第 39 卷 第 2 期 广西师范大学学报(自然科学版) Vol. 39 No. 2
2021 年 3 月 Journal of Guangxi Normal University (Natural Science Edition) Mar. 2021

编者按:本期的计算机学术会议专栏论文来自“第 26 届全国信息检索学术会议”(CCIR2020)。CCIR2020 由中国计算机学会(CCF)和中国中文信息学会(CIPS)联合主办,由西安电子科技大学承办,于 2020 年 8 月 14—16 日在互联网上召开。信息检索学术研究面向人类精准获取信息与知识的需求,研究成果将支撑国家战略决策,推动互联网和 IT 领域的发展,提升行业生产效率,并对社会生活各个领域产生重大影响。全国信息检索学术会议(CCIR)有着悠久的历史,是中国信息检索领域最重要的盛会。本次会议主题是“新形势、新能力、新责任”。

本刊非常欢迎国内计算机学术会议向本专栏推荐论文,对录用的论文我刊将优先安排。

DOI: 10.16088/j.issn.1001-6600.2020082603 http://xuebao.gxnu.edu.cn

杨州,范意兴,朱小飞,等. 神经信息检索模型建模因素综述[J]. 广西师范大学学报(自然科学版),2021,39(2):1-12. YANG Z, FAN Y X, ZHU X F, et al. Survey on modeling factors of neural information retrieval model[J]. Journal of Guangxi Normal University (Natural Science Edition), 2021, 39(2): 1-12.

神经信息检索模型建模因素综述

杨 州[1,2],范意兴[3],朱小飞[1*],郭嘉丰[3],王 越[2]

(1. 重庆理工大学 计算机科学与工程学院,重庆 400054; 2. 搜狐公司智能媒体研发中心,北京 100190;
3. 中国科学院计算技术研究所 网络数据科学与技术重点实验室,北京 100190)

摘 要:信息检索模型被广泛运用于搜索引擎中,且在工业领域被广泛应用。信息检索任务中,模型对信号量的侧重建模导致模型指标差异巨大。目前模型大部分基于以下部分或全部信息建模:精确信号量、相似信号量、信号量区分度、查询词权重、临近量、文本结构信息、不同分布假设。本文介绍各个建模因素的具体含义,并通过引用相关实验例证该因素对于建模起到的积极作用。基于以上实验及分析,最后对信息检索模型的未来发展及趋势作进一步讨论和分析。

关键词:信息检索;深度学习;卷积神经网络;循环神经网络;综述

中图分类号:TP391.3 文献标志码:A 文章编号:1001-6600(2021)02-0001-12

随着科技进一步发展,信息检索技术不仅被运用于常见的搜索引擎,而且被运用于问答[1-2]、社区问答[3]、对话等任务,信息检索技术的快速发展为人类生活提供了极大便利。近年来,随着深度学习[4]在词性标注[5]、语法分析[6]、情感分析等 NLP 任务的发展,不同深度学习在信息检索任务[7]上模型基于不同的假设被提出。相比于传统信息检索模型,深度学习模型指标有较大提升。但结合目前信息检索挑战,挖掘更重要建模因素,以构建更优越的模型成为信息检索的重要问题。

目前信息检索任务存在以下难题。

1) 匹配失误。

匹配失误指模型将 2 段意思相近的文本判断为不相关。匹配失误由多方面原因导致,例如,“特朗普近期干了什么”与“川普最近做了啥”,2 段文本所表示的意思相同,由于相同的词较少,模型往往会误判 2 段文本不相关,由此导致匹配失误。传统信息检索模型往往只考虑查询与文档共同出现的词,忽略近义词。由于缺失近义词的匹配,模型容易将相关的文档判定为无关。另一方面,一词多义往往也会带来同样的匹配失误问题。例如“苹果”一词既可表述一种电子产品,也可表述一种水果,模型往往不能明白用户搜索的意图而导致匹配出不满足需求的文章。深度学习模型中,大量模型将词用稠密的词嵌入[8]表示,然后通过计算词嵌入的向量之间夹角来作为词与词之间相似性的度量。该方式可一定程度缓解匹配失误问题。

收稿日期:2020-08-26 修回日期:2020-09-22

基金项目:国家自然科学基金(61722211,61502065);重庆市基础科学与前沿技术研究项目(cstc2017jcyjBX0059,cstc2017jcyjAX0339);重庆市教委语言文字科研项目重点项目(yyk20103)

通信作者:朱小飞(1979—),男,江苏扬州人,重庆理工大学教授,博士。E-mail:zxf@cqut.edu.cn

2) 查询与文档结构差异巨大。

文本检索任务中,查询与文档结构上存在异质性差异,即查询和文档在长度及结构方面差异巨大。常见查询中,用户输入的查询语句一般为简短的词语或短语,如用户若想搜索“乔布斯在苹果公司的事迹以及其设计理念”,用户的输入很短且不同文档间的字数差异巨大,从几百字到上万字不等;查询和文档在组织结构上差异巨大,导致文档长度问题[9]。查询可以由不同的关键字表示,也可以由关键字组成的短语表示。一般情况下,查询不具有十分复杂的语法结构,而文档为表达其主题而具有十分复杂的组织结构。由于结构不同,不同的准则被相应提出[10]。查询和文档的异质性差异导致文本检索模型结构不同于其他文本匹配模型[11]。

