ACRank: 在神经排序模型中引入检索公理知识

薄 琳[1] 庞 亮[2] 张朝亮[3] 王钊伟[3] 董振华[3]
徐 君[4),5)] 文继荣[4),5)]

[1)](中国人民大学信息学院 北京 100872)
[2)](中国科学院计算技术研究所 北京 100190)
[3)](华为技术有限公司诺亚方舟实验室 广东 深圳 518129)
[4)](中国人民大学高瓴人工智能学院 北京 100872)
[5)](新一代智能搜索与推荐教育部工程研究中心 北京 100872)

摘 要 传统的信息检索(Information Retrieval,IR)是知识驱动的方法,如以BM25、LMIR等为代表的检索模型在设计过程中考虑词频、逆文档频率、文档长度等关键因素计算查询-文档的相关性得分.这些关键因素被总结为IR公理,在传统模型的设计和评价中起到了至关重要的作用.如词频规则认为有更多查询词的文档更相关.与之相对,数据驱动的神经排序模型基于大量的标注数据与精巧的神经网络结构自动学习相关性评分函数,带来了显著的排序精度提升.传统IR公理知识是否能用来提升神经排序模型的效果是一个值得研究的重要问题且已有学者进行了初步探索,其首先通过公理指导增强数据生成,然后利用生成的标注数据直接训练神经网络.但IR公理的形式是通过比较匹配信号的强弱给出两个文档间相对的相关关系,而非直接给出文档的相关度标签.针对这一问题,本文提出了一种通过对比学习将IR公理知识引入神经排序模型的框架,称为ACRank.ACRank利用信息检索公理生成增强数据,抽取不同文档的匹配信号,利用对比学习拉开匹配信号间差距,使正样本匹配信号强于负样本,通过上述方式,ACRank将IR公理知识自然地融入到数据驱动的神经排序模型中.ACRank作为通用框架,可应用于不同规则,本文选择词频规则进行实验,基于大规模公开数据集上的实验结果表明,ACRank能够有效提升已有神经检索模型如BERT的排序精度,相关分析实验验证了该框架的有效性.

关键词 神经检索模型;信息检索公理;对比学习;知识驱动;数据驱动
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.02117

ACRank: Injecting IR Axiomatic Knowledge to Neural Ranking Models

BO Lin[1)] PANG Liang[2)] ZHANG Chao-Liang[3)] WANG Zhao-Wei[3)] DONG Zhen-Hua[3)]
XU Jun[4),5)] WEN Ji-Rong[4),5)]

[1)](*School of Information, Renmin University of China, Beijing* 100872)
[2)](*Chinese Academy of Sciences, Beijing* 100190)
[3)](*Huawei Noah's Ark Lab, Shenzhen, Guangdong* 518129)
[4)](*Gaoling School of Artificial Intelligence, Renmin University of China, Beijing* 100872)
[5)](*Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Beijing* 100872)

Abstract Traditional information retrieval (IR) models such as BM25 and language models for IR (LMIR) are knowledge-driven approaches that primarily focus on the term frequency,

收稿日期:2022-09-27;在线发布日期:2023-04-25. 本文受到国家重点研发计划项目(2019YFE0198200)、国家自然科学基金项目(62276248)、北京高校卓越青年科学家计划项目(BJJWZYJH012019100020098)、中国人民大学"双一流"跨学科重大创新规划平台"智能社会治理跨学科交叉平台"的支持.**薄 琳**,硕士研究生,主要研究领域为信息检索、深度学习.E-mail:bolin20@ruc.edu.cn. **庞 亮**,博士,副研究员,主要研究领域为深度学习、文本挖掘. **张朝亮**,硕士,主要研究领域为信息检索、推荐系统. **王钊伟**,硕士,主要研究领域为信息检索、推荐系统. **董振华**,博士,主要研究领域为信息检索、反事实学习. **徐 君**(通信作者),博士,教授,主要研究领域为信息检索、互联网搜索.E-mail:junxu@ruc.edu.cn. **文继荣**,博士,教授,主要研究领域为互联网大数据管理、信息检索、文本挖掘.

document length, and inverse document frequency. These models derive the relevance score of a document based on a combination of these factors. Experts summarize a set of IR axioms that these models should follow. The axioms describe the characteristics that an ideal IR model should have. One such famous axiom is the TFC constraint, which stands for "Term Frequency Constraint". According to this constraint, a document that contains more occurrences of the query terms should be considered more relevant. This means that the model should prioritize documents that have a high frequency of search terms over those with a lower frequency. Recently, there has been a significant amount of research and development focused on data-driven approaches to IR, particularly using neural models. These models can learn ranking functions from labeled data and have been extensively studied and improved in recent years. Researchers have preliminarily explored whether axiomatic knowledge can improve the neural ranking model. One approach that has been explored is to generate augmented data that follows IR axioms and use these data to train the neural models. By incorporating these axioms, it is believed that the models can learn to better distinguish between relevant and irrelevant documents. However, it is worth noting that IR axioms only provide information on the relative relevance of two documents based on the difference in their matching signals. They do not provide a definitive determination on whether a document is relevant or not. To better incorporate IR axioms' comparison nature of matching signals and inspired by the data generation process in contrastive learning, we proposed a framework that enhances axiomatic knowledge in the neural ranking model via contrastive learning, named ACRank. ACRank generates documents guided by the IR axioms that have similar content but differ in other features. By presenting these pairs to the neural model and training it to extract matching signals between them based on axiomatic knowledge. The key differences in matching signals between the generated documents and the original documents are highlighted through optimizing contrastive loss. Through this approach, the model can learn the comparison nature of the matching signals mentioned by IR axioms. This allows the model to better understand IR axioms for determining relevance and to leverage this knowledge when ranking documents. In this way, the IR axiomatic knowledge is naturally transferred to the models. Overall, the proposed framework is a general framework that can be applied to different axioms. It is a promising way to enhance the effectiveness of neural ranking models by leveraging insights from IR axioms. By combining the strengths of knowledge-driven and data-driven approaches, we hope to develop more effective search systems for a wide range of applications. To test the effectiveness of the framework, we use the TFC constraint. Experimental results showed that the proposed framework was able to effectively improve the ranking model in comparison to BERT and others. The analysis showed that the framework was able to significantly improve the effectiveness of the model.

Keywords neural ranking model; information retrieval axiom; contrastive learning; knowledge driven; data driven

1 引 言

信息检索(Information Retrieval,IR)的主要目标是向用户提供满足其信息需求的文档[1]. 因此,IR领域最基础的问题是衡量查询和文档的相关度. 现有方法可分为两类:知识驱动的方法和数据驱动的方法. 知识驱动的方法有BM25、LMIR等,这类方法通过设计相关性评分函数计算文档相关性. 在函数中考虑精确匹配信号:词频(Term Frequency,TF)、文档长度(Document Length,DL)和逆文档频率(Inverse Document Frequency,IDF)

等. 数据驱动的方法以标注数据作为输入，通过设计模型学习输入和标签之间的关系建模查询和文档的相关性.

经验表明，知识驱动方法的性能与检索启发式规则的使用密切相关. 学者将启发式规则总结为形式化表达的IR公理约束：词频相关的启发式规则总结为词频约束(Term Frequency Constrains，TFCs)，文档长度相关的总结为长度归一化约束(Length Normalization Constraints，LNCs)[2]. 这些IR公理规则通过比较特定的匹配信号大小来判断相关性的强弱. 例如，词频相关规则通过比较文档中查询词个数这一匹配信号判断相关性的高低，查询词数量越多，匹配信号越强，文档越相关. 比较也更符合真实用户行为[3]. 截至目前，学者们从词频、文档长度、语义等方面总结了20余种规则. 实验表明，IR公理约束可用来分析并评价相关性评分函数[2].

