考虑 MeSH 依赖性。这些特性极大地促进了 MeSHLabeler 获得更好的性能。这些新特性可能会揭示开发大量训练实例的多标签分类问题的高效算法，例如 ontology annotation。在将来，进一步探索 MeSHLabeler 的局限性、尝试组合其他不同类型的信息和证据提高索引的性能，试图使用不同的方法来表示文档本身，突破词袋方法的限制，将是十分值得考虑的。

参考文献

[1]Aronson, A. and Lang, F. (2004). An overview of MetaMap: historical perspective andrecent advances. J Am Med Infor Assoc, 17(3), 229–236.

[2]Aronson, A., Mork, J., Gay, C., Humphrey, S., and Rogers, W. (2004). The NLM indexing initiative's medical text indexer. Stud Health Technol Inform, 107(Pt 1),268–272.

[3]Balikas, G., Partalas, I., Ngomo, A., Krithara, A., and Paliouras, G. (2014). Results of the BioASQ track of the question answering lab at CLEF 2014. CLEF (Working Notes), pages 1181–1193.

[4]Burges,C.J.C.(2010).From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research.

[5]Chang, C. C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2, 27:1–27:27.

[6]Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear:A library for large linear classification. The Journal of Machine Learning Research,9, 1871–1874.

[7]Gu, J.,Feng, W., Zeng, J., Mamitsuka, H., and Zhu, S. (2013). Efficient semisupervised MEDLINE document clustering with MeSH semantic and global content constraints. IEEE Transactions on Cybernetics, 43 (4), 1265–1276.

[8]Huang, M., N ́ev ́eol, A., and Lu, Z. (2011a). Recommending mesh terms for annotating biomedical articles. Journal of the American Medical Informatics Association, 18(5), 660–667.

[9]Huang, X., Zheng, X., Yuan, W., Wang, F., and Zhu, S. (2011b). Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. Information Science, 181(11), 2293–2302.

[10]Jiang, J. and Zhai, C. (2007). An empirical study of tokenization strategies for

biomedical information retrieval. Information Retrieval, 10(4-5), 341–363.

[11]Jimeno-Yepes, A., Mork, J. G., Wilkowski, B., Demner Fushman, D., and Aronson, A. R. (2012a). MEDLINE MeSH indexing: lessons learned from machine learning and future directions. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pages 737–742. ACM.

[12]Jimeno-Yepes, A., Mork, J. G., Demner-Fushman, D., and Aronson, A. R. (2012b). A one-size-fits-all indexing method does not exist: Automatic selection based on meta-learning. JCSE, 6(2), 151–160.

[13]Jimeno-Yepes, A., Mork, J. G., Demner-Fushman, D., and Aronson, A. R. (2013a). Comparison and combination of several mesh indexing approaches. In AMIA Annual Symposium Proceedings, volume 2013, page 709. American Medical Informatics Association.

[14]Jimeno-Yepes, A. J., Plaza, L., Mork, J. G., Aronson, A. R., and D´ıaz, A. (2013b). MeSH indexing based on automatically generated summaries. BMC bioinformatics,14(1), 208.

[15]Lin, J. and Wilbur, W. (2007). Pubmed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics, 8(1), 423.

[16]Liu, T. (2011). Learning to Rank for Information Retrieval. Springer.

[17]Lu, Z., Kim, W., andWilbur, W. (2010). Evaluation of query expansion using MeSH in PubMed. Information Retrieval, 12(1), 69–80.

[18]Mao, Y. and Lu, Z. (2013). NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. Technical report.

[19]Mao, Y., Wei, C., and Lu, Z. (2014). NCBI at the 2014 BioASQ challenge task: Large-scale biomedical semantic indexing and question answering. CLEF (Working Notes), pages 1319–1327.

[20]Mork, J., Jimeno-Yepes, A., and Aronson, A. (2013). The NLM medical text indexer system for indexing biomedical literature. BioASQ@ CLEF.

[21]Mork, J., Demner-Fushman, D., Schmidt, S., and Aronson, A. (2014). CLEF

(Working Notes), pages 1328–1336.

[22]NCBI Resource Coordinators (2015). Database resources of the national center for biotechnology information. Nucleic Acids Res., 43(Database issue), D6–D17.