3) 不同匹配需求。

由于查询与文本的异质性问题,查询与文档间的匹配关系可以是全局或者局部的。冗长假设[12]认为文档主题集中,文档的每个部分都围绕该主题展开阐述,若查询与该文档相关,查询应该与整个文档内容相关;范围假设认为文档可分为多个主题,文档不同的部分围绕不同的主题进行阐述,若查询与该文档相关,查询应该与文档某个部分相关,而不是整体相关。以上 2 种假设在实际查询任务中都有所体现,而模型如何合理利用相应假设进行检索成为文本检索模型设计面临的重要问题。

4) 临近关系。

由于文档较长,不同查询词在文档中的分布不同,相关研究表明各个查询词在文档中的临近量十分重要。临近量是指若查询词在文档中较为集中,则查询与文档相关的可能性较大,反之可能性较小。

为缓解或解决以上问题,信息检索模型提出不同构建原则,例如:利用相似信号量以及信号量区分度原则以缓解“匹配失误”问题,但由于查询词数量较少,注重精确信号量与查询词权重也十分重要;利用不同粒度的结构信息可缓解“查询与文档结构差异巨大”问题;挑选不同匹配信息,同时考虑全局信息,有利于解决“不同匹配需求”问题;通过添加文档词位置信息或者将文档分段匹配能够缓解“临近量”问题。

本文内容组织结构如下:第 1 章简单介绍信息检索任务;第 2 章详细介绍不同建模原则在模型构建中所起的作用;第 3 章是展望;第 4 章是结语。

1 信息检索任务简介

1.1 问题描述

文本检索可简单描述为用户输入一个查询语句,模型将该查询与库中的文档打分,然后将文档按照相关度从高到低的排序返回 K 个最佳的文档给用户。返回最相关的 K 个最佳文档为该任务的目标。若将 $S_{\text{train}}=\{s_1^i, s_2^i, r^i\}_{i=1}^N$ 表示为模型训练时的文本,其中 $s_1^i \in S_1$, $s_2^i \in S_2$ 分别为查询项和文档,$r^i \in \mathbf{R}$ 表示查询 s_1^i 和文档 s_2^i 的相关程度。信息检索模型 $f: S_1 \times S_2 \to R$ 目标为对于测试数据 S_{test} 上的任意输入 $s_1 \in S_1$, $s_2 \in S_2$,模型能够较准确得出相关度分数 r,并通过该相关度的高低对文档进行排序并返回结果。

下面给出 1 个例子,其信息检索任务可描述为以下问题:

s_1^1:健康的生活方式。

s_2^1:健康生活方式是指有益于健康的习惯化的行为方式……

s_2^2:有的人往往是透支健康,再花大价钱医治。其实在生活中最好的医生是自己,应在健康时多投资,以防患病……

s_2^3:随着现在生活水平的提高,社会进步越来越快,人们精神压力也特别大,所以不管是工作还是生活上都养成一些坏习惯……

给出查询 s_1^1 以及待排序的文档 s_2^1、s_2^2、s_2^3,将查询与文档通过不同的方式组合放入模型进行训练,模型为待排序文档进行打分,并按照文档的相关度高低对文档排序。计算查询 s_1^i 与文档 s_2^i 的相关度得分 r^i 为信息检索建模问题。

1.2 数据集简介

信息检索中比较典型的评测数据集有如下 6 个。

1) Robust04: Robust04 是一个小型新闻数据集,该数据集来自 TREC Robust Track 2004,Robust04-Title 意味着文章的标题被用于查询。该集合包含 5×10^5 个文档和 250 条查询,词汇量大小为 6×10^5,文档大小为 252 MiB,详细描述见表 1。

2) ClueWeb-09-Cat-B 数据集: ClueWeb09 数据集为支持信息检索等任务研究而创建,该数据集的主题来自于 rec web tracks2009、rec web tracks2010 和 rec web tracks2011,被广泛应用于 TREC 的会议。

3) MQ2007 与 MQ2008 数据集: 使用了 Gov2 网页集合、来自 Trec2007 和 Trec2008 数据集的百万查询。本文简称这 2 个查询集为 MQ2007 和 MQ2008。在带有标签文档的 MQ2007 中有 1 692 条查询; 在带有标签文档的 MQ2008 中有 784 条查询。

4) Sogou-Log 数据集: 该数据集来自中国商业搜索引擎搜狗网的搜索日志提取的中文查询日志,包含 96 229 条查询,每条查询平均对应 12 个相关文档。由于结果来自商业搜索引擎,返回的文档往往具有较高的质量。该数据集的训练数据由 DCTR 模型过滤之后得到,测试数据使用不同的模型 DCTR、TACM[13] 过滤,形成了不同的数据集 Testing-SAME、Testing-DIFF、Testing-RAW(未经过滤的原数据)。