随着数据驱动的神经排序模型不断发展，IR公理约束同样也被用于分析评价这类方法. 此外，IR公理知识能否指导神经排序模型训练，提升排序精度是一个亟待研究的问题，已有学者通过构造符合公理规则的数据训练模型，使模型掌握IR公理知识，提高模型效果[4]. 然而该方法直接学习生成的数据和给定的标签之间的映射关系，忽略了IR公理是通过比较匹配信号的强弱进而判断相关性高低、而非直接将文档转化为相关性标签这一特点. 此外，文献[5]利用公理进行预训练，通过将公理更改为判断同一篇文档不同查询间的相对关系，而后设计预训练方法使模型学得IR公理知识. 更改后的公理仍在比较特定匹配信号，该方法通过预训练利用大量数据使模型直接学习公理，也忽略了公理中比较匹配信号的性质.

为了学习比较匹配信号的过程，让数据驱动的神经排序模型从本质上建模IR公理规则的特性，受对比学习强调正负样本间差距的启发，查询q与其相关的文档匹配信号更强，与其不相关的文档匹配信号更弱[2]，正样本和负样本之间匹配信号应有一定差距，基于此本文提出框架ACRank(Axiom Enhanced Neural Ranking Model via Contrastive Learning). ACRank构造符合IR公理规则的数据，抽取原始数据和构造数据的匹配信号，通过对比学习拉开正负样本的匹配信号差异，从而让模型学习IR公理规则比较匹配信号的本质性质，进而让模型掌握IR公理知识. ACRank框架包括四个阶段：(1)增强数据生成：通过给定查询和文档，根据规则构造相较原文档更相关的正样本或更不相关的负样本；(2)神经网络编码：将查询和文档进行编码，得到每个单词的表达；(3)匹配信号抽取：获得每篇文章基于特定规则的匹配信号；(4)对比学习训练：强调正负样本间匹配信号的差异.

ACRank作为通用框架，建模IR公理比较匹配信号的性质，可应用于多种IR公理规则. 考虑到传统信息检索模型均将文本看作词袋并考虑完全匹配信息，词频规则为最重要的IR公理知识之一，基于此，本文基于词频规则实现ACRank. 根据词频规则比较匹配词语数量的特点和训练深度神经网络的要求，ACRank在阶段1对词频规则进行扩展实现数据增强[4]，在阶段2使用BERT[6]获得每个单词的表达，在阶段3选用与匹配词数量相关的soft-TF[7]值作为核心匹配信号，在阶段4实现对比学习[8]，强调正负样本之间匹配信号的差异.

本文在中文数据集“Sogou-QCL”[9]和“TianGong-ST”[10]及英文数据集“MS MARCO”[11]上分别进行了测试. 实验表明ACRank优于底层神经网络模型和已有利用IR规则的方法，证明了抽取匹配信号并利用对比学习强调正负样本匹配信号的差距能让数据驱动的神经排序模型更为本质的学习IR公理规则.

本文的主要贡献包括：

(1)提出了引入检索公理知识的神经排序模型框架ACRank，框架在IR公理的指导下构造数据，抽取匹配信号，并利用对比学习的方法拉开不同文档匹配信号间差异；

(2)结合传统检索模型中多将文本看作词袋的特点，选择词频规则进行实现. 实验结果表明，ACRank能够提升神经排序模型的排序指标；实验分析表明，ACRank框架通过抽取文档的匹配信号，并利用对比学习强调正负样本间差距，使模型学得IR公理知识，提升模型性能.

本文组织结构如下：本节讨论全文背景，第2节介绍本文的相关工作；第3节介绍ACRank框架的具体结构，包括整体框架、各阶段实现细节和框架优势；第4节介绍基于词频规则的实现方式，详细展开ACRank框架4个阶段的实现细节；第5节介绍实验，包括数据集、实验设置、实验结果与分析；第6节展开本文的结论与展望.

2 相关工作

信息检索领域最基础的问题是如何衡量查询和文档之间的相关程度[1]. 传统的IR模型，如BM25，LMIR[12-13]等利用专家构造的相关性评分函数衡量文档相关性. 已有工作发现，这些函数的性能与其满足TF、IDF、DL等IR启发式规则密切相关. 学者将这一系列启发式规则总结为可形式化表达的IR公理约束. 截至目前，已有二十多种相关性规则被提出，涵盖词频[2,14-15]、词频下界[16-17]、文档长度[2,18]、语义相似度等[19-20].

近年来数据驱动的方法成为信息检索领域的研究热点，这类方法通过给定训练数据和其相关性标签训练神经网络，排序学习（Learning to Rank，LTR）以手工提取特征作为输入，如RankSVM[21]、LambdaMart[22]，深度学习方法以查询和文档原文作为输入，自动学习特征，如DSSM[23]、KNRM[7]等.

最初规则之于信息检索，承担了“理论”角色，应用理论评价相关性评分函数. 随着机器学习的发展，数据驱动的方法逐渐兴盛，学者利用IR规则评价数据驱动的神经网络模型. Rennings等人[24]构建分析数据，分析四种神经网络模型是否满足四类IR规则. Câmara等人[25]同样构造分析数据，分析BERT[6]模型是否满足九条IR规则. Chen等人[26]通过构造数据，并利用梯度集成进行解释分析. 此外，Völsk等人[27]分析神经网络模型能够有良好表现是因为满足了哪些规则.

除了利用IR规则评价数据驱动的神经网络方法外，学者开始利用IR规则指导数据驱动的神经网络训练. Hagen等人[28]利用规则对任意检索模型排序结果的前k个文档进行重排. Rosset等人[29]利用IR公理规则，约束神经网络模型中的参数，构建符合规则的数据，添加正则损失，验证了在少量数据和特定规则下公理的有效性. Li等人[30]将用户判断相关性时的阅读方式总结为启发式规则，重新组织召回模型. Cheng等人[4]利用IR规则构造数据，而后直接利用增强数据训练神经网络模型，然而这种方式没有考虑IR规则比较匹配信号的特点. Chen等人[5]提出ARES方法，将IR公理应用于预训练任务. 信息检索公理通过给定查询后比较不同文档间的匹配信号判断相对的相关关系，该方法将公理修改为判断给定文档后比较特定匹配信号以得到查询间的相关关系. ARES包括三个阶段，分别是伪查询采样，偏好预测器构造和公理正则化预训练. 具体方法如下：在第一阶段通过从文档中采样更具代表性的伪查询，在第二阶段构造公理偏好预测器，为查询对生成公理偏好标签，利用四种方式构造有序的查询对，并为每个查询对按照修改后的公理抽取特征，将公理特征与弱标签联合使用以训练基于公理的二分类偏好预测器，在阶段三选择利用偏好预测器约束后的查询进行预训练，通过联合优化该相关性损失和遮盖词预测损失训练模型. 在微调阶段使用预训练后的模型. 该方法主要以修改后的公理选择出的查询作为输入，修改后的公理仍然在比较某些特定匹配信号，因此为更好的建模公理，本文从原始规则出发，在模型框架中考虑ARES框架未考虑的IR公理比较匹配信号的特点，选取端到端的训练方式，设计不同阶段以比较匹配信号.

为了使模型学得比较匹配信号的性质，正样本中的匹配信号强于负样本，受对比学习的启发，ACRank框架通过优化对比学习损失进行训练. 对比学习通过拉近相似样本，拉远不相似样本来学习表达[31]，已被广泛应用于图像领域[32-33]. 对于自然语言处理（Natural Language Processing，NLP）领域而言，学者们通过构造正负[34-35]样本，拉开正样本和负样本之间的距离，以训练模型有更好的语言表达. IR任务的主要目标是衡量相关关系[1]，对于某一查询，IR公理通过比较匹配信号的强弱，认为相关的文档应较不相关的文档匹配信号更强，为了使模型掌握IR公理知识，在训练时以拉开相关文档和不相关文档之间的匹配信号差距为训练任务. Gao等人[8]实现了IR领域中的对比学习方法，该文主要探索稠密向量检索. 现有利用语言模型的方法通常是将查询和文档编码为一个稠密向量，而后进行检索，如使用BERT的CLS向量编码查询和文档，但该文发现现有训练稠密向量时的注意力机制对于检索而言并非最优，因此该方法在预训练阶段调整模型结构，优化注意力机制，以适应稠密向量检索，在微调阶段使用对比学习方法，进一步训练检索向量，拉开正负样本间检索向量的差距. 为了使模型学得信息检索公理比较匹配信号的性质，本文参考了该文微调阶段对比学习的实现方式，以匹配信号为对比对象，拉开单个正样本和其余所有负样本之间的差距.

