一种融合语义资源的生物医学查询理解方法

徐 博[1),3)] 林鸿飞[1)] 林 原[2)] 许 侃[1)]

[1)](大连理工大学 计算机科学与技术学院 辽宁 大连 116024)

[2)](大连理工大学公共管理与法学学院 辽宁 大连 116024)

[3)](认知智能国家重点实验室(科大讯飞) 合肥 230088)

摘 要 近年来,随着生物医学相关研究的快速发展,生物医学文献的数量与日俱增,相关人员从海量文献中获取所需信息变得越来越困难,信息检索技术能够为用户提供所需信息,但由于领域专业度高,术语庞杂,传统通用领域的检索技术往往很难胜任这项任务,而生物医学领域存在丰富的语义资源,这些资源涵盖该领域专业术语,可以对文献检索起到辅助和提升作用.因此,为进一步提升生物医学文献检索的性能,该文尝试基于词共现查询扩展模型,结合生物医学领域特征,利用医学主题词表衡量扩展词的重要性,综合权衡扩展词与查询词的共现关系和扩展词在医学主题词表中的分布情况,选择优质扩展词;并在此基础上提出一种基于组排序学习的监督式查询扩展方法,该方法根据候选扩展词对检索性能的影响和候选扩展词能否反映查询的主题信息两个方面对扩展词进行相关性标注,提取与扩展词相关的上下文特征和领域语义特征对扩展词进行向量化表示,最后采用组排序学习方法训练扩展词选择模型,完成查询扩展.在 TREC 基因任务数据集上的实验结果表明,该方法能够有效提升查询扩展性能,与基于排序学习方法 ListMLE 的监督式查询扩展方法相比,在文档平均准确率方面分别提升 4.41% 和 11.35%,有效提升了生物医学文献检索的综合性能.

关键词 生物医学文献检索;医学主题词表;词共现模型;查询扩展;组排序

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.02160

A Biomedical Query Understanding Method Based on Semantic Resources

XU Bo[1),3)] LIN Hong-Fei[1)] LIN Yuan[2)] XU Kan[1)]

[1)](*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

[2)](*School of Public Administration and Law, Dalian University of Technology, Dalian, Liaoning* 116024)

[3)](*State Key Laboratory of Congnitive Intelligence, IFLYTEK, Hefei* 230088)

Abstract In recent years, with the rapid progress in biomedical research, the number of biomedical literature increases rapidly, which becomes a big problem for researchers to obtain the needed information manually. Traditional information retrieval technologies can hardly achieve ideal performance for biomedical retrieval because of some domain-specific characteristics, especially the mismatching on biomedical terminologies. Query expansion method can deal with the problem by adding relevant terms to interpret users' query and fulfill the information need. Given that biomedicine domain has abundant semantic resources, which contain a large amount of terminologies and may assist the retrieval process, we first propose a novel query expansion model based on co-occurrence model and MeSH thesaurus. The model can help to choose the useful expansion terms by balancing the co-occurrences of terms and the distribution of terms in MeSH. Furthermore, based on the MeSH-based method, we obtain a large set of candidate expansion terms, and proposed to select high-quality expansion terms using group ranking methods for supervised query

收稿日期:2016-11-21;在线出版日期:2017-07-18. 本课题得到国家自然科学基金(61632011,61572102,61602078,61562080)、认知智能国家重点实验室开放基金(COGOS-20190001)、博士后科学基金面上项目(2018M641691)、教育部人文社会科学基金青年项目(19YJCZH199)及中央高校基本科研业务费专项资金(DUT18ZD102)资助.徐 博,博士,助理研究员,主要研究领域为信息检索、查询扩展和排序学习. E-mail: xubo@dlut.edu.cn. 林鸿飞,博士,教授,博士生导师,研究方向为搜索引擎、文本挖掘、情感计算和自然语言处理. 林 原,博士,副教授,研究方向为信息检索和排序学习. 许 侃,博士,高级工程师,研究方向为专利检索和查询扩展.

expansion. Compared with unsupervised query expansion, supervised query expansion takes much more information about candidate expansion terms at the same time to refine the set of expansion terms, and improve the quality of the expanded queries.

Specifically, we use a group-based modified ListMLE method to learn the term selection models. ListMLE is a listwise ranking method based on permutation likelihood probability between targeted ranking list and optimum ranking list, and group sampling divides its sample space further by taking one query term with higher relevance label and several terms with lower relevance labels as a group. The modified ListMLE model can thus be learned with more focus on the expansion terms with higher relevance label, which contributes much on the quality of the expanded queries. To give each candidate expansion term a ground truth label, we not only consider the latent impact of the term on retrieval performance, but also consider whether the term is contained in topic terms of the corresponding biomedical query. Therefore, we generate more accurate term labels, which will be taken as the learning targets of the term selection model. Besides, we extract term features based on both the context information and domain-specific information. The context information refers to term occurrences and co-occurrences with query terms in the context of retrieval, and the domain-specific information refers to the term importance or preference detected using biomedical semantic resources. Both feature sets can be useful to represent and describe the candidate expansion terms, and feature vectors of terms are then taken as the input of the group-based ListMLE method to train the term selection model.

We examine the effectiveness of our method on two TREC Genomics Track datasets. Experimental results show that our MeSH-based query expansion method can help to choose a set of high-quality candidate expansion terms, and expanded queries based on term selection models using the modified group-based ListMLE outperforms the term selection model using original ListMLE method, achieving 4.41% and 11.35% improvements in terms of document MAP on the two datasets, respectively.

Keywords　biomedical literature retrieval; medical subject headings; co-occurrence model; query expansion; group ranking

1　引　言

近年来，随着生物医学(Biomedicine)领域的快速发展，生物医学相关研究取得较多有价值的成果，这些成果不仅促成一些疾病的治疗，也推动了人类对于自身认识的深入发展. 与此同时，生物医学文献数量也与日俱增，文献中涵盖的信息量呈指数性增长，这些文献和所包含的信息能够辅助生物医学研究人员和相关从业人员把握最新研究进展，推动相关研究的进一步开展.

然而，海量的文献信息很难通过传统的手工方式获取，因此需要借助于现代信息检索技术和方法，协助相关人员获取所需的信息. 信息检索技术能够根据用户提交的查询，对文献进行相关性排序，并将排序结果返回给用户. 而直接将传统的信息检索方法应用于生物医学文献的检索任务中，很难取得较好的检索性能. 其原因在于未能充分考虑生物医学领域的固有特点，例如生物医学词汇的专业性和多样性等，同时，这些专业词汇往往存在很多同义词和缩写词的情况. 如果能在传统的信息检索方法中充分考虑生物医学领域的特点，在检索中融入语义资源，将会进一步提高生物医学信息检索的性能.

查询扩展技术是传统信息检索领域的关键技术之一. 它能够在用户提交的原始查询的基础上，根据用户的检索意图，对查询进行补充和完善，从而得到更符合用户检索意图的查询，提高检索的性能. 现有的查询扩展方法按照扩展词的来源不同可以分为两大类：一类是基于文档集合的查询扩展方法，这类方法以全部数据文档集合或者部分数据文档集合为研究对象，从中提取与查询相关的内容，完善原始查询；另一类是基于外部扩展资源的查询扩展技术，外

部资源主要包括有词典资源、检索系统查询日志、锚文本和维基百科等，很多研究表明利用外部扩展资源完善原始查询，可以更好地完成查询扩展任务，进而提升检索的性能.

查询扩展按照其扩展词选择过程不同可以划分为非监督式查询扩展和监督式查询扩展.传统经典查询扩展方法以非监督式查询扩展为主，这些方法以特定扩展词评估函数为依据，选择高质量扩展词扩展原始查询，进而提升检索效果；近年来，一些研究表明基于单一扩展词评估函数的非监督式查询扩展方法在泛化过程中具有一定局限性，其原因主要源自于单一评估函数很难充分考虑扩展词与原始查询在不同维度上的相关性，而监督式查询扩展方法能够在很大程度上弥补这些局限.

所谓监督式查询扩展是采用监督式机器学习方法选择扩展词，其优势在于能够以特征的形式从不同维度充分度量扩展词与原始查询的相关性，并通过损失函数最小化的方法训练得到扩展词选择模型，用于扩展词的选择和精炼.这类方法在多个检索任务中均被证明具有较好的查询扩展效果，因此本文拟采用监督式查询扩展提升生物医学文献检索的性能.

同时，考虑到生物医学领域存在较多语义资源，如果能在信息检索的过程中，充分利用这些资源对用户提交的查询进行补充和完善，检索的性能将有很大可能性得到进一步提升.基于以上考虑，本文提出一种基于医学主题词表(MeSH)的生物医学文献检索方法.该方法分为两个阶段：非监督式候选扩展词选择阶段和监督式查询扩展阶段.