表 1 数据集统计

Tab. 1 Data set statistics

数据集名称	查询数目	文档数目
Robust04	250	500 000
ClueWeb-09-Cat-B	150	34 000 000
MQ2007	1 692	65 323
MQ2008	784	14 384
Sogou-Log	96 229	1 189 436
Bing-Log	10 043	5 002150
Bing-Search-Weight	207 494	1 170 067
Bing-Search-Unweight	206 561	243 024

5) Bing-Log 数据集: 该数据来自于 Bing 搜索引擎在 2006 年的英文查询日志。同样地,该数据训练集用 DCTR 模型得到,测试数据通过 DCTR、TACM 以及未过滤的数据形成 3 种类型 Testing-SAME、Testing-DIFF、Testing-RAW。

6) Bing-Search-Weight 与 Bing-Search-Unweight 数据集: 该数据集是利用 Bing 搜索引擎在 2012 年 1 月和 2014 年 9 月的日志产生的,分为 weight 与 unweight 2 种类型,其训练集的查询与文档是同一份数据集,且长度分别为 199 753 与 998 765,区别是测试集中 weight 数据集按照用户点击频率选取 query,而 unweight 数据集以平均的概率选取。

2 信息检索模型建模原则

随着深度学习在自然语言处理任务的发展,信息检索模型也开始广泛使用深度学习。本文根据不同建模策略对深度信息检索模型进行探讨,主要包括以下 7 点: 精确信号量、相似信号量、混合信号量、查询词权重、临近量、文本结构信息、不同分布假设,其中精确信号量、相似信号量、混合信号量分别指查询与文档共有的词、文档中与查询相似的词、相似信号量与精确信号量的混合。精确信号量为模型提供准确的判断依据,而相似信号量则提供语义相似的依据。将 2 种信号量有区别的结合为混合信号量对模型也十分重要。考虑到查询较短,为不同查询词设置不同权重有重要意义。由于文档较长,合理引入文档结构信息

以及查询词在文档的分布情况、相分布位置为模型判断查询与文档是否相关提供了合理依据。

2.1 精确信号量

2.1.1 精确信号量简介

精确信号量是指出现在查询中的文档词,例如查询项为“比特币 新闻”,若文档中出现“比特币”或者“新闻”,则称该信号为精确匹配信号量。由于查询项长度较短,精确信号量十分重要。在文本检索任务中,有的传统模型例如 BM25 只利用精确信号量建模并取得一定效果。与之相对应的是一些深度学习模型在建模过程中忽略了精确信号量,反而导致了模型指标不佳。

2.1.2 实验对比

本节将只注重精确信号量的传统信息检索模型 BM25[14]、QL(query likelihood model)[15]与忽略精确信号量的深度学习模型 ARC-I[16]、DSSM[17]、CDSSM[18]对比,体现精确信号量在信息检索模型中的重要程度。

该实验来自 DRMM 模型[19],在 Robust-04 与 ClueWeb-09-Cat-B 数据集上进行。表 2 为模型实验,其中,只注重精确匹配信号量的模型 BM25 与 QL 指标较为接近,且远远超过遗失精确匹配信号量的模型。在遗失精确匹配信号量的模型中,ARC-I 指标最差,DSSM 指标最佳。在 Robust-04 Topic titles 与 Topic descriptions 上,以 MAP 指标衡量,QL 模型比 DSSM 模型绝对指标分别提升了 15.8%、16.8%; BM25 模型比 DSSM 模型绝对指标分别提升了 16%、16.3%; 在 ClueWeb-09-Cat-B Topic titles 与 Topic descriptions 数据集上(见表 3),对应的 MAP 指标,QL 模型比 CDSSM 模型绝对指标分别提升 3.6%、2%; BM25 模型比 CDSSM 模型绝对指标分别提升 3.7%、2.5%(由于该实验中 CDSSM 比 DSSM 指标高,故采用 CDSSM 作为对比)。所有模型与文档标题进行的匹配得分低于与文档简介的得分。因为文档简介长于文档标题,文档简介中的精确信号量多于标题。

表 2 Robust-04 数据集上各模型实验对比

Tab. 2 Comparison of model experiments on Robust-04 dataset

模型	Topic titles			Topic descriptions		
	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
QL	0.253	0.415	0.369	0.246	0.391	0.334
BM25	0.255	0.418	0.370	0.241	0.399	0.337
DSSM	0.095	0.201	0.171	0.078	0.169	0.145
CDSSM	0.067	0.146	0.125	0.050	0.113	0.093
ARC-I	0.041	0.066	0.065	0.030	0.047	0.045

表 3 ClueWeb-09-Cat-B 数据集上各模型实验对比

Tab. 3 Comparison of model experiments on ClueWeb-09-Cat-B dataset

模型	Topic titles			Topic descriptions		
	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
QL	0.100	0.224	0.328	0.075	0.183	0.234
BM25	0.101	0.225	0.326	0.08	0.196	0.255
DSSM	0.054	0.132	0.185	0.046	0.119	0.143
CDSSM	0.064	0.153	0.214	0.055	0.139	0.171
ARC-I	0.024	0.073	0.089	0.017	0.036	0.051