3 ACRank框架

本节首先介绍ACRank整体框架，而后介绍框

架中各阶段具体设计原因和实现方式及框架优势.

3.1　整体框架

ACRank框架的核心在于建模IR公理比较匹配信号的过程，框架共分四个阶段，阶段1结构如图1所示，阶段2至阶段4结构如图2所示. ACRank在训练时，以原始查询q和文档d输入阶段1，根据不同的生成规则，指导生成较d更相关的文档d^+或更不相关的文档d^-. 在阶段2，将原始数据和增强数据以查询q和文档d拼接的形式输入数据驱动的神经网络模型中进行编码，获得查询和文档交互后每个单词细粒度的表达$\mathbf{v}_t$. 在阶段3，将查询和文档交互后每个单词的表达$\mathbf{v}_t$输入匹配信号抽取模块，获得特定规则下的匹配信号$S(q,d)$. 最后在阶段4，将匹配信号$S(q,d)$输入对比学习损失函数中，通过优化该损失，强调不同文档间匹配信号的差异，使数据驱动的神经网络模型学到对应规则比较某一特定匹配信号知识.

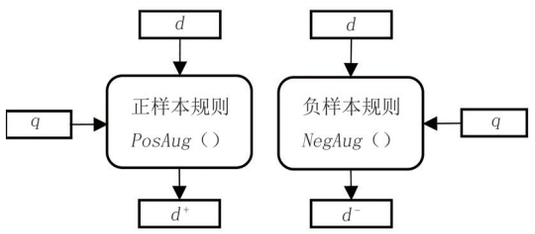

图1　ACRank框架阶段1实现方法

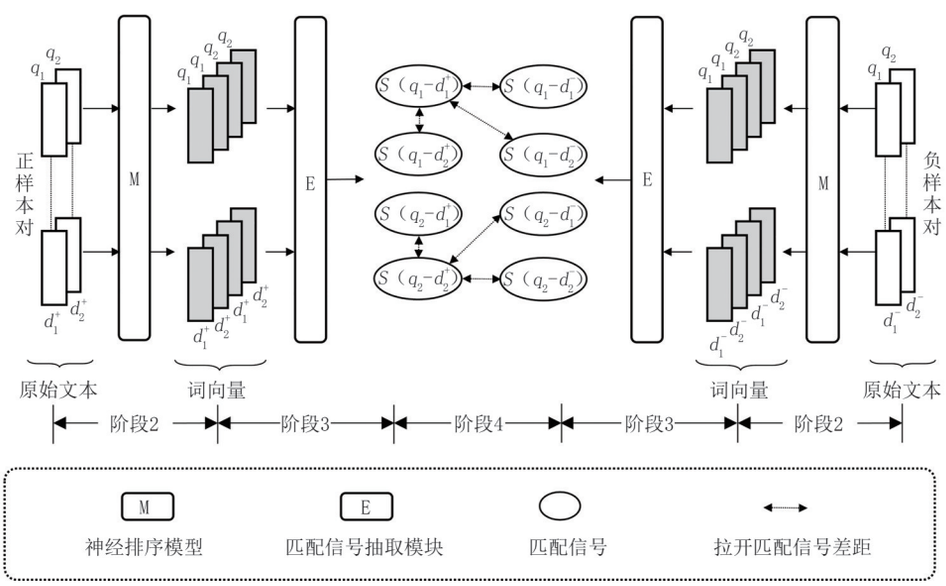

图2　ACRank框架阶段2至阶段4实现方法

ACRank在推理时，以原始测试数据进行输入，经过阶段2神经网络编码和阶段3匹配信号抽取，以匹配信号的大小进行排序，得到最终结果.

3.2　各阶段实现细节

本节将介绍ACRank框架每个阶段的设计思路及实现细节.

3.2.1　阶段1:增强数据生成

为更直接学到某一特定IR公理所蕴含的知识，ACRank首先在第一阶段构造数据. IR规则是严格数学形式表达的，难以直接构造完全符合定义的数据，因此构造尽可能满足规则直观要求的数据. 以原始相关的查询文档对作为输入，对文档施加扰动，输出较原始文档更相关或更不相关的构造文档.

以查询q和其相关文档d_{pos}作输入，根据生成正样本规则方法$PosAug()$或负样本规则方法$NegAug()$，对文档施加扰动，可分别生成比d_{pos}更相关的数据d_{pos}^+或更不相关的数据d_{pos}^-：

$$d_{pos}^+ \leftarrow PosAug\left(q, d_{pos}\right) \tag{1}$$

$$d_{pos}^- \leftarrow NegAug\left(q, d_{pos}\right) \tag{2}$$

3.2.2　阶段2:神经网络编码

IR规则偏向于规定词频、文档长度等细粒度文档性质，因此神经网络应尽可能保留词级别的信息：如文档长度、单词表达等，所以ACRank框架主要针对基于表达的神经排序模型. 以原始数据和构造数据作为输入，通过神经排序模型$M()$获得每个词的编码结果.

对于查询$q=\{t_1^q,\cdots,t_i^q,\cdots,t_n^q\}$和文档$d=\{t_1^d,\cdots,t_j^d,\cdots,t_m^d\}$，神经排序模型$M()$将$q$和$d$中每个词$t$编码为L维的向量$\mathbf{v}_t$：

$$\mathbf{v}_t = M(t) \tag{3}$$

3.2.3 阶段3:匹配信号抽取

IR规则通过比较特定的匹配信号来判断相关性的强弱,ACRank框架在第三阶段抽取该信号.根据不同的规则,该阶段有不同的实现方式.输入查询、文档编码后的结果,输出查询、文档对在该规则下的匹配信号.

匹配信号抽取模块$E()$将q和d中每个词经过编码后的词向量$\mathbf{v}_t$映射成针对特定规则的匹配信号$S(q,d)$:

$$S(q,d) = E(\mathbf{v}_t) \tag{4}$$

3.2.4 阶段4:对比学习训练

为建模IR规则比较匹配信号的性质,训练目标应使正样本的匹配信号强于负样本.受到对比学习的启发,ACRank框架使用对比学习损失作为优化目标.查询q和其使用规则后更相关的样本d^+匹配信号更强,反之使用规则后更不相关的样本d^-匹配信号更弱,d^-与其他标签为不相关的文档共同组成d_l^-.本文参考了Gao等人[8]的实现方式构造损失函数:

$$L = -\log \frac{\exp\left(S(q,d^+)\right)}{\exp\left(S(q,d^+)\right) + \sum_l \exp\left(S(q,d_l^-)\right)} \tag{5}$$

3.3 框架优势

IR公理规则重点在于比较匹配信号,通过信号的强弱确定文档的相关性高低.ACRank框架通过匹配信号抽取和对比学习训练两个阶段建模,相较于直接使用增强数据的方法更本质地建模了IR公理规则.同时ACRank作为通用框架,可应用于多种IR公理规则.

4 基于词频规则的ACRank实现

ACRank通过建模IR公理比较匹配信号的过程使框架学得公理知识,在实现中需考虑公理对查询、文档的限制条件和匹配信号的量化方式.传统模型通常将文本看作词袋,更关注词语的完全匹配,因此本文选择词频规则作为ACRank的实现.

词频规则主要包括TFC1、TFC2、TF-LNC等规则.他们的定义如下:使用d或d_i代表一篇文档,q代表查询,w或w_i代表一个查询词,$c(w,d)$代表词w在文档d中的出现次数,$|d|$代表文档d的长度.f代表检索方法,$f(d,q)$代表给定查询q和文档d后的评分.

(1) TFC1:假设q为只包含一个查询词w的查询,即$q=\{w\}$,假设文档d_1和文档d_2长度相同,即$|d_1|=|d_2|$.如果$c(w,d_1)>c(w,d_2)$,则有$f(d_1,q)>f(d_2,q)$.