在非监督式候选扩展词选择阶段，本文方法一方面通过伪相关反馈过程，从反馈文档集合中提取候选扩展词，并根据候选扩展词和原始查询词的共现关系对扩展词加权；另一方面基于医学主题词表中的词分布情况，对候选扩展词的重要性进行进一步衡量，从而综合权衡扩展词在文档集合和外部资源中的重要性；在监督式查询扩展阶段，本文提出一种基于组排序学习的扩展词选择模型，该模型以扩展词特征向量为输入，扩展词特征向量主要基于扩展词在上下文中的分布信息和扩展词在语义资源中的分布信息进行抽取，并根据扩展词和原始查询的相关性对扩展词进行自动化标注，作为模型训练的目标值，采用组排序学习方法，通过迭代优化组排序损失，训练得到扩展词选择模型，用于查询扩展和二次检索.

在TREC数据集上的实验结果表明，本文方法能够有效提升生物医学文献检索的性能，在非监督式候选扩展词选择阶段选出大量具有潜在有用性的扩展词，并通过监督式查询扩展过程，对扩展词进一步精炼和优化，更好地完成查询扩展，提升生物医学文献检索的整体性能.

本文第2节介绍相关工作；第3节详细阐述本文提出的基于医学主题词表的非监督式候选扩展词选择过程；第4节详细阐述本文提出的基于组排序学习的监督式查询扩展过程；第5节通过实验检验本文方法的性能，并对结果进行分析和讨论；第6节总结全文并对未来工作给予展望.

2 相关工作

2.1 查询扩展方法相关研究

查询扩展是一种有效的信息检索技术，其目的在于向用户提交的原始查询中补充一些相关词汇，丰富和完善原始查询，构建更加符合用户信息需求的扩展查询，更好地完成检索任务，提高信息检索的性能.所添加或补充的相关词汇通常来源于初次检索列表中排序靠前的文档，这类方法被称作伪相关反馈方法，该方法假设初次检索得到的排序靠前的文档为相关文档，称作伪相关文档，而这些文档中出现的词汇也与原始查询具有更高相关性，因此，通常选择其中的高频词对原始查询进行补充，实现查询扩展.经典的伪相关反馈模型已在多种信息检索模型中被实现，例如向量空间模型[1]、概率模型[2]、相关模型[3]和混合模型[4]等.

在扩展查询中，不同的扩展词对于检索性能的提升贡献程度也不一样，因而一些研究致力于寻求更为精准的方式对扩展词的重要程度进行权衡.例如，Chen等人[5]通过引入并定义大量基于语言和统计的词特征，探索查询词隐含的关联，进而有效衡量不同查询词与原始查询的相关性.近年来，监督式查询扩展方法得到广泛关注，并在不同检索任务中取得较好的效果[6-9].这类方法以经典的非监督式查询扩展方法为基础，以监督式机器学习方法为手段，通过监督学习过程训练得到扩展词选择模型.由于在扩展词表示中可以同时考虑与扩展词和原始查询相关的多维度信息，这类方法能够有效改善非监督式查询扩展方法的不足，改善扩展查询的质量.例如，Cao等人[6]将扩展词划分为好的扩展词和差的扩展词，进而通过支持向量机分类器对扩展词进行分类，

有效选择更为优质的扩展词.

考虑到监督式查询扩展的效果，本文将伪相关反馈文档作为扩展词的来源，面向生物医学文献检索任务，对现有查询扩展方法进一步完善，提出一种基于组排序学习监督式查询扩展方法改善生物医学文献检索的效果.

2.2　排序学习相关研究

排序学习算法[10-17]是信息检索领域热门研究内容之一，并被广泛应用于多种自然语言处理任务，并取得较好的效果，例如基于排序学习的社区问答[18]和推荐系统[19]等. 排序学习以传统的机器学习为方法，以经典检索模型为特征，通过监督学习的过程训练排序模型.

与传统的机器学习方法相比，其优势一方面在于能够同时考虑多种不同检索模型，并将其作为特征用于模型选择，另一方面可以定义基于排序的损失函数，更有针对性地优化排序结果. 查询扩展中扩展词的选择同样涉及扩展词的多维度特征，而扩展词的精炼同样可以转化为排序问题，即选择排序最为靠前的扩展词作为扩展查询中的元素.

排序学习按照其损失函数计算方法的不同可以划分为三类：点级方法、对级方法和列表级方法. 这三类方法分别以单个文档、具有偏序关系的文档对和整个文档列表作为损失函数计算的依据. 一些研究表明列表级方法在很多任务中具有最好的效果. 组排序学习[20]是在列表级方法的基础上进一步优化样本空间，根据具有不同相关性的样本对原始输入数据进一步划分，从而使得模型更倾向于将具有较高相关性的样本排列在样本列表的前面，达到提升排序性能的目的. 因此本文拟采用基于组排序学习的监督式查询扩展方法，用于生物医学文献检索任务.

2.3　生物医学文献检索相关研究

近年来，在生物医学文献检索任务中，查询扩展方法已被引入用于提升检索的性能. 早在 1996 年，Srinivasan[21]就在 MEDLINE 数据集中率先引入查询扩展方法用于提升检索的有效性，并将其应用于 SMART 检索系统；近几年 Xu 等人[22]在基因文献检索任务中，综合对比基于全局、基于局部以及基于本体的查询扩展方法，为后续进一步探索针对该领域的信息检索模型奠定了基础；类似地，Rivas 等人[23]对生物医学信息检索中的查询扩展方法进行了综合的评估，所使用的扩展词来源包括基于查询的扩展词、基于语料的扩展词和基于语法的扩展词. 这些方法证明查询扩展能够有效提升生物医学文献检索的性能，而在检索模型中考虑领域特点能够很大程度上改善现有检索性能.

由于生物医学领域存在大量语义资源和词典，这些资源中涵盖领域内丰富的专业词汇，因而一些研究致力于研究基于语义资源的查询扩展. Drame 等人[24]提出在向量空间模型中利用 MeSH 词表进行查询扩展，并在 2014 年 ShARe/CLER eHealth 评测中获得较好成绩；Oh 等人[25]使用外部资源来改善传统伪相关反馈模型的效率，提升检索性能；Mao 等人[26]在语言模型架构中引入 MeSH 概念层，充分利用概念关联提升检索效果；Jalali 等人[27]提出一种语义查询扩展方法，用以对查询和文档所对应的概念进行匹配. 这些方法的初衷主要在于将外部资源作为查询扩展中扩展词的来源，由于基于外部资源的扩展词具有较高的领域依赖性，因而能够丰富原始查询的领域特征，提升检索性能. 然而，基于外部资源的查询扩展方法通常是通过直接衡量查询和扩展词的相似度来判断扩展词是否有效，一定程度上忽略了扩展词在外部资源中的重要程度.

为克服这一问题，本文通过在现有伪相关反馈方法中融入 MeSH 词表中的词信息，完善查询扩展过程，一方面考虑扩展词与原始查询的相关性，另一方面充分挖掘扩展词在 MeSH 词表中的重要程度，综合权衡二者关系，选择更为优质候选扩展词，该方法与现有方法的主要区别在于扩展词的来源不同，现有基于语义资源的方法通常将外部资源作为扩展词的来源，这一定程度上忽略了扩展词在检索集合中的词分布信息，而本文研究是将语义资源作为扩展词重要性度量的依据，同时结合扩展词在检索集合中的分布信息，更加全面地衡量扩展词的有用性；同时考虑到监督式查询扩展方法的良好性能，本文拟采用基于组排序学习方法的监督式查询扩展方法训练扩展词选择模型，提升查询扩展的效果和生物医学文献检索的整体性能.

3　基于医学主题词表的查询扩展方法

本章对本文提出的基于医学主题词表的查询方法进行详细介绍，该方法以语言模型作为基础检索模型，伪相关反馈方法作为查询扩展方法，通过基于扩展词共现和 MeSH 词表两种扩展词加权策略对扩展词的重要性进行综合权衡，来完成查询扩展.

3.1　基础检索模型

本文采用语言模型作为基本检索模型，在初次

检索和查询扩展后的二次检索中进行应用.根据用户给定的查询进行初次检索后,可以得到文档排序列表,选择列表中前 N 篇文档作为伪相关文档,提取扩展词,根据伪相关反馈基本假设,伪相关文档与原始查询具有较高相关性,因而,其中蕴含的词汇也与原始查询较为相关,从中提取扩展词能够起到丰富原始查询的作用.

3.2 扩展词选择模型

在获取伪相关反馈文档集合后,需要采取有效的方式对候选扩展词的重要性进行综合评估,因此本文方法主要从以下两个角度对候选扩展词的重要性进行度量.