2.1.3 实验结论

以上实验说明,在文本检索任务中,精确信号量具有十分重要的作用,主要原因有:① 查询语句较短而文档较长,若忽略精确的关键信息,模型很难学习出较为准确的信息进行判别;② 深度学习模型非常依赖于数据,而实验中的数据量较小,导致模型指标较差。

2.2 相似信号量

2.2.1 相似信号量简介

大量深度学习模型利用词嵌入表示每个单词。词嵌入为低维稠密的向量，且相似的词利用词嵌入计算的 Cosine 值较高。例如“猫”与“狗”的相似度高于“猫”与“凳子”的相似度。文本表示方式的多样性造成了文本结构的多样性，同一意思可由不同词、不同语法结构表示；相同的词在不同语义环境表示不同的含义。用户想搜索关于“宠物猫”的文章，若很少有文章包括该关键词，模型会利用与其较为相似的信号量进行判别，例如文章中包含“动物”、“狗”等词语会更容易被判断为相关文章，而包含“凳子”、“椅子”等词语的文章被判断为相关的概率较低。若用户搜索“川普”，模型如何判断用户需要搜索关于美国总统“特朗普”的文档还是关于“四川普通话”的文档是一个需要解决的问题。若忽略上下文环境，模型将较难判定用户搜索的文档是否为用户所需，而结合文档中的大量相似信号量利于文档被合理检索。

2.2.2 实验对比

相比于 BM25 和 QL 模型，DRMM 模型注重精确信号量的同时，也考虑了全局的相似信号量。目前很多深度学习模型考虑了相似信号量，但为保证文档长度一致，模型几乎都采用截断文档的方式。DRMM 模型保留了文档所有的信号量，故用该模型在同等实验环境下与 BM25 与 QL 进行对比，说明相似信号量在信息检索模型中发挥的作用。值得注意的是，本节选取 DRMM 在实验结果上最优的版本进行对比。该实验在 Robust-04、ClueWeb-09-Cat-B 数据集上进行，实验取自文献[19]，由表 4 可以看出，在 Robust-04 Topic titles 与 Topic descriptions 上，以 MAP 指标衡量，DRMM 模型比 QL 模型绝对指标分别提升 0.026、0.029；DRMM 模型比 BM25 模型绝对指标分别提升 0.024、0.034；由表 5 可以看出，在 ClueWeb-09-Cat-B collection Topic titles 与 Topic descriptions 数据集上，对应的 MAP 指标上，DRMM 模型比 QL 模型绝对指标分别提升 0.013、0.012；DRMM 模型比 BM25 模型绝对指标分别提升0.012、0.007。

表 4 Robust-04 数据集上各模型实验对比

Tab. 4 Comparison of model experiments on Robust-04 dataset

模型	Topic titles			Topic descriptions		
	Model	MAP	NDCG@20	MAP	NDCG@20	P@20
QL	0.253	0.415	0.369	0.246	0.391	0.334
BM25	0.255	0.418	0.37	0.241	0.399	0.337
DRMM	0.279	0.431	0.382	0.275	0.437	0.371

表 5 ClueWeb-09-Cat-B 数据集上各模型实验对比

Tab. 5 Comparison of model experiments on ClueWeb-09-Cat-B dataset

模型	Topic titles			Topic descriptions		
	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
QL	0.100	0.224	0.328	0.075	0.183	0.234
BM25	0.101	0.225	0.326	0.080	0.196	0.255
DRMM	0.113	0.258	0.365	0.087	0.235	0.310

相比于互联网上巨大的数据量，Robust-04 与 ClueWeb-09-Cat-B 数据较小。由于深度学习模型都是数据驱动，数据量的大小对模型指标影响很大。以下使用以 Bing 搜索的日志为基础生成的数据集作为实验数据集。相对于以上 2 个数据集，该数据集较大且分为 2 个版本，区别在于采样时，weighted 数据集对查询考虑频次抽取，而 unweighted 则没有。从表 4 与表 5 中可以看到，在数据量较大时，注重精确匹配信号量的模型 BM25、QL 与注重相似信号量的模型指标相当。该实验说明在数据量较大的情况下，相似信号量作用较大。

2.2.3 实验结论

本节在较小与较大2种数据集上测试了相似信号量的作用:① 在较小数据集上,相似信号量有一定作用,但是重要程度不及精确信号量;② 在较大数据集上,相似信号量与精确信号量同样重要。

2.3 混合信号量

2.3.1 混合信号量度简介

当模型使用了相似信号量与精确信号量时,如何混合2种信号量需要进一步探讨。目前相关研究[20]表明,将2种信号量加以区分并结合为混合信号量为模型匹配起到积极作用。

2.3.2 实验对比

本节使用HiNT模型[20]内部实验与Duet[21]等模型实验对比,模型实验分别在MQ2007与Bing-Search数据集上进行对比。

利用HiNT模型的不同版本进行对比,实验详见表6。其中:第1个模型代表该模型只使用精确信号量;第2个模型代表使用Cosine计算查询与文档所有的信号量,并没有对精确信号量与相似信号量进行区分;第3个模型将所有信号量生成histogram[19]进行区分,只利用精确信号量;第4个模型将精确信号量与Coinse生成的信号量分开进行学习,其中spatial GRU[22]为特殊的GRU[23],能够从左到右、从上至下对矩阵进行扫描[24]。