[23]]Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J.-L., and Arluk, N. (2004). The MeSH translation maintenance system: structure, interface design, andimplementation. Medinfo, 11(Pt 1), 67–69.

[24]Partalas, I., Gaussier, ´E., Ngomo, A.C. N., et al. (2013). Results of the first BioASQ workshop. In BioASQ@ CLEF, pages 1–8.

[25]Ruch, P. (2006). Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics, 22(6), 658–664.

[26]Stokes, N., Li, Y., Cavedon, L., and Zobel, J. (2010). Exploring criteria for successful query expansion in the genomic domain. Information Retrieval, 12(1), 17–50.

[27]Tang, L., Rajan, S., and Narayanan, V. K. (2009). Large scale multi-label classification via metalabeler. In Proceedings of the 18th international conference on World wide web, pages 211–220. ACM.

[28]Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., and Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics, 25(11), 1412–1418.

[29]Tsoumakas, G., Laliotis, M., Markantonatos, N., and Vlahavas, I. P. (2013). Large-scale semantic indexing of biomedical publications. In BioASQ@ CLEF.

[30]Zhang, M. and Zhou, Z. (2014). A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819–1837.

[31]Zhu, S., Zeng, J., and Mamitsuka, H. (2009a). Enhancing MEDLINE document clustering by incorporating mesh semantic similarity. Bioinformatics, 25(15),1944–1951.

[32]Zhu, S., Takigawa, I., Zeng, J., and Mamitsuka, H. (2009b). Field independent probabilistic model for clustering multi-field documents. Information Processing

& Management, 45(5), 555–570.

[33]D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res., 5:361 – 397, 2004.

[34]T.-Y. Liu, Y. Yang, H. Wan, Q. Zhou, B. Gao, H.-J.Zeng, Z. Chen, and W.-Y. Ma. An experimental study on large-scale web categorization. In WWW ’05: Special interest tracks and posters of the 14th international conference on World Wide Web, pages 1106 – 1107, New York, NY, USA, 2005. ACM.

[35] I. Katakis, G. Tsoumakas, and I. Vlahavas. Multilabel text classification for automated tag suggestion. In Proceedings of the ECML/PKDD 2008 Discovery Challenge, 2008.

[36] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In ICML ’04: Proceedings of the twenty-first international conference on Machine learning, page 104, New York, NY, USA, 2004. ACM.

[37] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In NIPS, pages 721 – 728, 2002.

[38] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. Pattern Recogn.,40(7):2038 – 2048, 2007.

[39] N. Ghamrawi and A. McCallum. Collective multi-label classification. In CIKM’05: Proceedings of the 14th ACM international conference on Information and knowledge management, pages 195 – 200, New York, NY, USA, 2005. ACM Press.

[40]S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In SIGIR, 2005.

[41]R. Yan, J. Tesic, and J. R. Smith. Model-shared subspace boosting for multi-label classification. In KDD’07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 834 – 843, 2007.

[42]Ji, Tang, Yu, and Ye. Extracting shared subspace for multi-label classification. In

KDD ’ 08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 381 – 389, New York, NY, USA, 2008. ACM.

[43]Huang, M., Névéol, A., and Lu, Z. (2011). Recommending MeSH terms for annotating biomedical articles. Journal of the American Medical Informatics Association, 18(5),660-667.

[44] Mao, Y., and Lu, Z. (2013). NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. Technical report.

[45] Liu, K., Wu, J., Peng, S., Zhai, C., Zhu, S. (2014) The Fudan-UIUC Participation in the BioASQ Challenge Task 2a: The Antinomyra system. In Working Notes for CLEF 2014 Conference, Sheffield, UK (pp. 1311-1318).

[46] Partalas, I., Gaussier, É., and Ngomo, A. C. N. (2013). Results of the First BioASQ Workshop. In BioASQ@ CLEF (pp. 1-8).