3.2.1 基于词共现的扩展词加权

在一篇文档中,如果两个词共同出现的次数较多,可以认为这两个词具有较强的关联,因而可以通过考虑查询词和文档中候选扩展词的共现关系来衡量扩展词的重要性,当查询词和扩展词共现次数越多时,该扩展词与原始查询具有更高的相关性,本文使用共现词频对这一指标进行定量描述,其计算方法如下所示.

$$tf_{doc}(t,q)=\frac{\sum_{d\in D}\log(freq(t,d)+1.0)\cdot\log(freq(q,d)+1.0)}{\log|D|} \tag{1}$$

其中,d 代表伪相关反馈文档集合 D 中的一篇文档,$freq(t,d)$ 和 $freq(q,d)$ 分别代表在文档 d 中候选扩展词 t 和查询词 q 出现的次数,$|D|$ 代表伪相关反馈文档集合的大小.该指标可以衡量查询词 q 与扩展词 t 共现得分,该得分主要基于局部文档共现的情形,为进一步衡量扩展词和查询词在全局文档集合中的重要性,借鉴词频逆文档频率(tf-idf)加权的思想,本文引入逆文档频率,其定义如下.

$$idf_{doc}(t)=\log\frac{N-n(t)+1.0}{n(t)+1.0} \tag{2}$$

其中,N 代表全局文档集合中文档总数,$n(t)$ 代表整个文档集合中包含词 t 的文档个数,该指标可以衡量扩展词 t 在整个文档集合中的重要性,出现该词的文档数越少则认为该词越重要.结合以上两个指标,可以采用如下方式对整个查询与候选扩展词的重要性进行评估.

$$TFIDF_{DOC}(t,Q)=\sum_{q\in Q}idf_{doc}(q)\cdot idf_{doc}(t)\cdot\log(tf_{doc}(t,q)+1.0) \tag{3}$$

其中,Q 代表用户提交的原始查询,q 代表原始查询 Q 中的一个查询词.式(3)结合查询词与候选扩展词的共现词频、查询词的逆文档频率和候选扩展词的逆文档频率,在所有查询词上进行累加操作,该公式可以对扩展词 t 在文档集合中的重要性进行度量.

3.2.2 基于 MeSH 词表共现的扩展词加权

为充分考虑扩展词在生物医学领域内的重要性,本文采用 MeSH 词表对扩展词的重要进行进一步的评估,主要考虑扩展词在 MeSH 中的分布信息.在介绍该方法前,首先简要介绍 MeSH 词表,该词表全称医学主题词表,是由美国国立图书馆所管理的医学词汇资源,以树状层次化组织,涵盖大量专业术语和词条,2016 年最新发布的 MeSH 词表涵盖 27 883 个描述符和超过 87 000 个术语词,主要用于对生物医学文献数据库 MEDLINE 和生物医学搜索引擎 PubMed 中的文档索引和信息管理等.

由于 MeSH 词表中涵盖大量专业词汇,本文方法首先以候选扩展词在 MeSH 中出现的词频作为扩展词领域依赖性的度量,该指标定义如下.

$$tf_{MeSH}(t)=\frac{\log(freq(t,MeSH)+1.0)}{\log|T|} \tag{4}$$

其中,$freq(t,MeSH)$ 代表 MeSH 词表中出现该扩展词的频率,$|T|$ 表示 MeSH 中出现的词的总数.在此基础上,进一步考虑包含该扩展词的唯一词条的个数,这里唯一词条是指包含该扩展词的不重复出现的词条,类比逆文档频率的计算方法,可以认为如果包含某一扩展词的唯一词条的数目越少,则说明该扩展词具有更高的重要性,具体量化方式如下所示.

$$idf_{MeSH}(t)=\frac{M-m(t)+1.0}{m(t)+1.0} \tag{5}$$

其中,M 代表 MeSH 中包含的词条的总数,$m(t)$ 代表出现扩展词 t 的唯一词条的个数.将式(4)和式(5)进行结合,可以得到如下公式对扩展词在 MeSH 中的重要性进行评估.

$$TFIDF_{MeSH}(t)=idf_{MeSH}(t)\cdot\log(tf_{MeSH}(t)+1.0) \tag{6}$$

该加权策略借鉴信息检索领域词频逆文档频率 tf-idf 的加权策略,用以衡量候选扩展词 t 在整个 MeSH 词表中的重要性.

3.3 基于 MeSH 和词共现融合的查询扩展

上文介绍的两种扩展词加权策略,一方面是以扩展词在文档集合中与查询的共现关系为基础,另一方面是以扩展词在 MeSH 词表中的分布情况为基础,从两个不同方面对扩展词的重要性进行度量.本节拟结合以上两种加权策略,采用线性插值方式,从而实现扩展词重要性的综合评估.

$$score(t|Q)=\lambda\cdot\frac{TFIDF_{DOC}(t,Q)}{\sum_{t}TFIDF_{DOC}(t,Q)}+(1-\lambda)\cdot\frac{TFIDF_{MeSH}(t)}{\sum_{t}TFIDF_{MeSH}(t)} \quad (7)$$

公式(7)所示为最终的扩展词选择模型，该模型中 λ 为线性插值参数，取值范围为区间[0,1]，当 λ 取 0 时该模型退化为仅基于 MeSH 词表对扩展词加权，当 λ 取 1 时，该模型退化为只使用词共现模型对扩展词加权.

本文采用上述模型对伪相关文档中出现的所有词汇进行打分，按照分数高低对扩展词进行排序，进而选择排序最为靠前的 k 个词作为最终的扩展词，构造扩展查询，进行二次检索，完成整个查询扩展流程.

在基于医学主题词表的查询扩展方法处理中，首先采用用户原始查询进行初次检索，基于伪相关反馈假设，认为初次检索结果中排序靠前的文档与原始查询具有更大的相关性，因此从初次检索文档列表中选择前 N 篇文档作为扩展词的来源；扩展词选择过程主要基于上文提出的两种选择策略，综合考虑扩展词在文档集合和扩展词在医学主题词表中的分布信息，从而选择与原始查询最为相关的扩展词重构原始查询，得到扩展查询；最后，基于扩展查询，进行二次检索，以期获得最好的检索效果.

此外，由于扩展词的得分能够反映出扩展词相对于原始查询的重要程度，因此不同扩展词在扩展查询中发挥的作用也不尽相同. 基于这点考虑，本文进一步在查询扩展中采用扩展词的得分对不同扩展词进行加权，从而构造出具有不同词项权重的扩展查询，最终完成二次检索. 实验中，本文分别对比词项加权扩展查询与词项不加权扩展查询对检索性能的影响，并尝试采用更加合理有效的词项权重改善扩展查询的整体质量.

4　基于组排序学习的监督式查询扩展

4.1　监督式查询扩展流程

本节简要介绍监督式查询扩展的算法流程，并在后续小结详细介绍如何改进该方法并将其应用于生物医学文献检索任务，算法流程如算法 1 所示.

算法 1.　监督式查询扩展流程.

·训练扩展词选择模型 M：

1：基于非监督查询扩展方法为每一个训练查询 q 选择 k 个候选扩展词；

2：扩展词有用性标注；

3：扩展词特征抽取，将每一个扩展词表示为特征向量；

4：基于组排序学习算法训练扩展词选择模型.

·基于扩展词选择模型 M 的查询扩展：

1：基于非监督查询扩展方法为每一个训练查询 q 选择 k 个候选扩展词；

2：扩展词特征抽取，将每一个扩展词表示为特征向量；

3：基于模型 M 选择前 m 个扩展词用于查询扩展；

4：基于扩展查询二次检索.

监督式查询扩展算法主要分为两个阶段：模型训练阶段和基于模型的查询扩展阶段. 在模型训练阶段，首先通过非监督查询扩展方法为每一个训练查询选择大量候选扩展词，其中非监督查询扩展方法采用本文第 3 节提出的基于 MeSH 的查询扩展方法；然后根据扩展词对检索性能的影响，对扩展词质量进行有用性标注；同时抽取扩展词特征，将每一个扩展词都表示为特征向量；最后基于组排序学习算法训练得到扩展词选择模型. 在基于模型的查询扩展阶段，针对于每一个测试查询采用非监督查询扩展方法获取大量候选扩展词，并将每一个扩展词表示为特征向量；之后应用训练好的模型，选择其中最为有效的扩展词，重构原始查询，实现二次检索.

接下来主要针对扩展词有用性标注、扩展词特征抽取以及基于组排序学习的扩展词选择模型训练过程进行详细介绍.