(2) TFC2:假设q为只包含一个查询词w的查询,即$q=\{w\}$,假设文档d_1、文档d_2和文档d_3长度相同,即$|d_1|=|d_2|=|d_3|$,同时$c(w,d_1)>0$.如果$c(w,d_2)-c(w,d_1)=1$,$c(w,d_3)-c(w,d_2)=1$,则有$f(d_2,q)-f(d_1,q)>f(d_3,q)-f(d_2,q)$.

(3) TF-LNC:假设q为只包含一个查询词w的查询,即$q=\{w\}$,如果有$c(w,d_1)>c(w,d_2)$,同时$|d_1|=|d_2|+c(w,d_1)-c(w,d_2)$,则有$f(d_1,q)>f(d_2,q)$.

这些规则对查询q、文档d、完全匹配词$c(w,d)$有严格限制,难以完全按照规则生成数据.从直观上理解,TFC1规则认为有更多查询词的文档更相关,TFC2规则认为有更多区分性的查询词的文档更相关,TF-LNC规则控制了长度和词频之间的关系.本文根据Fang等人[2]关于各规则的直观论述,放宽词频规则对文档长度需完全一致、完全匹配词个数需满足特定等式关系的限制,保留规则的核心要素——比较完全匹配词的情况.

这些规则比较的匹配信号均与完全匹配词的个数相关,为了更好地适应神经排序模型,以soft-TF[7]值作为匹配信号.Soft-TF值以每个单词的表达作为输入,为了获得更好的单词表达,使用BERT[6]获得每个单词的表达.

4.1 阶段1:增强数据生成

词频规则认为若文档d_1是通过向d_2添加更多查询项来生成的,那么d_1相关性高于d_2,本文参考了Fang等人[14]和Rosset等人[29]的实践,实现两种增强方法:生成更相关文档的TFC-A规则,生成更不相关文档的TFC-D规则,具体如下:

TFC-A:给定原始查询q及其正样本d_{pos},在原始文档最前端以一定概率插入查询词,生成增强样本.生成的样本更相关,记为d_{pos}^+.

TFC-D:给定原始查询q及其正样本d_{pos},以一定概率删除文档中的查询词,生成增强样本.生成的样本更不相关,记为d_{pos}^-.

表1展示了不同概率下TFC-A规则和TFC-D

表 1　不同概率下TFC-A规则和TFC-D规则对文档标题的扰动情况

查询	如何卸载瑞星杀毒软件
文档标题原文	关于瑞星杀毒软件如何卸载的说明
TFC-A(10%)	**卸载**关于瑞星杀毒软件如何卸载的说明
TFC-D(100%)	关于~~瑞星杀毒软件~~如何~~卸载~~的说明

规则对文档标题的扰动情况. 其中"()"中百分数代表插入或删除的概率,加粗代表插入该词,删除线代表删除该词. 将在后续实验中探索只使用TFC-A规则、TFC-D规则和联合使用两种规则的效果.

4.2　阶段2:神经网络编码

针对词频规则的ACRank框架在神经网络编码阶段需要获得每个单词的具体表达,为了获得更高质量的表达,本文使用目前广泛应用的BERT[6]作为框架实现. 实现情况如图3所示,将查询q文档d与特殊字符[CLS]和[SEP]拼接得到最终输入:$[\mathrm{CLS}], t_1^q, \cdots t_i^q, \cdots, t_n^q, [\mathrm{SEP}], t_1^d, \cdots t_j^d, \cdots, t_m^d, [\mathrm{SEP}]$.

最大输入长度按照BERT的要求进行处理,输出每个字符t最后一层的表达:

$$\mathbf{v}_t = \mathrm{BERT}(t) \tag{6}$$

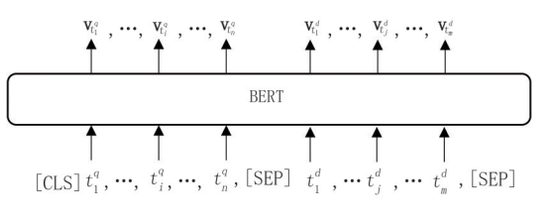

图3　基于词频规则的阶段3实现方法

4.3　阶段3:匹配信号抽取

针对词频规则的ACRank框架在匹配信号抽取阶段需要获得与完全匹配词个数相关的匹配信号. 原始词频规则考虑完全匹配的情况,属于词袋模型,ACRank框架在阶段2使用神经网络模型自动学习单词的表达,为充分挖掘语义信息,本文使用soft-TF的变化值为匹配信号,参考Xiong等人[7]的实现方式,首先计算得到矩阵$\boldsymbol{T}$,$\boldsymbol{T}_{ij} = \cos\left(\mathbf{v}_{t_i^q}, \mathbf{v}_{t_j^d}\right)$,而后利用核函数$\phi()$生成特征,并通过多层前馈神经网络(Multi Layer Perception,MLP)计算最终匹配信号$S(q,d)$:

$$S(q,d) = \mathrm{MLP}\left(\phi(\boldsymbol{T})\right) \tag{7}$$

TFC-A规则生成的文档匹配信号应强于原文的匹配信号,TFC-D规则生成的文档匹配信号应弱于原文的匹配信号.

4.4　阶段4:对比学习训练

在对比学习训练阶段,生成的数据在损失函数中的位置依据规则的不同而不同,后续将探索单独使用TFC-A规则、TFC-D规则,和联合使用两种规则的效果.

对于TFC-A规则,生成的数据d_{pos}^+相较于其原始文档d_{pos}更相关,因此$d_{pos} \in d_l^-$,有损失函数:

$$L = -\log \frac{\exp\left(S\left(q, d_{pos}^+\right)\right)}{\exp\left(S\left(q, d_{pos}^+\right)\right) + \sum_l \exp\left(S\left(q, d_l^-\right)\right)} \tag{8}$$

对于TFC-D规则,生成的数据d_{pos}^-相较于其原始文档d_{pos}更不相关,因此$d_{pos}^- \in d_l^-$,有损失函数:

$$L = -\log \frac{\exp\left(S\left(q, d_{pos}\right)\right)}{\exp\left(S\left(q, d_{pos}\right)\right) + \sum_l \exp\left(S\left(q, d_l^-\right)\right)} \tag{9}$$

对于联合使用TFC-A规则和TFC-D规则对框架进行训练的情况,使用TFC-A规则生成的数据d_{pos}^+相较于其原始文档d_{pos}更相关,使用TFC-D规则生成的数据d_{pos}^-相较于其原始文档d_{pos}更不相关$d_{pos}^- \in d_l^-$,损失函数表达式同(8).

在训练过程中通过阶段3得到匹配信号,按照使用规则的不同分别带入不同的公式计算对比学习损失,通过优化该损失,拉开词频更高的文档和词频更低的文档之间匹配信号的差距,进而使模型学得词频相关规则.

5　实　验

5.1　数据集

本文在三个公开数据集上进行实验,分别是中文数据集Sogou-QCL[9]和TianGong-ST[10]及英文数据集MS MARCO[11],各数据集大小如表2所示.

表 2　不同数据集训练和测试数据量大小

数据集	训练集		测试集	
	查询数量	文档数量	查询数量	文档数量
Sogou-QCL	310 380	3 495 134	1 570	38 511
TianGong-ST	40 596	288 181	610	5 547
MS MARCO	367 013	3 213 835	5 193	519 300

(1)　Sogou-QCL[9]

取样自商业搜索引擎搜狗的查询日志,共包括537 366条查询,超过9百万的中文网页. 每条记录包括查询文本和一系列文档,每个文档包括了标题、

全文内容、五种基于点击模型(TCM、DBN、PSCM、TACM以及UBM)自动生成的相关性标签等,该数据集同时包括一个4级人工标注的子集.

本文参考Li等人[36]的实现方式,使用PSCM的相关性标签生成训练数据,去除重复查询和空文档后,训练集包括310 380条查询,3 495 134个文档.测试集包括1 570条查询,38 511个文档.其中,人工标注的测试数据集记为"Sogou-QCL(HUMAN)",PSCM点击模型标注的测试数据集记为"Sogou-QCL(PSCM)".