表6 HiNT不同版本在MQ2007数据集上实验对比

Tab. 6 Experimental comparison of different versions of HiNT on MQ2007 dataset

模型	P@10	NDCG@10	MAP
Mxor+MLP	0.384	0.435	0.461
Mcos+MLP	0.329	0.344	0.386
Mhist+MLP	0.393	0.447	0.469
Mxor+Mcos+spatial GRU	0.405	0.470	0.484

如表6所示,HiNT 2种不同版本都表明将精确信号量与相似信号量分开学习有利于模型学习。相比于同时包含精确信号量与相似信号量且不加以区分的模型,模型指标在MAP上提升21.5%。表7与表8代表不同模型在Bing-Search-Unweight与Bing-Search-Weight的指标对比,这些模型都为深度模型,其中Duet模型将精确信号量与相似信号量加以区分学习,而DSSM与CDSSM模型则将所有信号量混合学习。在Bing-Search-Weight的指标数据集上,Duet模型指标相比于指标较好的DSSM模型,指标NDCG@1和NDCG@10分别提升10.2%、3.1%;在Bing-Search-Weight数据集上,Duet模型指标相比于指标较好的CDSSM模型,指标NDCG@1和NDCG@10分别提升17.9%、9.9%。

表7 模型在Bing-Search-Unweight数据集上的指标对比

Tab. 7 Comparison of model experiments on Bing-Search-Unweight dataset

模型	NDCG@1	NDCG@10
DSSM	0.343	0.644
CDSSM	0.343	0.640
Duet	0.378	0.664

表8 模型在Bing-Search-Weight数据集上的指标对比

Tab. 8 Comparison of model experiments on Bing-Search-Weight dataset

模型	NDCG@1	NDCG@10
DSSM	0.258	0.482
CDSSM	0.273	0.482
Duet	0.322	0.530

2.3.3 实验结论

本节对比了HiNT 2个变种版本及Duet模型与DSSM和CDSSM,说明了在深度学习模型中,区分精确信号量与相似信号量对模型指标提升起着一定的作用。

2.4 查询词权重

2.4.1 查询词权重简介

文本检索任务中,查询往往较短且结构简单,文档较长且结构复杂。查询词在文档出现的占比较低,又由于查询词具有不同的重要度,因此,将不同查询词加以区分十分必要。若用户输入查询词“比特币新闻”,用户想获取的信息是偏向于比特币的,而不是偏向于新闻的。若将比特币与新闻视为同等重要,模型将检索出许多关于新闻的文档,而这并非用户所需,由此导致指标欠佳。

2.4.2 实验对比

本节使用 HiNT 模型内部实验,该实验对比了是否具有查询词权重的模型,其中模型 1 为未注重查询词权重的模型,模型 2 为注重查询词权重的模型,见表 9。实验表明模型注重查询词权重后,指标有较大提升。相比于未注重查询词权重模型,该模型在 MQ2007 数据集上 P@10、NDCG@10、MAP 指标分别提升 1.3、2.0 和 3.7 个百分点。

表 9 NiHT 不同模型版本在 MQ2007 数据集上实验指标

Tab. 9 Experimental comparison of different versions of HiNT on MQ2007 dataset

模型	P@10	NDCG@10	MAP
M_{xor}+M_{cos}+spatial GRU	0.405	0.470	0.484
S_{xor}+S_{cos}+spatial GRU	0.418	0.490	0.502

2.4.3 实验结论

本节实验表明查询词权重在文本检索任务中较为重要。目前模型中,有 3 种方式计算查询词权重: 利用 IDF 生成查询词权重; 利用词嵌入学习 Term Gate [19] 生成查询词权重; 利用词嵌入的低维表示与交互信号量拼接生成查询词权重。在不同模型中,不同的查询词权重会取得不同的效果。通过 IDF 作为权重能减小模型参数,在词嵌入不能很好表达查询词时该方法较好; 通过 Term Gate 方式作为查询词权重的优点是减少了 IDF 计算时间,灵活性较高,且能够在训练中不停更正; 利用拼接方法的优点是当神经网络不支持以上 2 种方式时,该方式可作为强调查询词权重的方式,例如,有卷积神经网络而导致不能使用 Term Gate 时,该方式能够较好生成查询词权重。

2.5 临近量

2.5.1 临近量简介

查询词在文档中的分布不同,有研究 [25] 表明相关文档的查询词更加临近。目前模型中,利用临近量有 2 种方法: 第一,将文档按照不同的方法切分; 第二,将文档词的位置加入模型,由于文档词的位置为自然数,为凸显相近词的位置关系,很多模型将相近词的位置关系用不同的方式做了平滑处理。

文档切分方法包括文档分片与联合分片 2 类,其中有 3 种方式可进行文档分片,即: ① 基于语义划分。由于文档可能由不同主题组成,该方式将不同的主题段落划分成不同的分片。② 基于自然段落。文章具有自然段落,该方式将原始的自然段落作为分隔进行不同的分段。③ 基于固定长度。该方式将文档基于固定长度 N 进行划分,其中 N 为模型超参数。而联合分片首先考虑文档与查询的交互,通过交互信息考虑划分的片段。