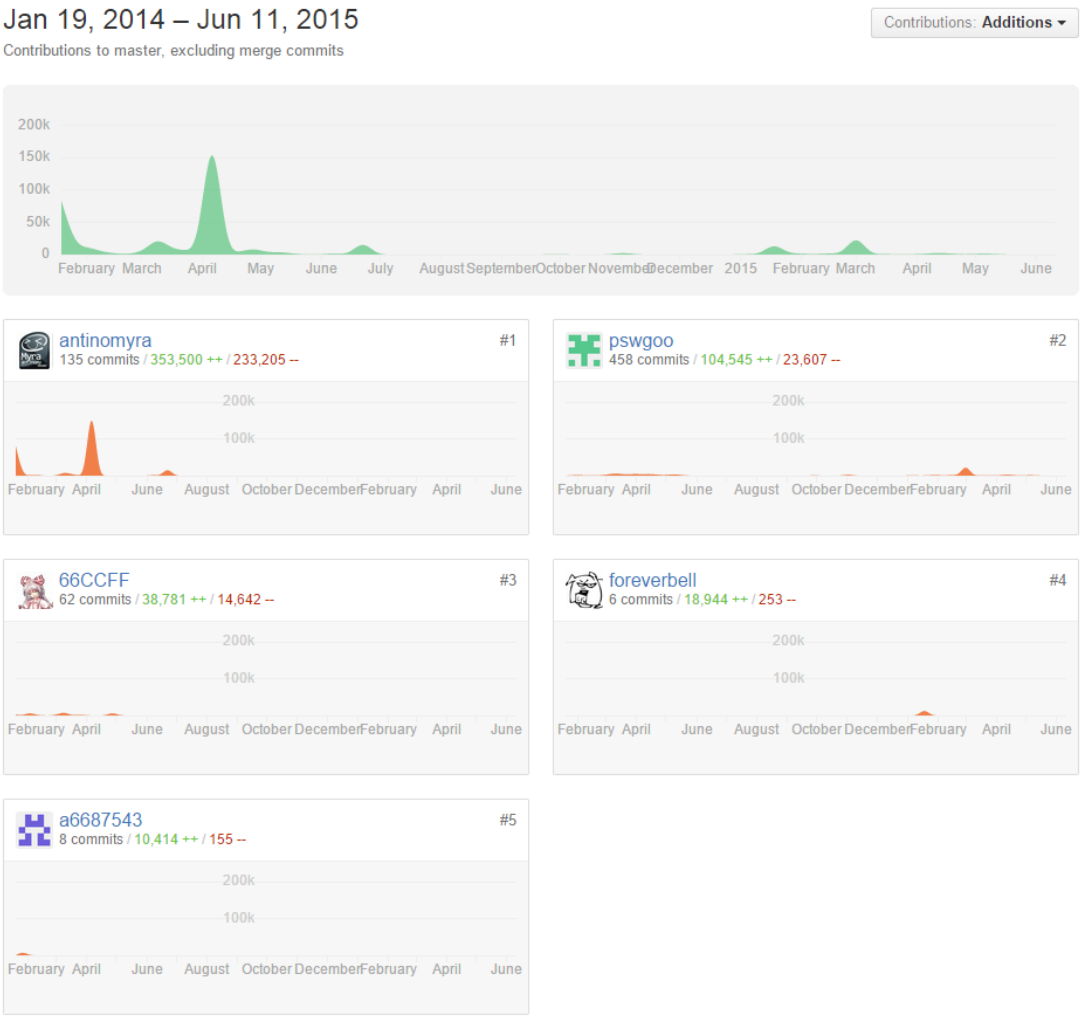
[47] Balikas, G., Partalas, I., Ngomo, A. C. N., Krithara, A., Gaussier, E., and Paliouras, G. (2014). Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. CLEF (Working Notes), pages 1181–1193.

[48] Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., and Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics, 25(11), 1412-1418.

[49] Tsoumakas, G., Laliotis, M., Markantonatos, N., and Vlahavas, I. P. (2013). Large-Scale Semantic Indexing of Biomedical Publications. In BioASQ@ CLEF.

致　谢

感谢以下同学为直接这篇文章的项目贡献代码，包括现复旦大学计算机系硕士在读的彭声闻同学，现在中南大学智能科学系本科在读的吴隽逍同学，现复旦大学数学系本科在读的董麒麟同学和现清华大学交叉信息研究院本科在读的彭天翼同学。



感谢朱山风老师一直以来对我的关心和支持，一直以来给我的非常耐心的指导。感谢我的家人，他们几乎一直支持我的所有决定，并在本文完成期间给予直接的帮助。