(2) TianGong-ST[10]

取样自搜狗搜索18天的日志,包含147 155个搜索会话,40 596条不重复查询,297 597个网页正文.每条会话记录包括用户查询及搜索引擎返回的前10个文档,每个文档包括了标题、正文、六种基于点击模型(TCM、DBN、PSCM、THCM、TACM以及UBM)自动生成的相关性标签等.同时包括2 000个会话内容的5级人工标注子集.

训练数据的预处理同Sogou-QCL,处理后训练集包括40 596条查询,288 181个文档.测试集包括610条查询,5 547个文档.测试标签由人工标注,记做"Tiangong-ST(HUMAN)".

(3) MS MARCO[11]

取样自必应搜索日志,包含3 213 835个文档.训练数据包括367 013个查询,每条查询含一条人工标注的相关文档.所有实验均采用官方提供的Top100文件进行负样本采样.本文关注文档重排任务,使用官方提供的验证集做测试,共5 193条查询,和每条查询利用BM25召回的前100个文档.

5.2 对比模型和本文提出的方法

本文选用两类对比方法:1、神经网络模型;2、已有利用IR规则的方法.

(1) 神经网络模型

KNRM[7]:是一种基于交互的神经排序模型.首先生成查询和文档的交互矩阵,而后引入核函数,在交互矩阵转化的直方图上通过kernel-pooling的方式提取不同粒度的匹配信号.

Conv-KNRM[37]:是一种基于卷积核的神经排序模型.在KNRM的基础上使用卷积来融合周围单词的上下文信息,而后利用排序学习层计算匹配分数.

BERT[6]:是一个预训练语言模型.该模型利用遮盖词预测(Masked Language Modeling,MLM)和下一句预测(Next Sentence Prediction,NSP)任务进行预训练.

Condenser[8]:是一种针对检索的预训练方法.该方法通过在预训练阶段调整Transformer内部结构以适应稠密向量搜索,在微调阶段使用对比学习训练检索任务.

(2) 已有利用IR规则的方法

Axiomatic Perturbation method(AP)[4]是一种直接利用IR规则生成数据并训练的方法.通过对数据集施加扰动,构建满足公理性质的数据,利用生成的数据训练已有神经网络模型.

ARES[5]是一种将IR公理应用于预训练任务中的方法.该方法对公理进行了更改,衡量同一文档不同查询间的关系,修改后的公理仍通过比较特定匹配信号进行训练.模型设计三个阶段五类公理进行训练①从文档中采样伪查询;②根据改造后针对查询的公理构建偏好预测器;③利用偏好预测器的结果对模型进行正则化训练.

根据使用规则的不同,本文提出的ACRank有三个版本:使用TFC-A规则的相关实验记作ACRank(TFC-A);使用TFC-D规则的相关实验记作ACRank(TFC-D);联合使用TFC-A规则和TFC-D规则的相关实验记作ACRank(TFC-A&D).同一数据集下的所有实验所使用的数据输入长度、负例选择均相同.

5.3 实现细节及评价指标

在数据生成阶段,TFC相关规则需要获取查询和文档中完全匹配的词语,中文数据使用jieba进行分词,使用哈工大发布的停用词表去除停用词,英文数据使用nltk中停用词表去除停用词,而后根据TFC-A规则及TFC-D规则生成数据.

本文所有实验均使用PyTorch实现,对于对比模型,KNRM、Conv-KNRM使用OpenMatch[38]进行实现,其中英文词向量使用300维GloVe[39],中文使用300维通过百度百科训练的词向量[40].BERT[6]使用pytorch-pretrained-bert进行实现.Condenser使用其公开的checkpoint进行实验.Axiomatic Perturbation method使用本文所提出的TFC-A和TFC-D规则与神经排序模型BERT进行重新实现,其中使用TFC-A规则的相关实验记作AP(TFC-A),使用TFC-D规则的相关实验记作AP(TFC-D),联合使用TFC-A规则和TFC-D规则的相关实验记作AP(TFC-A&D).ARES使用公开代码与checkpoint进行实验.Condenser与ARES仅有基于英文的预训练开源checkpoint,因此相关对比实验将

在MS MARCO上进行. 所有针对MS MARCO的实验均从官方提供的top 100文件中进行负样本的采样. 基于词频规则的ACRank框架第2阶段以BERT作为编码器，BERT对输入长度存在限制，在实现中保证了原始词频规则对文档长度的限制. ACRank各数据集实验批次大小∈[16，32，64]，学习率取值范围为[2E-5，2E-3]，实验中我们使用Adam优化器进行参数更新.

在实验中，我们使用NDCG@1、NDCG@10、MAP分别作为评价指标.

5.4　实验结果

英文数据集实验结果如表3所示，中文数据集实验结果如表4所示. 表3和表4均分为三部分，分别展示神经网络模型结果、已有利用IR规则的方法及本文提出的ACRank框架结果. 其中TFC-A添加概率、TFC-D删除概率在不同数据集下均保持一致. 不同概率的影响将在5.5.2节中展开分析. 通过实验，可以得到以下结论：

表3　ACRank及对比模型实验结果(英文数据集)

类型	方法	MS MARCO		
		NDCG@1	NDCG@10	MAP
神经网络模型方法	KNRM	0.128*†‡	0.251*†‡	0.174*†‡
	Conv-KNRM	0.167*†‡	0.302*†‡	0.219*†‡
	BERT	0.200*†‡	0.376*†‡	0.321*†‡
	Condenser	0.210*†‡	0.382*†‡	0.323*†‡
已有利用IR规则的方法	AP(TFC-A)	0.179*	0.358*	0.302*
	AP(TFC-D)	0.193†	0.369†	0.315†
	AP(TFC-A&D)	0.187‡	0.360‡	0.308‡
	ARES	0.221	**0.417**	**0.363**
ACRank框架方法	ACRank(TFC-A)	0.226	0.414	0.357
	ACRank(TFC-D)	0.227	0.413	0.356
	ACRank(TFC-A&D)	**0.229**	0.413	0.355

注：本表展示英文数据集下神经网络方法、已有利用IR规则的方法和ACRank框架方法的实验结果. 其中TFC－A和TFC－D为本文所提出的数据增强方法. *表示该方法显著差于ACRank(TFC－A)方法(t检验，$p<0.05$)，“†”表示该方法显著差于ACRank(TFC－D)方法(t检验，$p<0.05$)，“‡”表示该方法显著差于ACRank(TFC－A&D)方法(t检验，$p<0.05$).

表4　ACRank及对比模型实验结果(中文数据集)

类型	方法	Sogou-QCL(HUMAN)			Sogou-QCL(PSCM)			Tiangong-ST(HUMAN)		
		NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP
神经网络模型方法	KNRM	0.762*†‡	0.789*†‡	0.910*†‡	0.320*†‡	0.432*†‡	0.433*†‡	0.664*†‡	0.852*†‡	0.841*†‡
	Conv-KNRM	0.778*†‡	0.793*†‡	0.913*†‡	0.372*†‡	0.458*†‡	0.455*†‡	0.666*†‡	0.856*†‡	0.854*†‡
	BERT	0.788*†‡	0.811*†‡	0.919*†‡	0.373*†‡	0.465*†‡	0.462*†‡	0.674*†‡	0.861*†‡	0.850*†‡
已有利用IR规则的方法	AP(TFC-A)	0.801*	0.816*	0.924*	0.420*	0.510*	0.500*	0.689*	0.864*	0.852*
	AP(TFC-D)	0.798†	0.827†	0.937†	0.448†	0.520†	0.506†	0.697†	0.869†	0.861†
	AP(TFC-A&D)	0.804‡	0.833‡	0.904‡	0.427‡	0.502‡	0.491‡	0.709‡	0.872‡	0.862‡
ACRank框架方法	ACRank(TFC-A)	0.834	0.848	0.943	**0.535**	0.586	0.557	0.736	0.880	0.864
	ACRank(TFC-D)	**0.851**	**0.862**	**0.952**	0.527	**0.589**	**0.558**	**0.739**	0.882	0.868
	ACRank(TFC-A&D)	0.843	0.853	0.945	0.516	0.579	0.551	**0.739**	**0.886**	**0.885**

注：本表展示中文数据集下神经网络方法、已有利用IR规则的方法和ACRank框架方法的实验结果. 其中TFC-A和TFC-D为本文所提出的数据增强方法. *表示该方法显著差于ACRank(TFC-A)方法(t检验，$p<0.05$)，“†”表示该方法显著差于ACRank(TFC-D)方法(t检验，$p<0.05$)，“‡”表示该方法显著差于ACRank(TFC-A&D)方法(t检验，$p<0.05$).