2.5.2 实验对比

相关研究人员提出了临近量的不同计算方式,给出了相关的统计数据,详见表 10,其中 MinDist、MinCover 代表临近量的不同计算方式。在 MinCover 方式下计算临近量时,除 FR88-89 数据集以外,相关文档比无关文档的平均临近量都要小。而用 MinDist 方式计算临近量时,所有相关文档的平均临近量都小于无关文档。

表 11 来自 DeepRank [26] 模型不同版本对比实验。以下模型的区别在于查询词在文档的位置计算函数的差异,详细地假设词 w_{qi} 在文档中的位置为 p, DeepRank-Const 采用线性映射函数 $g(p)=p$; DeepRank-Linear 采用线性映射函数 $g(p)=(L-p)/L$(L 为常数); DeepRank-Exp 采用 $g(p)=a/(p+b)$ 映射函数(a、b

为常数);DeepRank-Recip 采用 $g(p)=a\times\exp(-p/b)$(其中 a、b 为常数)。实验表明采用不同的位置信息对模型指标影响较大,当模型使用的位置信息处理函数能够将临近词的位置映射得较为相近(DeepRank-Recip 模型)时,模型指标最佳。该实验也说明临近量在模型中所起的作用。

表 10 计算邻近量不同方法及指标

Tab. 10 Different methods and indexes for calculating proximity

数据集	MinDist		MinCover	
	无关文档	相关文档	无关文档	相关文档
AP88-89	30.64	16.18	50.78	46.43
FR88-89	39.83	39.35	104.13	150.90
TREC8	31.77	19.15	56.25	57.43
WEB2g	67.91	61.20	108.38	153.48
DOE	11.68	7.66	108.38	153.48

表 11 DeepRank 不同版本在 MQ2007 数据集上指标对比

Tab. 11 Experimental comparison of different versions of DeepRank on MQ2007 dataset

模型	NDCG@1	NDCG@5	MAP
DeepRank-Const	0.384	0.384	0.473
DeepRank-Linear	0.431	0.445	0.492
DeepRank-Exp	0.441	0.454	0.494
DeepRank-Recip	0.441	0.457	0.497

2.5.3 实验结论

本节从添加位置信息与文档切分的角度说明位置信息所起的作用。实验表明临近量对文本检索模型具有一定贡献,但定义与利用临近量是信息检索模型的待解决问题。

2.6 文本结构信息

2.6.1 文本结构信息简介

信息检索任务中,文档具有复杂的结构特性,文档结构的层次性给任务提出了一大挑战。文档由段落组成,段落由句子组成,句子由短语组成,短语由词组组成,词组由字组成[27]。而查询结构较为简单,一般由句子组成,句子由词组成,词由字组成。将查询的不同层次结构信息与文档的层次信息较好利用起来是信息检索的一大问题。卷积神经网络[28]能够较好利用局部小范围的文本层次信息,循环神经网络[29-30]能够很好利用文字中的序列信息,但如何将语句中同层次、不同层次以整体序列信息递归提取[24]并合理利用仍然较难解决。

2.6.2 实验对比

该实验来自 Conv_KNRM 模型[31],在搜狗与 Bing 搜索的日志信息上进行,其中 Conv_KNRM 为 KNRM 模型[32]的多粒度版本,Conv_KNRM 通过一维卷积将查询与文档的多粒度信息提出,并且将查询与文档对应的粒度信息进行交互,再将交互之后的信息交给 KNRM 模型处理。因此,相比于 KNRM 模型,Conv_KNRM 模型加入了多粒度信息,由此提升了模型指标。相对于 KNRM 模型,Conv_KNRM 模型的 NDCG@1、NDCG@10、MRR 指标在 Sogou-Log 数据集上分别提升 27.2%、12.3%、5.9%;在 Bing-Log 数据集上分别提升 44.2%、30.8%、33.5%,具体结果见表 12。

2.6.3 实验结论

本节实验表明利用文本的结构信息对模型的构造起着积极作用,但目前模型只利用局部信息,如何同时利用好大范围结构信息与局部结构信息仍然是该任务的一大难点。

表 12 模型在 Sogou-Log 和 Bing-Log 数据集上的指标对比

Tab. 12 Comparison of model experiments on Sogou-Log and Bing-Log dataset

模型	Sogou-Log			Bing-Log		
	NDCG@1	NDCG@10	MRR	NDCG@1	NDCG@10	MRR
KNRM	0.264	0.428	0.338	0.208	0.334	0.265
Conv_KNRM	0.336	0.481	0.358	0.300	0.437	0.354

2.7 不同分布假设

2.7.1 不同分布假设简介

文本匹配有 2 种假设。冗长假设认为文档主题集中,文档的每个部分都围绕该主题展开阐述,若查询与该文档相关,查询应该与整个文档内容相关;范围假设认为文档可分为多个主题,文档不同的部分围绕不同的主题进行阐述,若查询与该文档相关,查询应该与文档某个部分相关,而不是整体相关。由此导致不同的匹配需求。如何结合 2 种需求是信息检索任务要解决的问题。