4.2　候选扩展词有用性标注

候选扩展词有用性标注是为每一个候选扩展词赋予一个真实标签，该标签能够反映出该扩展词相对于原始查询的作用，同时该标签也作为模型训练过程中的真实值，用于损失函数计算和模型优化. 在监督式查询扩展中，当前常用的一种候选扩展词有用性标注策略是根据候选扩展词对检索性能的潜在影响将其标注为相关或者不相关[6]. 具体标注策略可通过如下公式描述.

$$label(t)=\begin{cases}0, Eval(t,q)\leq Eval(q);\\ 1, Eval(t,q)>Eval(q)\end{cases} \quad (8)$$

其中，t 为某一个候选扩展词，q 为原始查询，$Eval$ 为某一种检索性能评价指标，例如平均准确率(MAP)等. 该标注策略的核心思想在于首先根据原始查询进行初次检索，并记录其检索性能值为 $Eval(q)$；然后将与之相关的某一候选扩展词加入原始查询，构成扩展查询，并基于该扩展查询进行检索，记录其检索性能值为 $Eval(t,q)$；若原始查询的检索效果优于该扩展查询，则说明该候选扩展词未能提升原始查询的效果，因此将该扩展词标注为 0，即不相关，反之，则将该扩展词标注为 1，即查询相关.

在将该标注策略应用于生物医学文献检索时，充分考虑该任务的特点，本文将该策略进行改进.与其他检索任务不同，生物医学文献检索任务的文档相关性标注不仅标注了与某一查询相关的一系列文档，同时也针对每一篇相关文档，标注其所能反应出的查询的主题信息，因此生物医学文献检索任务的目标就转化为获得相关文档列表，同时该列表尽可能多地覆盖相应查询的主题信息，从而为用户提供更加全面的检索结果，满足其信息需求.例如针对生物医学查询“How does P53 affect apoptosis?”(蛋白质 P53 如何影响细胞凋亡)，所标注的主题短语包括“apoptosis regulatory proteins”(凋亡调节蛋白质)“tumor suppressor protein P53”(肿瘤抑制蛋白 P53)和“gene expression”(基因表达)等.考虑到主题短语与原始查询的相关性，本文在式(8)标注策略的基础上进一步考虑候选扩展词在主题短语中出现的情况，对标注策略做如下改进，如表 1 所示.

表 1 候选扩展词标注策略

标注	$Eval(t,q)>Eval(q)$	$TopicTerm(t,q)$
0	否	否
1	是	否
1	否	是
2	是	是

其中，$TopicTerm(t,q)$表明候选扩展词 t 是否出现在相应查询 q 的主题词中，若出现则说明该词有助于获取某一主题相关的文档，因此与查询较为相关.该标注策略同时兼顾候选扩展词对检索性能的影响和与查询主题的相关性，将既能提升检索性能又可以作为查询主题词的候选扩展词标注为 2，即确定相关，将仅能提升检索性能或者仅能覆盖查询主题的候选扩展词标注为 1，即可能相关，将既不能提升检索性能也不能作为查询主题词的候选扩展词标注为 0，即不相关.该设置既考虑了候选扩展词对检索性能的潜在影响，又考虑生物医学文献检索任务的特点，因此能对候选扩展词的有用性给出更为精确的标注.

4.3 候选扩展词特征抽取

监督学习算法的输入需要将输入样本转化为特征向量的形式，特征需要充分反映样本在不同维度上的特性.同样，监督式查询扩展需要将候选扩展词表示为特征向量的形式，而所选取的扩展词特征需要从不同维度上反映出扩展词与原始查询的相关性.而基于监督式查询扩展的生物医学文献检索中所提取的候选扩展词特征一方面需反映扩展词与原始查询的相关性，另一方面也需体现扩展词的领域依赖性，即所选扩展词在查询所涵盖领域内的重要性，因此本文定义两类候选扩展词特征：基于上下文的词特征和基于领域的词特征.

4.3.1 基于上下文的候选扩展词特征

基于上下文的候选扩展词特征，主要考虑扩展词在检索语料中的词分布信息，及其与查询词分布信息之间的差异.因此本文主要从两个方面抽取基于上下文的候选扩展词特征：扩展词分布信息和查询词与扩展词共现信息.

扩展词分布信息主要考虑扩展词的词频 TF、逆文档频率 IDF 和其组合，这类特征定义如下.

$$tf(t_j)=\frac{\sum_{d\in D} freq(t_j,d)}{\sum_{d\in D} length(d)} \tag{9}$$

其中，t_j表示任意候选扩展词，d 表示文档集合 D 中的任意一篇文档，$freq(t_j,d)$表示文档 d 中词 t_j出现的次数，$length(d)$表示文档 d 中总共包含的词数.该公式用来计算候选扩展词 t_j在文档集合 D 内的词频.

$$idf(t_j)=\log\frac{Num(D)-n(t_j)+1.0}{n(t_j)+1.0} \tag{10}$$

$Num(D)$表示文档集合 D 中总共包含的文档个数，$n(t_j)$表示集合 D 中出现候选扩展词 t_j的文档个数.该公式用于计算候选扩展词 t_j的逆文档频率，逆文档频率可以反映扩展词在整个文档范围内的重要程度.

$$tfidf(t_j)=\frac{tf(t_j)\times idf(t_j)}{\sum_{\forall t_i w.r.t^q} tf(t_i)\times idf(t_i)} \tag{11}$$

其中，该公式将候选扩展词的词频和逆文档频率相结合，可以更为全面度量扩展词 t_j在文档集合内的重要性.以上三类特征主要根据扩展词在检索集合中的分布信息进行抽取.而扩展词与查询的共现可以更大程度地反映出扩展词对于给定查询的重要性，如果扩展词与查询词在同一文档上下文中共现频率较高，则可以认为该扩展词更有可能与原始查询具有较高相关性.基于这点考虑，本文进一步定义并抽取基于词共现的扩展词特征，其定义可以表示如下.

$$cooc(t)=\sum_{q\in Q}\sum_{d\in D} cooccurrence(q,t,d) \tag{12}$$

其中，$cooccurrence(q,t,d)$表示扩展词 t 与某一查询词 q 在文档 d 范围内共同出现的次数，在所有文档和所有查询词上对该共现频次进行累加，可以得到

扩展词 t 相对于整个查询的共现次数，并将其作为一类特征.

在文档集合 D 的选取上，本文在实验中尝试从整个检索语料集合和伪相关反馈文档集合两个方面着手，并抽取不同的词特征，用于候选扩展词的向量化表示.

4.3.2　基于领域的候选扩展词特征

在生物医学文献检索中，专业术语往往能够在很大程度上辅助刻画用户的信息需求，因此术语中所包含的词很有可能对原始查询进行补充和完善，从而改善检索效果. 因此本文基于生物医学语义资源提取候选扩展词特征，这些特征能够反映出扩展词的领域依赖性和重要性，更加准确地实现扩展词的向量表示. 本文提取的基于领域的候选扩展词特征包括两类：基于 MeSH 词表的扩展词特征和基于术语概念的扩展词特征.

基于 MeSH 词表的扩展词特征主要包括包含扩展词的唯一词条的个数、扩展词的 MeSH 词频和二者的组合，具体定义如第 3.2 节式(4)、式(5)和式(6)所示.

为抽取基于术语概念的扩展词特征，本文基于生物医学领域自然语言处理工具 MetaMAP[28] 识别扩展查询所涵盖的生物医学概念. MetaMAP 由美国国家医学图书馆开发并发布，用于识别给定生物医学文本片段中所涵盖的术语和概念. 具体来说，首先将一个候选扩展词加入原始查询构成扩展查询，然后将该扩展查询输入 MetaMap，可以得到若干与该扩展查询相关的概念，若结果中包含较多概念，则可以认为该查询能够涵盖更多有用信息，而该候选扩展词更有可能被选作扩展词，用于查询扩展[29]. 基于上述考虑，定义如下扩展词特征.

$$concept(t)=count(t,Q_{expand}(t)) \tag{13}$$

其中，t 表示任意候选扩展词，$Q_{expand}(t)$ 表示在原始查询中加入候选扩展词 t 的扩展查询. 该特征累计基于候选扩展词 t 的扩展查询所涵盖的概念的总数，概念基于 MetaMap 进行识别. 此外，MetaMap 在识别文本片段所涵盖的概念的同时，还会根据文本中所出现的词，返回若干候选概念，候选概念的个数同样可以作为扩展词领域重要性的度量依据，因此本文进一步定义如下特征.

$$candidate(t)=\frac{\sum_{q\in Q_{expand}(t)}|R(q)|}{|Q_{expand}(t)|} \tag{14}$$

其中，$R(q)$ 表示基于扩展词 t 的扩展查询中任意一个查询词所返回的候选概念的个数，$|Q_{expand}(t)|$ 表示扩展查询中所涵盖的概念总数.