（1）　对比使用神经网络模型的各个方法

中文数据集与英文数据集中均有预训练语言模型BERT的效果整体优于Conv-KNRM，Conv-KNRM整体优于KNRM，说明了查询与文档交互越充分，利用文本信息越充分，所能带来的效果越好. 英文数据集中Condenser的效果优于BERT及其他方法，说明预训练阶段针对检索任务特性更改Transformer内部结构更适于稠密向量搜索.

（2）　对比已有利用IR规则的各个方法

在中文数据集上，AP方法中利用TFC-D的方

法大部分情况下优于利用TFC-A规则的方法，联合使用两种规则的方法大部分情况下优于利用TFC-A规则的方法．在英文数据集上，ARES方法能超过各类AP方法，说明了利用信息检索公理进行预训练比直接使用增强数据更为有效．

（3） 对比基于词频规则的ACRank框架各个方法

不同数据集下不同规则的表现按照数据集的不同具有一定规律：大部分情况下，TFC-D规则优于TFC-A规则．对于不同数据集，Sogou-QCL绝大多数情况下TFC-D效果最优；Tiangong-ST联合使用两种增强样本效果最优；MS MARCO中NDCG@1指标下联合使用两种规则的效果最优，其他指标下TFC-A规则最优．说明了对于中文数据，向文本中添加查询词的方法在经过BERT编码后对语义的影响小于删除查询词，导致匹配信号变化幅度较弱，因而效果最弱．不同数据集下不同规则的实验结果间的性能差距不一，与数据集大小、所使用标签情况均无明显关联．

（4） 对比已有利用IR规则的方法与使用神经网络模型的方法

中文数据集上AP方法较神经网络模型的三种方法均能带来显著提升，证明了直接利用IR规则生成数据，便可使神经网络模型掌握一定的IR公理知识，验证了Cheng等人[4]的结论．在英文数据集上，AP方法优于KNRM与Conv-KNRM，劣于BERT，这主要是因为本文使用了MS MARCO全量数据进行训练，而Cheng等人[4]使用了20 000条进行训练，根据Rosset等人[29]的结论，MS MARCO在原始数据量逐渐增加的情况下加入规则的效果减弱，导致AP方法在该数据集上表现并不理想．利用公理进行预训练的ARES方法明显优于AP方法和各类神经网络方法，证明了在预训练阶段让模型学得IR公理比直接使用数据增强更为有效．

（5） 对比ACRank框架方法与使用神经网络模型的方法

对于ACRank框架方法，在中文数据集上，分别使用TFC-A规则、TFC-D规则和联合使用两种规则相较于三种神经网络模型均有显著提升．说明了ACRank框架利用增强数据抽取并比较匹配信号的有效性．

在英文数据集上，可对比ACRank方法与Condenser方法：ACRank三种实现结果均显著优于Condenser，说明了本文对比匹配信号的有效性．ACRank与Condenser方法均采用了对比学习损失函数，除此之外，本文与Condenser的训练方式与实现思路均不同：Condenser使用预训练-微调框架，在预训练阶段修改Transformer结构，并在微调阶段使用对比学习对比基于Transformer编码的上下文语义，而ACRank为端到端的训练方法，在阶段4中利用对比学习对比匹配信号．通过实验结果可以看到不论对比哪一种信号，相较于BERT都能带来效果的提升，说明了对比学习方法的有效性；ACRank最终结果优于Condenser，说明了对比匹配信号能够使模型学习IR公理知识并带来效果提升．

（6） 对比ACRank框架方法与已有利用IR规则的方法

在中英文数据集上，均有在相同规则下使用ACRank方法效果显著优于AP方法，说明了只使用增强数据的有限性，抽取并比较匹配信号能够使模型更好地学习IR公理知识．

在英文数据集上，可对比ACRank方法与ARES方法：ACRank框架方法最优情况与ARES实验效果在NDCG@1指标上有大幅提升，其余两个指标略弱，这主要是因为ARES方法利用了五类公理规则，本文只针对词频一种规则进行实现，说明了基于IR公理的预训练方法和本文建模比较匹配信号的方法都能有效使模型学得IR公理知识．ACRank与ARES都为使用IR公理指导神经网络训练的方法，但二者实现方式上存在诸多不同：①训练方法不同，ARES为针对信息检索的预训练方法，可应用于多种任务更为灵活，但需要大量数据进行训练，本文直接从数据出发关注下游任务，训练代价相对较小；②使用公理方式不同，ARES对公理进行了修改，衡量同一文档更适合哪一查询，修改后的公理也通过比较特定匹配信号判断查询间相对的相关关系，本文从公理原文出发，建模公理比较匹配信号的性质．可以看到，两种方法各有优劣，分别从不同角度在神经排序模型中引入IR公理知识．

5.5 实验分析

5.5.1 抽取匹配信号的影响

本文提出的ACRank框架是按照IR公理的特性进行构建，核心为阶段3匹配信号抽取和阶段4对比学习训练．本节分别讨论这两个阶段对框架最终效果的影响，将分别从使用原始数据和使用增强数据两个维度进行实验．

表5为直接使用原始数据进行实验的结果，表

表5　原始数据消融实验结果

方法	Sogou-QCL(HUMAN)			Sogou-QCL(PSCM)			Tiangong-ST(HUMAN)			MS MARCO		
	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP
BERT	0.788	0.811	0.919	0.373	0.465	0.462	0.674	0.861	0.850	0.200	0.376	0.321
w/ E()	0.815	0.831	0.932	0.412	0.488	0.480	0.657	0.866	**0.898**	0.195	0.377	0.322
w/ L()	0.814	0.833	0.935	0.438	0.512	0.499	0.699	0.866	0.846	**0.230**	0.407	0.352
w/ E()&L()	**0.827***	**0.838***	**0.936**	**0.449***	**0.526***	**0.511***	**0.730***	**0.880***	0.873	0.224	**0.412***	**0.353***

注：最优情况加粗表示. *表示“w/E()&L()”方法显著优于其他方法(t检验, $p<0.05$).

表6　增强数据消融实验结果

方法	Sogou-QCL(HUMAN)			Sogou-QCL(PSCM)			Tiangong-ST(HUMAN)			MS MARCO		
	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP	NDCG@1	NDCG@10	MAP
TFC-A												
BERT	0.801	0.816	0.924	0.420	0.510	0.500	0.689	0.864	0.852	0.179	0.358	0.302
w/ E()	0.828	0.842	0.940	0.407	0.496	0.488	0.653	0.867	**0.885**	0.196	0.374	0.319
w/ L()	0.813	0.836	0.935	0.483	0.562	0.539	0.692	0.868	0.853	0.224	0.407	0.351
ACRank	**0.834***	**0.848***	**0.943***	**0.535***	**0.586***	**0.557***	**0.736***	**0.880***	0.864	**0.226***	**0.414***	**0.357***
TFC-D												
BERT	0.798	0.827	0.937	0.448	0.520	0.506	0.697	0.869	0.861	0.193	0.369	0.315
w/ E()	0.831	0.851	0.944	0.432	0.512	0.498	0.670	0.869	**0.897**	0.195	0.372	0.318
w/ L()	0.836	0.850	0.947	0.505	0.567	0.544	0.736	0.877	0.858	0.225	0.412	0.354
ACRank	**0.851**†	**0.862**†	**0.952**†	**0.527**†	**0.589**†	**0.558**†	**0.739**†	**0.882**†	0.868	**0.227**†	**0.413**†	**0.356**†
TFC-A&D												
BERT	0.804	0.833	0.940	0.427	0.502	0.491	0.709	0.872	0.862	0.187	0.360	0.308
w/ E()	0.820	0.846	0.939	0.416	0.499	0.488	0.714	0.878	0.868	0.210	0.382	0.328
w/ L()	0.829	0.852	**0.947**	0.504	0.567	0.543	0.727	0.879	0.870	0.218	0.403	0.347
ACRank	**0.843**‡	**0.853**	0.945	**0.516**‡	**0.579**‡	**0.551**‡	**0.739**‡	**0.886**‡	**0.885**‡	**0.229**†	**0.413**†	**0.355**†

注：TFC-A、TFC-D、TFC-A&D最优情况加粗表示，整体最好情况以下划线表示. *表示TFC-A情况下“w/E()&L()”方法显著优于其他方法(t检验, $p<0.05$),“†”表示TFC-D情况下“w/E()&L()”方法显著优于其他方法(t检验, $p<0.05$),“‡”表示TFC-A&D情况下“w/E()&L()”方法显著优于其他方法(t检验, $p<0.05$).