2.7.2 实验对比

关于不同分布假设的问题,目前模型给出 2 类解决方案。第一,将文档所有信号量提取出,只要提取信号量的函数足够有效,无论有效信号量是集中分布还是全局分布都可提取到。例如 DRMM 模型、PACRR [33] 的 k-window 版本都为该方案。第二,选取 K 个有效信号量与统计全局信号量方法相结合,无论相关文档为局部相关或全局相关,该模型都可获取有效信号量。

表 13 中: $HiNT_{ID}$ 模型只考虑局部信息; $HiNT_{AD}$ 只考虑全局信息; $HiNT_{HD}$ 同时考虑 2 种信息。实验表明同时考虑局部信号量与全局信号量的模型具有优越性。

表 13 不同分布假设对 HiNT 模型影响

Tab. 13 Impact of different matching requirements on HiNT model

模型	P@10	NDCG@10	MAP
$HiNT_{ID}$	0.389	0.405	0.418
$HiNT_{AD}$	0.446	0.472	0.490
$HiNT_{HD}$	0.464	0.483	0.502

2.7.3 实验结论

本节实验说明不同分配假设在信息检索任务中的贡献。目前模型针对该问题提出了 2 种解决方案:第 1 种方案将文档所有信号量按照相关度由高到低排序取出,重要信号量在文档任何位置都可以被取出,但该方法忽略了高维度语义信息在文本中的作用;第 2 种方案为取出重要信号量的同时也取出高维度语义信息,让单粒度重要信息与高维度语义信息信号竞争,模型选取单粒度与高纬度最优信息,较前者考虑更加符合匹配习惯。

3 展望

本文主要阐述了信息检索模型的 7 个建模要素:精确信号量、相似信号量、信号量区分度、查询词权重、临近量、文本层次结构信息、不同分布假设。针对以上分析,得出目前模型存在的问题,并提出信息检索模型未来发展方向。

1) 充分利用相似信号量与文本层次结构信息。在信息检索模型中,如何利用相似信号量的语义环境,并结合文档的层次结构信息,将查询与文档的不同层次的结构信息更好匹配是信息检索模型的一个突破点。

2) 临近量定义及更好利用是模型的一个突破点。目前模型利用对文档进行分片的方式让临近量在

建模过程得以体现。大量模型采用文档分片而不是联合分片方式,本文认为构造有效联合分片函数对文档进行切分将更有效地引入临近量。模型考虑文档词的位置信息也是引入临近量的体现,与上文提到的临近量的2种定义MinDist、MinCover相比,该方式并不能很直接地将临近量引入,以该方式引入临近量并不一定最合适。

3)信息检索任务存在2种不同分布假设,HiNT模型提出解决2种不同假设的方法。将该假设与语义信息结合考虑,即通过语义信息判断该文档是全局或局部相关,并通过相应鉴别函数来完成相应的选取功能。

4)额外信息的加入。一词多义与同义词广泛存在,通过上下文语义环境可缓解"匹配失误"问题,但仍存在许多现实中相关实体在文本中无法清晰表示的情况,目前模型利用WordNet、Knowledge Graph[34]、Wiki-pedia[35]等额外知识对模型进行拓展。利用额外信息对模型进行改进仍是很有前景的方向。

5)模仿人类检索行为。人类会根据自身习惯判断文档是否相关[36],例如,人们会先搜索查询词出现的文档部分,并阅读相关部分;当文档过长时往往不会读完文档,而是跳跃性阅读;阅读完文档内容之后,会根据之前阅读的记忆推测文档是否相关。

4 结语

本文主要阐述了信息检索模型的7个建模要素:精确信号量、相似信号量、信号量区分度、查询词权重、临近量、文本层次结构信息、不同分布假设。目前模型全部或部分考虑了以上建模因素,但仍存在许多可改进的地方。本文在详细阐述了以上建模因素之后提出未来研究方向,希望新的建模因素不断被提出,以往的建模因素能更好地被利用。

参 考 文 献

[1] YANG Y, YIH S W, MEEK C. WikiQA: a challenge dataset for open-domain question answering [C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudburg, PA: ACL, 2015: 2013-2018.

[2] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100,000+ questions for machine comprehension of text [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudburg, PA: ACL, 2016: 2383-2392.

[3] YANG L, QIU M, GOTTIPATI S, et al. CQARank: jointly model topics and expertise in community question answering [C] // Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. New York: ACM, 2013: 99-108.

[4] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[6] VINYALS O, KAISER L, KOO T, et al. Grammar as a foreign language [EB/OL]. (2015-06-09) [2020-08-26]. https://arxiv.org/abs/1412.7449.

[7] LI H, XU J. Semantic matching in search [J]. Foundations and Trends in Information Retrieval, 2014, 7(5): 343-469.

[8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-09-07) [2020-08-26]. https://arxiv.org/abs/1301.3781.

[9] PANG L, LAN Y Y, GUO J F, et al. A deep investigation of deep IR models [EB/OL]. (2017-07-24) [2020-08-26]. https://arxiv.org/abs/1707.07700.

[10] FANG H, TAO T, ZHAI C X. Diagnostic evaluation of information retrieval models [J]. ACM Transactions on Information Systems, 2011: 7.