基于以上特征定义，可以将扩展词表示为特征向量的形式，用于监督学习算法的输入，接下来具体阐述模型的训练过程.

4.4　基于组排序学习的扩展词选择模型

组排序学习方法在经典排序学习方法的基础上对其样本空间进一步划分，从而使得训练模型具有更好的泛化能力和优质的排序性能[20]. 因此本文拟采用组排序学习方法用于扩展词选择模型的训练. 在样本组空间的划分上，根据候选扩展词的标注级别不同将候选扩展词划分为三种类别的组空间，分别是相关-可能相关词分组、可能相关-不相关词分组和相关-不相关词分组. 一个相关-可能相关词分组包含一个标注为 2 的相关扩展词和若干标注为 1 的不相关扩展词；一个可能相关-不相关词分组包含一个标注为 1 的可能相关扩展词和若干标注为 0 的不相关扩展词；一个相关-不相关词分组包含一个标注为 2 的相关扩展词和若干标注为 0 的不相关扩展词. 通过以上划分可以使得训练的模型更有针对性地选择相关扩展词，避免选择不相关扩展词，同时扩充原始的样本空间.

在学习方法选择上，本文以 ListMLE 排序学习方法[30]为基础改进组排序损失函数. ListMLE 是一种基于序列似然概率计算排序损失的列表级排序学习方法，它的损失函数基于 Luce 模型定义如下.

$$L(f;t^q,y^q)=\sum_{s=1}^{n-1}(-f(t^q_{y^q(s)}))+\ln(\sum_{i=s}^{n}\exp(f(t^q_{y^q(i)}))) \tag{15}$$

其中，y 是随机选择的最优扩展词排序序列，满足对于任意扩展词 t_i 和 t_j，若 t_i 的标注值大于 t_j 的标注值，则 t_i 排列于 t_j 之前.

由于原始的 ListMLE 损失在区分不同级别相关性的扩展词上具有局限性，而基于组样本空间的划分能够在原始查询空间划分的基础上，进一步更有针对性地将具有不同相关性级别的扩展词划分在不同的分组，从而增强模型在不同相关性扩展词上的区分能力. 因此本文采用组样本空间策略对该损失函数进行改进，具体定义如下.

$$L(f;t^g,y^g)=\sum_{s=1}^{n-1}(-f(t^g_{y^g(s)}))+\ln(\sum_{i=s}^{n}\exp(f(t^g_{y^g(i)}))) \tag{16}$$

该函数仅代表一个分组的损失，在模型训练中需要将所有分组的排序损失累计，采用梯度下降等

策略优化排序损失,以获得最优排序性能的扩展词选择模型,具体学习过程如算法 2 所示.

算法 2. 基于组排序学习的 ListMLE 方法.

输入:训练查询集合 Q,每个查询所对应的候选扩展词集合 T_q 以及相关性标注 Y_q;迭代次数 C,学习率 η

1. 对于集合 Q 中的任一查询 q,构造组样本空间.
2. 初始化模型参数 ω
3. for $c=1$ to C do
4. 计算排序损失梯度 $L(f;t^g,y^g)$
5. 计算损失梯度值 $\Delta\omega=\dfrac{\partial L(f;t^g,y^g)}{\partial\omega}$
6. 更新参数 $\omega=\omega-\eta\cdot\Delta\omega$
7. end for

输出:模型参数 ω

通过算法 2 可以训练得到扩展词选择模型,用于在候选扩展词集合中进一步甄选高质量扩展词,完成监督式查询扩展过程.

5 实 验

5.1 实验设置

实验数据集采用 2006 和 2007 两年的 TREC 会议 Genomics Track 评测数据集.该数据集涵盖来自于 49 个生物医学期刊的 162259 篇文献[31].文献按照预先定义的边界被划分为超过 1000 万个篇章,并进行篇章级别的检索任务.本文根据该任务执行篇章级别检索.查询集合共 62 个查询,其中包括 2006 年任务给出的 26 个查询(除去两个未标注出相关文档的查询)和 2007 年任务所给出的 36 个查询.该数据集作为公开的生物医学文献评测数据集,多年来被广泛应用于生物医学文献检索研究.实验中评价方法采用该 *TREC* 任务所提供的四种评价方式:Document MAP,Passage MAP,Aspect MAP 和 Passage2 MAP,分别对文档级、主题级和篇章级的检索结果平均准确率进行定量的评估,评价指标的详细定义可参阅文献[31].

基本检索环境采用 Indri 搭建,采用其提供的语言模型作为基准的检索方法.提前采用 Porter 算法对原文和查询进行词干化处理,并去除停用词,采用 Indri 提供的结构化查询语言构造扩展查询,具体方式如下,其中,α 代表原始查询的权重,取值范围为[0,1].

$$\#weight(\alpha\ Q_{original}\ (1.0-\alpha)\ \#combine\ (\#weight(w_1\ term_1\ w_2\ term_2\ \cdots\ w_k\ term_k))) \tag{17}$$

其中,$Q_{original}$ 代表用户提交的原始查询,($term_1$,$term_2$,…,$term_k$)代表最终选择的扩展词列表,(w_1,w_2,…,w_k)代表每个扩展词所对应的权重.

实验中,为获得监督式查询扩展方法在所有查询上的平均性能,本文分别在两组查询集合上采用五倍交叉验证,即根据查询编号将查询划分为训练集、测试集和验证集,其中训练集用于模型训练,测试集用于模型效果的测试,验证集用于模型参数的选择.最终给出的实验结果是在五个测试集上的平均值.

5.2 基于医学主题词表的查询扩展效果

本节基于 TREC 会议 2006 年基因任务的查询集合和 2007 年任务的查询集合进行实验.实验中对比方法包括:语言模型检索[32](Language Model),相关模型查询扩展检索[3](Relevance Model),词依赖模型检索[33](Term Dependency)和基于聚类的查询扩展模型(Cluster-based Model).其中相关模型是一种经典伪相关反馈方法,词依赖模型在查询扩展中考虑查询词和扩展词的完全独立和顺序依赖两种情形,基于聚类的查询扩展模型在医学文献查询扩展中引入 K-means 聚类选择扩展词,可以看作是一个较强的对比方法.

实验结果中的本文方法包括:对比仅考虑 MeSH 得分时的检索结果(MeSH-based Model),以及基于 MeSH 和词共现融合的查询扩展方法(Proposed PRF).在扩展词加权中的扩展词权重采用本文方法的最终打分(式(7))获得.实验结果如表 2 和表 3 所示.

表 2 基于 2006 查询集合的检索结果

检索模型	Document	Passage	Aspect	Passage2
Language Model	0.3178	0.0205	0.1983	0.0239
Relevance Model	0.3194	0.0207	0.2023	0.0240
Term Dependency	0.3198	0.0208	0.1785	0.0254
Cluster-based Model	0.3089	**0.0235**	**0.2644**	0.0258
MeSH-based Model	0.3176	0.0204	0.1902	0.0241
Proposed PRF	**0.3237**	0.0212	0.2037	**0.0260**

表 3 基于 2007 查询集合的检索结果

检索模型	Document	Passage	Aspect	Passage2
Language Model	0.2587	0.0646	0.2000	0.0876
Relevance Model	0.2678	**0.0720**	**0.2302**	0.0963
Term Dependency	0.2804	0.0683	0.1974	0.0939
Cluster-based Model	0.2651	0.0673	0.1987	0.0905
MeSH-based Model	0.2634	0.0706	0.2263	0.0941
Proposed PRF	**0.2810**	0.0705	0.1995	**0.0991**

从表 2 的实验结果可以看出,在三种对比方法语言模型、相关模型和词依赖模型的检索效果中,词依赖模型在大多数评价指标上取得最优值;而单纯依赖 MeSH 的扩展词选择方法效果未好于对比方

法，这说明在查询扩展中仅考虑 MeSH 分布对词项选择未能将检索性能产生提高；而本文基于 MeSH 和词共现融合的查询扩展方法能够取得更好的效果，结果表明本文提出的融合扩展词在文档集中分布和扩展词在 MeSH 中分布的查询扩展方法能够有效提升检索的效果.

从表 3 中可以发现类似的实验结果，从实验结果可以看出本文方法能够有效提升文献检索的性能. 这组实验中，相关模型方法在大部分评价指标中可获得较好的性能，基于 MeSH 的检索方法虽未优于其他对比方法，但同时考虑共现和 MeSH 词表的检索可以有效提升现有方法的性能.