6为分别使用TFC-A规则、TFC-D规则以及联合使用两个规则生成数据的实验结果. 其中“w/E()”代表直接抽取匹配信号，不利用对比学习损失函数而利用与基线BERT相同的损失函数进行训练；“w/L()”代表使用BERT的[CLS]表达经过前馈神经网络后的结果作为匹配信号，并用对比学习强调正负样本差异；“w/E()&L()”代表抽取匹配信号，并用对比学习强调正负样本差异，“ACRank”代表本文方法.

通过分别比较表5中基线BERT结果与“w/E()”、“w/L()”的结果可以看到，分别抽取匹配信号，或利用对比学习强调正负样本间差异都能在原始数据上带来效果提升. “w/E()”和“w/L()”在训练上损失函数不同、输入损失函数的量也不同，二者并不可直接对比. 分别对比“w/E()&L()”和“w/E()”、“w/L()”可发现“w/E()&L()”带来了明显的效果提升，说明了仅抽取或仅对比的有限性，验证了在深度排序模型中通过抽取匹配信号、比较该信号来建模IR公理特性的有效性.

分别对比表6中TFC-A下四种情况、TFC-D下四种情况与TFC-A&D联合使用的四种情况，有和表5类似的现象：“w/E()”、“w/L()”的结果相较于BERT都有一定提升，“w/E()&L()”较“w/E()”、“w/L()”有更明显的效果提升. TFC-A下四种实验情况和TFC-D下相同实验设置的四种情况可看到，TFC-D规则的效果普遍优于TFC-A，删除相关词语构造负样本的效果优于添加词语构造正样本的效果，说明语义上的缺失带来的匹配信号变化更能使模型学到IR公理规则. TFC-A&D下四种实验情况大部分略优于TFC-A，差于TFC-D. 理论上，经TFC-A规则指导生成的数据较原始数据，应比在TFC-D规则指导下生成的数据有更强的匹配信号，对比学习所能学到的差异理应更明显. 但基于词频规则的实现中，以BERT进行编码，使用原始

数据作为输入.TFC-D规则指导生成的数据删除文本中的查询词,对语义的影响较大,TFC-A规则指导生成的数据向文本中添加查询词,对语义同样存在影响.当正负例均为构造数据时,两者均为非可读语言,经过BERT编码的结果可能存在一定误差,使得匹配信号抽取的结果存在波动,没有达到最优效果.

此外,分别将表6下TFC-A的情况与表5对比,可发现大部分情况下使用增强数据的效果优于使用原始数据的效果;将表6下TFC-D的情况与表5对比,可发现除个别情况外,使用增强数据的效果整体优于使用原始数据的效果;将表6下TFC-A&D联合使用的情况与表5对比,有同TFC-D相同的实验结果;说明了使用规则增强数据,而后抽取匹配信号并进行比较可以使模型掌握IR公理知识,并且TFC-D规则优于联合使用TFC-A&D规则,优于TFC-A规则.

5.5.2 TFC-A规则增加概率及TFC-D删除概率的影响

本文通过抽取匹配信号并比较匹配信号使神经排序模型学习IR公理知识,本节讨论在数据增强时,不同增加或删除的概率对最终结果的影响.

如图4所示,分别展示了Sogou-QCL数据集用人工标注的测试数据"Sogou-QCL(HUMAN)"和用PSCM点击模型标注的测试数据"Sogou-QCL(PSCM)"在TFC-A规则不同增加概率和TFC-D规则不同删除概率NDCG@1、NDCG@5、NDCG@10的变化情况.从图中可以看出,对于Sogou-QCL(HUMAN),不论TFC-A还是TFC-D,各指标整体呈现先波动上升后下降的趋势.对于Sogou-QCL(PSCM),不论TFC-A还是TFC-D,各指标整体有微弱上升的现象.这说明不同构造规则删除或添加的概率对模型的结果会产生影响.

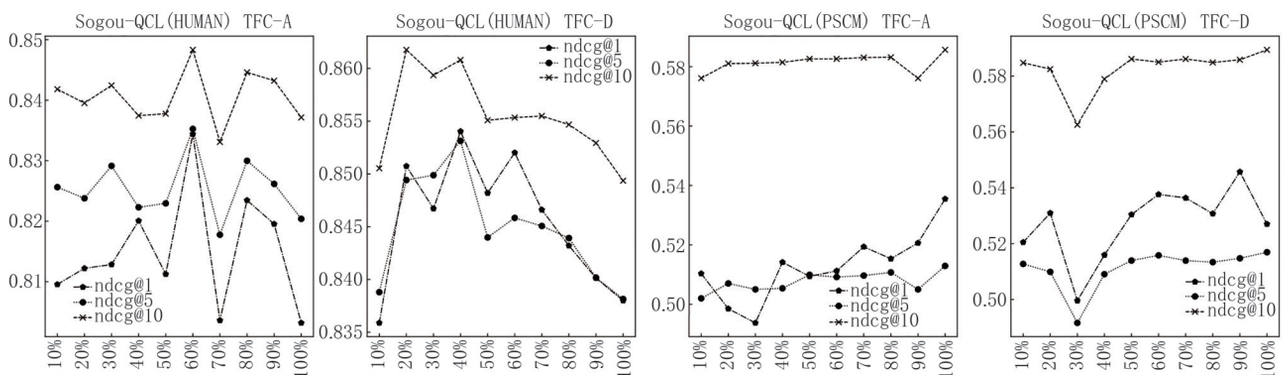

图4 不同增加概率及删除概率对结果的影响

对于Sogou-QCL(HUMAN),由于训练数据利用PSCM作相关性标签,训练和测试分布不一致,导致效果随着删除或添加的概率提高而逐渐下降;对于Sogou-QCL(PSCM),训练和测试分布一致,随着删除或添加的概率提高有微弱的上升趋势,说明了删除或添加的词概率越大,匹配信号的变化越明显,模型更能学到IR公理比较匹配信号的性质.

6 结论与展望

本文提出了一种使神经排序模型能够学习IR公理知识的对比学习框架ACRank,其通过构造数据,而后抽取并比较匹配信号的差异来判断不同文档间的相关关系,并基于词频规则对ACRank进行了具体实现.基于两个中文数据集和一个英文数据集上的实验结果与相关分析表明,ACRank相较于基线方法有显著的排序精度提升,验证了通过比较匹配信号,IR公理知识能够指导数据驱动的神经排序模型获得进一步的效果提升.

在下一步工作中,将完善基于词频规则的验证实验,并将尝试将其他IR公理规则如长度规则、语义规则应用于ACRank并进行相关实验.

致 谢 本文受到国家重点研发计划项目(2019YFE0198200)、国家自然科学基金项目(62276248)、北京市卓越科学家计划项目(BJJWZYJH012019100020098)、中国人民大学"双一流"跨学科重大创新规划平台"智能社会治理跨学

科交叉平台”的支持.