[11] 庞亮,兰艳艳,徐君,等. 深度文本匹配综述 [J]. 计算机学报, 2017, 40(4): 985-1003.

[12] CHUKLIN A, MARKOV I, RIJKE M D. Click models for Web search [J]. Synthesis Lectures on Information Concepts Retrieval & Services, 2015, 7(3): 1-115.

[13] LIU Y Q, XIE X H, WANG C, et al. Time-aware click model [J]. ACM Transactions on Information Systems, 2016, 35(3): 16.

[14] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval [C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berlin: Springer-Verlang, 1994: 232-241.

[15] ZHAI C, LAFFERTY J. A study of smoothing methods for language models applied to Ad Hoc information retrieval [C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2001: 333-342.

[16] HU B T, LU Z D, LI H, et al. Convolutional neural network architectures for matching natural language sentences [EB/OL].(2015-03-11) [2020-08-26]. https://arxiv.org/abs/1503.03244v1.

[17] HUANG P S, HE X D, GAO J F, et al. Learning deep structured semantic models for web search using clickthrough data [C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM, 2013: 2333-2338.

[18] SHEN Y L, HE X D, GAO J F, et al. Learning semantic representations using convolutional neural networks for web search [C]// Proceedings of the 23rd International Conference on World Wide Web. New York: ACM, 2014: 373-374.

[19] GUO J F, FAN Y X, AI Q Y, et al. A deep relevance matching model for ad-hoc retrieval [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2016: 55-64.

[20] FAN Y X, GUO J F, LAN Y Y, et al. Modeling diverse relevance patterns in Ad-hoc retrieval [C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2018: 375-384.

[21] MITRA B, DIAZ F, CRASWELL N. Learning to match using local and distributed representations of text for web search [EB/OL]. (2016-10-26) [2020-08-26]. https://arxiv.org/abs/1610.08136.

[22] GRAVES A. Offline handwriting recognition with multidimensional recurrent neural networks [M]// MÄRGNER V, EL ABED H. Guide to OCR for Arabic Scripts. London: Springer, 2012: 297-313.

[23] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudburg, PA: ACL, 2014: 1724-1734.

[24] WAN S X, LAN Y Y, XU J, et al. Match-SRNN: modeling the recursive matching structure with spatial RNN [J]. Computers & Graphics, 2016, 28(5): 731-745.

[25] TAO T, ZHAI C X. An exploration of proximity measures in information retrieval [C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 295-302.

[26] PANG L, LAN Y Y, GUO J F, et al. Deeprank: a new deep architecture for relevance ranking in information retrieval [C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 257-266.

[27] PANG L, LAN Y Y, GUO J F, et al. Text matching as image recognition [C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 2793-2799.

[28] LECUN Y, BOTTOU L. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29] LEVIN E. A recurrent neural network: limitations and training [J]. Neural Networks, 1990, 3(6): 641-650.

[30] DATAR M, Immorlica N, Indyk P, et al. Locality sensitive hashing scheme based on p-stable distributions [C]// Proceedings of the Twentieth Annual Symposium on Computational Geometry. New York: ACM, 2004: 253-262.

[31] DAI Z Y, XIONG C Y, CALLAN J, et al. Convolutional neural networks for soft-matching N-grams in Ad-hoc search [C]//

Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. New York: ACM, 2018: 126-134.

[32] XIONG C Y, DAI Z Y, CALLAN J, et al. End-to-end neural ad-hoc ranking with kernel pooling [C] // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 55-64.

[33] HUI K, YATES A, BERBERICH K, et al. PACRR: a position-aware neural IR model for relevance atching [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudburg, PA: ACL, 2017: 1049-1058.

[34] PONTES J, JOÃO D, CARVALHO R A D, et al. Information retrieval to knowledge retrieval: reflections and proposals [J]. Perspectives em Ciência da Informacão, 2013, 18(4): 2-17.

[35] GABRILOVICH E, MARKOVITCH S. Wikipedia-based semantic interpretation for natural language processing [J]. Journal of Artificial Intelligence Research, 2009, 34: 443-498.

[36] WU H C, LUK R W P, WONG K F, et al. A retrospective study of a hybrid document-context based retrieval model [J]. Information Processing & Management, 2007 43(5): 1308-1331.

Survey on Modeling Factors of Neural Information Retrieval Model

YANG Zhou[1,2], FAN Yixing[3], ZHU Xiaofei[1*], GUO Jiafeng[3], WANG Yue[2]

(1. School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China;
2. Intelligent Media R & D Center SOHU, Beijing 100190, China;
3. CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Information retrieval models are widely used in search engines. In the task of information retrieval, these models focuses on the different semaphores, which leads to great differences in model performance. At present, most models are based on part or all of the following information: exact signals, similar signals, signals differentiation, query word weight, proximity, text structure, and different distribution assumptions. This paper introduces the specific meaning of each modeling factor, and exemplifies the positive effect of this factor on modeling through relevant experiments. Based on the above experiments and analysis, this paper finally discusses and analyzes the future development and the trend of information retrieval model.

Keywords: information retrieval; deep learning; convolutional neural network; recurrent neural network; survey

(责任编辑　吴佃华)