上述结果表明，本文提出方法可以有效改善和提升现有检索性能，这是由于该方法在查询扩展中同时考虑扩展词在文档集合中的分布信息和扩展词在医学主题词表 MeSH 中的相对重要性. 扩展词在文档集合中的分布信息可以反映扩展词与原始查询之间的相关程度；而扩展词在医学主题词表 MeSH 中的相对重要性可以反映扩展词在生物医学领域中的重要程度. 因此通过上述设置，在查询扩展中兼顾扩展词与查询的关系和扩展词在领域中的重要性，从而更好地实现查询重构，并进一步提升检索的性能. 而通过该非监督式查询扩展方法选择的候选扩展词能够更大程度上覆盖高质量扩展词，便于在后续的监督式扩展词选择模型中对扩展词进一步筛选和精炼.

5.3　监督式查询扩展效果对比

本节实验在非监督式查询扩展选出大量候选扩展词的基础上，进一步采用监督式查询扩展方法，训练扩展词选择模型，选出查询相关的高质量扩展词，更好地完成查询扩展，提升检索的效果.

实验中的对比方法包括：基于两阶段监督式查询扩展方法（Two-stage SQE）[8]，该方法将监督式查询扩展分为查询选择阶段和模型训练阶段，作为最新的监督式查询扩展方法，可以看作是一个较强的对比方法；基于支持向量机分类器选择扩展词的方法（SVM）[6]，该方法是一种经典的监督式查询扩展方法，通过 SVM 将候选扩展词分类为好的扩展词和坏的扩展词，并根据分类器输出的后验概率得分选出最为有用的扩展词；基于迭代决策树的扩展词选择方法（MART）[12]，该方法和基于 SVM 的方法类似，用以选择候选扩展词，以上两种方法可以看作是点级的排序学习方法；而 RankNet[13] 和 RankBoost[15] 是两种对级的排序学习方法，RankNet 以神经网络模型为基础，以具有偏序关系的扩展词对间的序列概率计算排序损失，RankBoost 以梯度提升算法为基础，通过迭代的方式结合多个弱排序器，构成最终的扩展词选择模型；ListMLE 方法[30] 是列表级的排序学习方法，该方法以最优排序和预测排序之间的概率似然为依据计算排序损失，并以此为基础给出扩展词排序列表；Group-ListMLE 是本文提出的基于组排序学习的扩展词选择模型. 具体实验结果如表 4 和表 5 所示.

表 4　基于 2006 查询集合的监督式查询扩展性能

方法	Document	Passage	Aspect	Passage2
Proposed PRF	0.3242	0.0212	0.2040	0.0260
Two-stage SQE	0.3503	0.2417	**0.2612**	0.0289
SVM	0.3435	0.0249	0.2527	0.0306
MART	0.3434	0.0247	0.2505	0.0308
RankNet	0.3420	0.0236	0.2432	0.0292
RankBoost	0.3452	0.0251	0.2523	0.0309
ListMLE	0.3424	0.0241	0.2376	0.0300
Group-ListMLE	**0.3575**	**0.0263**	0.2587	**0.0337**

表 5　基于 2007 查询集合的监督式查询扩展性能

方法	Document	Passage	Aspect	Passage2
Proposed PRF	0.2818	0.0706	0.1996	0.0992
Two-stage SQE	0.3204	0.7652	0.2713	0.1056
SVM	0.3185	0.0809	0.2639	0.1112
MART	0.3140	0.0816	0.2589	0.1111
RankNet	0.2997	0.0769	0.2365	0.1070
RankBoost	0.3293	0.0832	0.2685	0.1153
ListMLE	0.3021	0.0791	0.2419	0.1052
Group-ListMLE	**0.3364**	**0.0847**	**0.2723**	**0.1192**

从表 4 的实验结果可以看出，相比于本文提出的非监督式查询扩展方法，监督式查询扩展方法在所有评价指标上均可以获得较高的检索效果；而在不同的监督式查询扩展方法中，对级的排序学习方法相比于点级的方法能够取得更好的检索效果，而基于 RankBoost 的排序学习方法要优于基于 RankNet 的方法；列表级方法 ListMLE 的性能介于 RankBoost 和 RankNet 之间，基于两阶段的监督式查询扩展方法相比于上述方法具有更好的检索效果，而本文提出的基于 Group-ListMLE 的扩展词分组选择方法获得了最优的检索效果，相比于基于 ListMLE 的方法提升幅度为 4.41%.

从表 5 的实验结果也可以看出类似的趋势，监督式查询扩展方法的检索性能均优于非监督的查询扩展方法，在点级方法中基于 SVM 的方法优于基于 MART 的扩展词选择方法，在对级方法中基于 RankBoost 的方法优于基于 RankNet 的方法，而基于组排序学习的列表级方法获得了最佳

的检索效果，相比于基于 ListMLE 的方法提升幅度为 11.35%.

上述结果表明，在生物医学文献检索任务中，监督式扩展词选择过程能够选出更多质量较好的扩展词，而包含所选扩展词的扩展查询能够更为清晰准确地描述用户信息需求，从而提升检索的效果；而基于组排序学习的方法相比于其他排序学习方法更加有效，其原因在于组排序学习对原始的查询相关的扩展词样本空间进一步划分，使得排序损失优化更具有针对性，从而将具有较高相关性级别的扩展词排列在预测列表的前面，改善查询扩展的效果.

5.4 不同候选扩展词标注策略对比

为进一步分析本文方法性能提升的原因，本节实验对第 4.2 节中的两种候选扩展词标注策略进行对比，其中二级标注指式(8)中提到的仅根据扩展词对检索效果潜在影响标注扩展词相关与否的方法，三级标注指本文提出的候选扩展词标注策略，即同时考虑扩展词对检索性能的影响和扩展词是否可作为查询的主题词. 实验结果如表 6 所示.

表 6 不同扩展词相关性标注策略对检索性能的影响

查询集合	标注策略	Document	Passage	Aspect	Passage2
2006 查询	二级标注	0.3129	0.0221	0.2579	0.0271
	三级标注	0.3575	0.0263	0.2587	0.0337
2007 查询	二级标注	0.3091	0.0796	0.2552	0.1093
	三级标注	0.3364	0.0847	0.2723	0.1192

从表 6 的实验结果可以看出，本文所提出的候选扩展词三级相关性标注策略均明显优于经典的二级相关性标注策略，这说明查询的主题信息能够反映出扩展词的有用性，而在标注中考虑候选扩展词是否出现在查询的主题词中可以进一步甄别扩展词的有用性，从而增强所训练的扩展词选择模型的性能.

5.5 候选扩展词特征选择

本节进一步分析和对比监督式查询扩展中所选择的不同候选扩展词特征集合对于检索性能的影响. 其中，文本特征指第 4.3.1 节所定义的基于上下文的候选扩展词特征集合，领域特征指第 4.3.2 节所定义的基于领域的候选扩展词特征，全部特征指上述两个特征集合的并集. 这组实验分别基于上述三个扩展词特征集合，采用 Group-ListMLE 方法，训练扩展词选择模型，用以对比不同特征集合在扩展词选择模型中的作用，除特征集合不同外，其它实验设置均相同. 实验结果如表 7 所示.

表 7 不同候选扩展词特征集合对检索性能的影响

查询集合	特征	Document	Passage	Aspect	Passage2
2006 查询	文本特征	0.3327	0.0241	0.2433	0.0300
	领域特征	0.3411	0.0244	0.2447	0.0280
	全部特征	0.3575	0.0263	0.2587	0.0337
2007 查询	文本特征	0.3145	0.0805	0.2552	0.1101
	领域特征	0.3250	0.0849	0.2606	0.1161
	全部特征	0.3364	0.0847	0.2723	0.1192

从表 7 的实验结果可以看出，仅基于领域特征的扩展词选择模型相比于仅基于文本特征的扩展词选择模型在检索性能上有小幅度提升，这说明生物医学领域的语义资源可以有效衡量扩展词的重要度，并且相比于仅基于上下文信息训练的扩展词选择模型具有更好的效果；而使用全部特征训练的扩展词选择模型取得了最佳的检索性能，从这点可以看出文本特征和领域特征二者可以从不同角度刻画扩展词的有用性，二者互相补充，对扩展词选择模型效果的提升具有相辅相成的作用，因此取得了较好的检索准确率.

5.6 扩展词加权效果对比

前文第 3.3 节曾提到本文方法在将所选择的扩展词加入到原始查询中时，可以选择是否根据扩展词得分对扩展词进行加权，如果得分能够有效反映扩展词的重要性程度，则加权后的扩展查询可以取得更好的检索效果，基于这点考虑，本节对扩展词加权与否进行实验和分析. 表 8 和表 9 分别给出在 2006 查询集合和 2007 查询集合上的实验结果，其中的方法包括本文提出的基于 MeSH 和词共现的非监督式查询扩展方法在扩展词加权和不加权情况下的检索性能，分别记作 UQE-weighted 和 UQE-unweighted；本文提出的基于组排序学习的监督式查询扩展方法在扩展词加权和不加权情况下的检索性能，分别记作 SQE-weighted 和 SQE-unweighted.