参 考 文 献

[1] Fan Y, Xie X, Cai Y, et al. Pre-training methods in information retrieval. Foundations and Trends in Information Retrieval, 2022, 16(3): 178-317

[2] Fang H, Tao T, Zhai C X. A formal study of information retrieval heuristics//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 2004: 49-56

[3] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback //Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 2005: 154-161

[4] Cheng Z, Fang H. Utilizing axiomatic perturbations to guide neural ranking models//Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, 2020: 153-156

[5] Jia Chen, Liu Yiqun, Yan Fang, Mao Jiaxin, Hui Fang, Yang Shenghao, Xie Xiaohui, Min Zhang, and Ma Shaoping. Axiomatically regularized pre-training for Ad hoc search// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 2022:1524-1534

[6] Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, USA, 2019: 4171-4186

[7] Xiong C, Dai Z, Callan J, et al. End-to-end neural ad-hoc ranking with kernel pooling//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, 2017: 55-64

[8] Gao L, Callan J. Condenser: a Pre-training architecture for dense retrieval//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 981-993

[9] Zheng Y, Fan Z, Liu Y, et al. Sogou-qcl: A new dataset with click relevance label//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor, USA, 2018: 1117-1120

[10] Chen J, Mao J, Liu Y, et al. TianGong-ST: a new dataset with large-scale refined real-world web search sessions// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 2485-2488

[11] NGUYEN T, ROSENBERG M, SONG X, et al. MS MARCO: A human generated machine reading comprehension dataset//Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-Located with the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016, 1773

[12] Robertson S E, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval// Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval. Dublin, Ireland, 1994: 232-241

[13] Croft W B, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice. New York, USA: Pearson Education, 2009

[14] Fang H, Tao T, Zhai C. Diagnostic evaluation of information retrieval models. ACM Transactions on Information Systems (TOIS), 2011, 29(2): 1-42

[15] Fang H, Zhai C X. An exploration of axiomatic approaches to information retrieval//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005: 480-487

[16] Lv Y, Zhai C X. Lower-bounding term frequency normalization//Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 2011: 7-16

[17] Lv Y, Zhai C X. A log-logistic model-based interpretation of TF normalization of BM25//European Conference on Information Retrieval, Berlin, Germany, 2012: 244-255

[18] Cummins R, O'Riordan C. A constraint to automatically regulate document-length normalisation//Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, USA, 2012: 2443-2446

[19] Fang H, Zhai C X. Semantic term matching in axiomatic approaches to information retrieval//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA, 2006: 115-122

[20] Fang H. A re-examination of query expansion using lexical resources//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus, USA, 2008: 139-147

[21] Herbrich R., Graepel T. and Obermayer K. Support vector learning for ordinal regression. 1999 9th International Conference on Artificial Neural Networks. Edinburgh, UK, 1999:97-102

[22] Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions// Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2006:193-200

[23] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, USA, 2013: 2333-2338

[24] Rennings D, Moraes F, Hauff C. An axiomatic approach to diagnosing neural IR models//European Conference on Information Retrieval, Cologne, Germany, 2019: 489-503

[25] Câmara A, Hauff C. Diagnosing BERT with retrieval

heuristics//European Conference on Information Retrieval. Lisbon, Portugal, 2020: 605-618

[26] Chen L, Lan Y, Pang L, et al. Toward the understanding of deep text matching models for information retrieval. arXiv preprint arXiv:2108.07081, 2021

[27] Völske M, Bondarenko A, Fröbe M, et al. Towards axiomatic explanations for neural ranking models//Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, 2021: 13-22

[28] Hagen M, Völske M, Göring S, et al. Axiomatic result re-ranking//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis, USA, 2016: 721-730

[29] Rosset C, Mitra B, Xiong C, et al. An axiomatic approach to regularizing neural ranking models//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris, France, 2019: 981-984

[30] Li X, Mao J, Wang C, et al. Teach machine how to read: reading behavior inspired relevance estimation//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris, France, 2019: 795-804

[31] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 1735-1742

[32] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9729-9738

[33] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations//International Conference on Machine Learning, 2020: 1597-1607

[34] Gao T, Yao X, Chen D. SimCSE: Simple contrastive learning of sentence embeddings//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 6894-6910

[35] Wang D, Ding N, Li P, et al. CLINE: Contrastive learning with semantic negative examples for natural language understanding//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2332-2342

[36] Li X, Mao J, Ma W, et al. A cooperative neural information retrieval pipeline with knowledge enhanced automatic query reformulation//Proceedings of the 15th ACM International Conference on Web Search and Data Mining, 2022: 553-561

[37] Dai Z, Xiong C, Callan J, et al. Convolutional neural networks for soft-matching n-grams in ad-hoc search//Proceedings of the 11th ACM International Conference on Web Search and Data Mining. Marina Del Rey, USA, 2018: 126-134

[38] Liu Z, Zhang K, Xiong C, et al. Openmatch: An open source library for neu-ir research//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 2531-2535

[39] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[40] Qiu Y, Li H, Li S, et al. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Changsha, China, 2018: 209-221

BO Lin, M. S. candidate. Her main research interests include information retrieval and deep learning.

PANG Liang, Ph. D., associate professor. His main research interests include deep learning and text mining.

ZHANG Chao-Liang, M. S. His research interest covers information retrieval and recommender system.

WANG Zhao-Wei, M. S. His main research interests include information retrieval and recommender system.

DONG Zhen-Hua, Ph. D., His main research interests include information retrieval and counterfactual learning.

XU Jun, Ph. D., professor. His main research interests include information retrieval, internet search.

WEN Ji-Rong, Ph.D., professor. His main research interests include internet big data management, information retrieval and text mining.

Background

This work was supported by the National Key Research & Development Program of China No. 2019YFE0198200 entitled "Research on the Key Technologies in Intelligent Legal Assistant and Judicial Cases Analysis for Diversified Dispute Resolution" and National Natural Science Foundation of China No.

62276248.

IR models aim to rank retrieved documents according to their relevance to the given queries. How to design a model, which accurately measures query-document relevance has long been a core research topic in the IR community. Traditionally, relevance is measured by handcrafted formulations which reflect the experts′ knowledge on relevance ranking. Representative models such as BM25 and LMIR, whose performances are closely related to the use of various retrieval heuristics, such as term frequency (TF), inverse document frequency (IDF), document length (DL), etc. Researchers summarized these heuristics knowledge as a set of desirable properties and expressed as formal constraints, referred to as IR axioms. For example, the Term Frequency Constraint (TFC) favors a document with more occurrence of a distinct query term and TF is an essential matching signal in the constraint. Another constraint, length Normalization Constraint (LNC), penalizes a long document when the two documents′ TFs are identical. Axiomatic thinking provides theoretical guidance of training models. Recently, with the development of Learning to Rank (LTR), neural ranking models directly learn from labeled data have become the mainstream of IR ranking. Though state-of-the-art performance has been achieved in a variety of applications, these data-driven models usually rely heavily on the quality of data. Theoretically, both the expert knowledge encoded in IR axioms and the information in labeled training data should provide useful information for relevance ranking, from two different while complementary aspects. One intuitive approach to leverage both of them is generating augmented training instances under IR axioms, then training the models on both the augmented data and the labeled data. However, the IR axioms follow the comparison nature of relevance ranking and judge the relative relevance of two different documents by comparing the matching signals, rather than directly judging the relevance label of a query-document pair. For example, the TFC constraint only requires a document with more occurrences of a query term should have a higher score. In this paper, we propose a framework to enhance the neural ranking model with IR axiomatic knowledge by using contrastive learning, called ACRank. ACRank mainly consists of four stages : 1) Augmented data generation, 2) Embedding generation with neural ranking model, 3) Matching signal extracting 4) Training with contrastive learning loss. ACRank highlights the difference of matching signals from both original relevant and irrelevant documents and also from the constructed one to learn the relative orders through optimizing contrastive learning loss. ACRank can be implemented based on multiple types of IR axioms. In this paper, we focus on the TFC axioms. To generate axiomatic knowledge enhanced data, we perturb the documents in training data along the lines of axioms. Specifically, the matching signal is a function related to the number of query terms. We use a score related with soft TF as matching signal. We tested ACRank on two Chinese datasets Sogou-QCL and TianGong-ST, and one English dataset MS MARCO. The experiments showed that the TFC constraints can guide the training of ACRank to achieve better performances than its underlying model. Empirical analysis showed that ACRank can learn from the IR axiomatic knowledge through modeling the relative order from the difference of matching signals, leading to better performances than the underlying neural models.