表 8 基于 2006 查询集合的扩展词加权效果对比

扩展方法-加权	Document	Passage	Aspect	Passage2
UQE-unweighted	0.3242	0.0212	0.2040	0.0260
UQE-weighted	0.3237	0.0212	0.2037	0.0260
SQE- unweighted	0.3439	0.0250	0.2540	0.0309
SQE- weighted	0.3575	0.0263	0.2587	0.0337

表 9 基于 2007 查询集合的扩展词加权效果对比

扩展方法-加权	Document	Passage	Aspect	Passage2
UQE-unweighted	0.3242	0.0212	0.2040	0.0260
UQE-weighted	0.3237	0.0212	0.2037	0.0260
SQE- unweighted	0.3273	0.0850	0.2638	0.1163
SQE- weighted	0.3364	0.0847	0.2723	0.1192

从实验结果可以看出，对于非监督式查询扩展方法，扩展词不加权的扩展查询相比于扩展词加权的扩展查询具有更好的检索性能，且二者的性能相差较小，这说明非监督式查询扩展方法所给出的扩展词得分对于衡量扩展词在扩展查询中的重要性作用较小；而对于监督式查询扩展方法，扩展词加权的扩展查询相比于扩展词不加权的扩展查询具有更好的检索性能，这说明组排序学习所给出的扩展词得分更为准确，能够更加精确地反映出扩展词在扩展查询中的重要性，进而有效提升检索的整体性能.

5.7　参数选择

在验证本文方法的有效性之后，本节对实验中的一些参数设置和选择的过程进行深入讨论. 本文方法参数包括 4 个，分别是伪相关反馈文档个数 N、扩展查询中扩展词的个数 k、扩展查询中原始查询权重 α 和本文方法中的线性插值参数 λ. 为保证实验参数选择的客观性，本文在参数选择中采用在两个数据集合上交叉验证的方式，即通过 2006 年任务所对应的查询选择 2007 年任务的最优参数，并用 2007 年任务所对应的查询来选择 2006 年任务的最优参数，评价指标选用文档平均准确率(Document MAP). 图 1、图 2、图 3 和图 4 给出上述参数在两组查询集合上的选择过程.

图 1 给出的是反馈文档个数 N 对检索性能的影响，在 2006 年任务的查询集合上，选择不同反馈文档个数时的检索性能波动较大，并在反馈文档个数为 60 时达到性能的最优值；而在 2007 年任务的查询集合上，检索性能在不同反馈文档个数的情况下波动较小，根据性能变化选择反馈文档个数为 10 个.

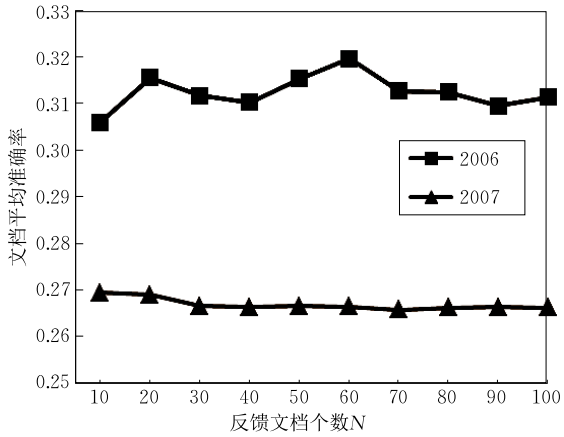

图 1　查询扩展中反馈文档个数 N 对检索性能的影响

图 2 给出的是扩展词个数 k 对检索性能的影响，从实验结果可以看出，在两组查询集合上，该参数的变化趋势较为一致，并在扩展词个数为 30 个时，检索性能达到最优.

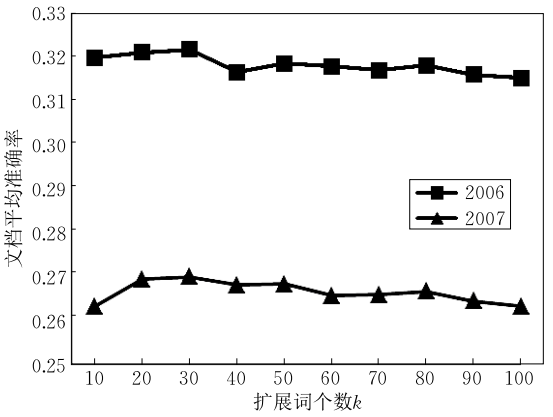

图 2　查询扩展中扩展词个数 k 对检索性能的影响

图 3 给出的是原始查询权重 α 对检索性能的影响，从图中可以看出，在两组查询集合上该参数的变化趋势较为一致，并分别在 0.8 和 0.7 时达到检索性能的最优值.

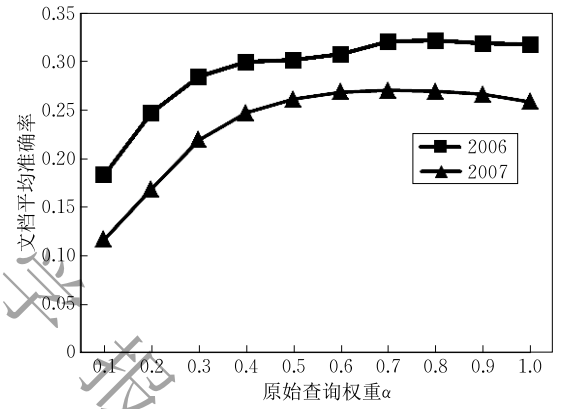

图 3　查询扩展中原始查询权重 α 对检索性能的影响

图 4 给出的是线性插值参数 λ 对检索性能的影响，从图中可以看出该参数在两组查询集合上的变化趋势并不一致，并分别在 0.3 和 0.7 时取得最优的检索性能.

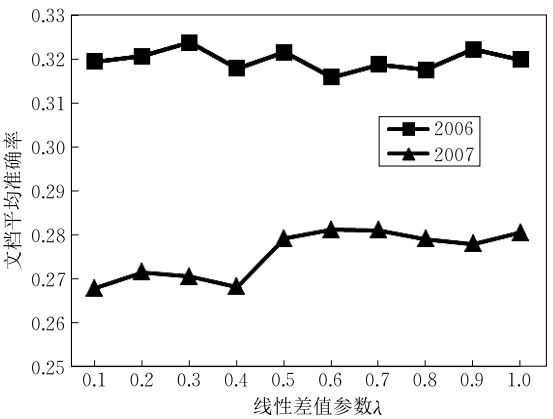

图 4　查询扩展中线性插值参数 λ 对检索性能的影响

5.8 分析与讨论

从上述实验结果可以看出本文所提出的基于组排序学习的监督式查询扩展方法能够有效提升生物医学文献检索的性能，本节对实验结果展开深入分析和讨论，总结本文方法的优势和不足，以期在未来工作中对该方法进一步优化，提升生物医学文献检索性能. 基于组排序学习的监督式查询扩展方法的有效性主要得益于四个方面的改进.

一是基于 MeSH 和词共现的候选扩展词选择. 该方法是一种融合语义资源的非监督式查询扩展方法，实验表明相比于其他非监督式查询扩展方法，该方法能够有效选择较高质量的候选扩展词，作为监督式查询扩展中扩展词选择的来源. 尽管该方法能够一定程度上提高查询扩展的效果，但该方法提升检索性能的幅度有限，同时所给出的扩展词权重不够准确，因此仍需进一步选择优质扩展词；

二是扩展词有用性标注策略. 本文提出的有用性标注策略不仅考虑扩展词对于检索性能的潜在影响，同时也考虑扩展词是否能够覆盖查询的主题信息，从而更有针对性地标注生物医学查询意图相关的扩展词. 本文实验中所用数据集显式给出了查询相关的主题词，而对于未显式标注查询主题词的数据集，可在预处理阶段采用主题模型对伪相关文档集进行主题分析，从而获取相应的主题词.

三是扩展词的特征提取. 本文方法同时提取与扩展词相关基于上下文信息的文本特征和基于语义资源信息的领域特征，两类特征相辅相成，能够从不同维度上刻画扩展词与原始查询的相关性，为训练更加有效的扩展词选择模型奠定基础.

四是基于组排序学习的监督式学习方法. 该方法相比于其他监督式学习方法，在原本的以查询划分的训练集的基础上基于分组对训练集进一步划分，丰富训练中的扩展词样本空间，使得模型更倾向于选择具有更高级别相关性的扩展词，从而提高扩展查询的质量.

以上四个方面共同作用，使得本文方法有效提升生物医学文献检索的效果，并优于其他对比方法. 相比于非监督式查询扩展，监督式查询扩展另一个值得考虑的问题是时间开销，其时间开销主要来源于模型训练和扩展词的特征提取，文献[8]指出，在实际应用中模型训练过程离线完成后，可直接应用于不同场景，因此监督式查询扩展在测试阶段的时间开销主要来源于扩展词特征的提取. 在未来工作中我们将探索并使用有效方法降低监督式查询扩展在扩展词特征提取上的时间开销，以进一步完善和优化本文方法，在提高检索效果的同时降低方法的时间开销.

6 结论和展望

本文提出一种基于医学主题词表 MeSH 的生物医学文献查询扩展方法，该方法基于伪相关反馈查询扩展，在扩展词选择中综合权衡扩展词与查询词共现以及扩展词在 MeSH 词表的词分布信息，从而选择更为优质的扩展词用于查询扩展；并在此基础上提出一种基于组排序学习的监督式查询扩展方法，该方法能够对非监督式查询扩展所选择的扩展词进一步甄别，选取其中高质量的扩展词用于查询扩展，以提升检索的整体性能. 在 TREC 数据集上的实验结果表明，本文方法相比于现有查询扩展方法能够有效提升生物医学查询扩展的效果，提高文献检索的整体性能.

未来工作可以从两个方面展开：一方面探索更为有效的监督式查询扩展方法，并将其改进和应用于生物医学文献检索上，提升检索的有效性和效率；另一方面尝试使用其它可用的生物医学语义资源，用于衡量扩展词的有用性，并提取候选扩展词相关文本特征，优化查询扩展过程，提高扩展查询的质量.

参 考 文 献

[1] Rocchio J. Relevance feedback in information retrieval//Proceedings of the SMART Retrieval System: Experiments in Automatic Document Processing, 1971: 313-323

[2] Robertson S E, Walker S, Beaulieu M, et al. Okapi at trec-4//Proceeding of the 4th Text Retrieval Conference. Gaithersburg, USA, 2016: 73-97

[3] Lavrenko V, Croft W B. Relevance based language models//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information retrieval. New Orleans, USA, 2001: 120-127

[4] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval//Proceedings of the 10th International Conference on Information and Knowledge Management. Atlanta, USA, 2001: 403-410

[5] Lee C J, Chen R C, Kao S H, Cheng P J. A term dependency-based approach for query terms ranking//Proceedings of the 18st ACM International Conference on Information and Knowledge Management. Hong Kong, China, 2009: 1267-1276

[6] Cao G, Nie J Y, Gao J, Robertson S. Selecting good expansion terms for pseudo-relevance feedback//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 243-250

[7] Lv Y, Zhai C X, Chen W. A boosting approach to improving pseudo-relevance feedback//Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 165-174

[8] Zhang Z, Wang Q, Si L, Gao J. Learning for efficient supervised query expansion via two-stage feature selection//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 265-274

[9] Xu B, Lin H, Lin Y. Assessment of learning to rank methods for query expansion. Journal of the Association for Information Science and Technology, 2016, 67(6): 1345-1357

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

[11] Qin T, Liu T Y, Xu J, Li H. LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval, 2010, 13(4): 346-374

[12] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2001: 1189-1232

[13] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 89-96

[14] Cao Z, Qin T, Liu T Y, et al. Learning to rank: From pairwise approach to listwise approach//Proceedings of the 24th International Conference on Machine Learning. Oregon, USA, 2007: 129-136

[15] Freund Y, Iyer R, Schapire R E, Singer Y. An efficient boosting algorithm for combining preferences. The Journal of Machine Learning Research, 2003, 4: 933-969

[16] Burges C J. From RankNet to LambdaRank to LambdaMART: An overview. Learning, 2010, 11: 23-581

[17] Quoc C, Le V. Learning to rank with nonsmooth cost functions//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2007, 19: 193-200

[18] Ji Z, Wang B. Learning to rank for question routing in community question answering//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, USA, 2013: 2363-2368

[19] Sun J, Wang S, Gao B J, Ma J. Learning to rank for hybrid recommendation//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 2239-2242

[20] Lin Y, Lin H, Ye Z. Learning to rank with groups//Proceedings of the 19nd ACM International Conference on Information & Knowledge Management. Toronto, Canada, 2010: 1589-1592

[21] Srinivasan P. Query expansion and MEDLINE. Information Processing & Management, 1996, 32(4): 431-443

[22] Xu X, Zhu W, Zhang X, et al. A comparison of local analysis, global analysis and ontology-based query expansion strategies for bio-medical literature search//Proceedings of IEEE International Conference on Systems, Man and Cybernetics. Taipei, China, 2006: 3441-3446

[23] Rivas A R, Iglesias E L, Borrajo L. Study of query expansion techniques and their application in the biomedical information retrieval. The Scientific World Journal, 2014(1): 132-158

[24] Dramé K, Mougin F, Diallo G. Query expansion using external resources for improving information retrieval in the biomedical domain//Proceeding of the conference and labs of the evaluation forum (Working Notes). Sheffield, UK, 2014: 189-194

[25] Oh H S, Jung Y. Cluster-based query expansion using external collections in medical information retrieval. Journal of Biomedical Informatics, 2015, 58: 70-79

[26] Mao J, Lu K, Mu X, Li G. Mining document, concept, and term associations for effective biomedical retrieval: Introducing MeSH-enhanced retrieval models. Information Retrieval Journal, 2015, 18(5): 413-444

[27] Jalali V, Borujerdi M R M. The effect of using domain specific ontologies in query expansion in medical field//Proceedings of International Conference on Innovations in Information Technology. Al Ain, Arab, 2008: 277-281

[28] Aronson A R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program//Proceedings of the AMIA Symposium. Washington, USA, 2001: 17

[29] Zhu D, Carterette B. An adaptive evidence weighting method for medical record search//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 1025-1028

[30] Xia F, Liu T Y, Wang J, et al. Listwise approach to learning to rank-theory and algorithm//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 192-199

[31] Hersh W R, Cohen A M, Roberts P M, Rekapalli H K. TREC 2006 genomics track overview//Proceedings of the Fifteenth Text Retrieval Conference. Maryland, USA, 2006: 14-23

[32] Ponte J M, Croft W B. A language modeling approach to information retrieval//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 275-281

[33] Lin Y, Lin H, Jin S, Ye Z. Social annotation in query expansion: A machine learning approach//Proceedings of the 34st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 405-414

XU Bo, Ph. D. postdoctoral. His research interests include information retrieval, query expansion, learning to rank.

LIN Hong-Fei, Ph. D., professor. His main research interests include search engine, text mining, sentiment analysis and natural language processing.

LIN Yuan, Ph. D., associate professor. His main research interests include information retrieval and learning to rank.

XU Kan, Ph. D., senior engineer. His main research interests include patent retrieval and query expansion.

Background

Biomedical information retrieval is becoming more and more important for biologists to capture their needed information, which becomes a hot research issue in the intersection of information retrieval field and biomedical field. Many studies focus on improving the performance of biomedical retrieval utilizing existing information retrieval technologies in combination with domain-specific semantic resources. Biomedical resources, such as Medical Subject Headings, are important domain-specific semantic resources to be used during the retrieval, and how to incorporate them attracts much attention from researchers.

In this work, we tackle the problem based on query expansion, a classic and effective information retrieval technique. In our method, we propose to obtain high-quality expansion terms not only based on the term distribution in corpus, but also based on semantic resources. We propose two query expansion methods for this task. One is based on unsupervised query expansion, and the other is based on supervised query expansion, which has not been applied for this task before. We modify existing supervised query expansion using well-defined term features and group-based ranking method. We examine the effectiveness of our method on two TREC Genomics Track datasets. Experimental results show that our MeSH-based query expansion method can help choose a set of high-quality candidate expansion terms, and expanded queries based on term selection model using the modified group-based ListMLE outperforms original ListMLE method, achieving 4.41% and 11.35% improvements in terms of document MAP on the two datasets, respectively.

Since our method is general, and future work about this study can be carried out from various aspects. For example, other domain-specific semantic resources can be introduced to extract useful term features to fulfill diverse query intents in different circumstances. Meanwhile, the optimization of term selection model can also be made for other related tasks to meet different requirements in these tasks by modifying the loss function with task-specific constraints.

This work is supported by grant from the Natural Science Foundation of China (Nos. 61632011, 61572102, 61602078, and 61562080), and the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK, P. R. China (No. COGOS-20190001), Postdoctoral Science Foundation of China (No. 2018M641691), the Ministry of Education Humanities and Social Science Project (No. 19YJCZH199), and the Fundamental Research Funds for the Central Universities (No. DUT18ZD102).