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[1)]

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

论文摘要

信息检索模型被广泛运用于搜索引擎中。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。本文介绍了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。

论文简介

随着科技进一步发展，信息检索技术被运用于常见的搜索引擎，且为人类生活提供了极大的便利。相比于传统的信息检索模型，深度学习模型指标有较大的提升。但如何结合目前信息检索的挑战挖掘更重要的建模因素，以构建更优越的模型成为信息检索的一大难点。

目前信息检索任务存在以下难点：

（1）匹配失误

匹配失误指模型将两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”。传统的信息检索模型往往只考虑到查询与文档共同出现的词，而忽略近义词。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。

（2）查询与文档结构差异巨大

MQ2007数据集			
	训练集	测试集	总计
问题数量	1165	336	1501
句子数量	45428	13302	58730
问题平均长度	-	-	5.869385343
句子平均长度	-	-	1944.343661
问题最长长度	-	-	19
句子最长长度	-	-	35390

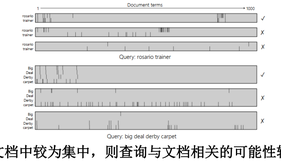
MQ2008数据集			
	训练集	测试集	总计
问题数量	459	156	615
句子数量	9601	2837	12438
问题平均长度	-	-	7.2126
句子平均长度	-	-	2479.682912
问题最长长度	-	-	22
句子最长长度	-	-	36033

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。**1)**查询和文档长度差异巨大，例如在**MQ2007**和**MQ2008**数据集中，查询的平均长度在**10**个字以下，而文档的平均长度在**2000**左右；**2**）查询和文档结构差异巨大，查询一般为简短的几个字组成，而文档为表达其主题通常具有十分复杂的组织结构。

（3）不同匹配需求

相关研究表明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。

（4）临近关系



临近关系指，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。图中为查询**“rosario trainer”**，**“big deal derby carpet”**在文档中的位置关系图，相关实验表明，查询中不同的词更临近则文档更可能相关。

实验仿真

•精确信号量简介

表 3.1 模型在 Robust-04 数据集上各模型实验对比

Model	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
	Topic titles			Topic descriptions		
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
DSSM	0.095	0.201	0.171	0.078	0.169	0.145
CDSSM	0.067	0.146	0.125	0.05	0.113	0.093
ARC-I	0.041	0.066	0.065	0.03	0.047	0.045

表 3.2 ClueWeb-09-Cat-B 数据集上各模型实验对比

Model	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
	Topic titles			Topic descriptions		
QL	0.1	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

•在**Robust-04**和**ClueWeb-09-Cat-B**两个数据集上，注重精确信号量的**QL**和**BM25**模型比不注重精确信号量的**DSSM**、**CDSSM**、**ARC-I**模型**MAP**指标高出了不少，说明了精确信号量在信息检索任务中的重要作用。

•相似信号量

表 3.3 Robust-04 数据集上各模型实验对比

Model	MAP	NDCG@20	P@20	MAP	NDCG@20	P@20
	Topic titles			Topic descriptions		
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 3.4 ClueWeb-09-Cat-B 数据集上各模型实验对比

Model	MAP	nDCG@20	P@20	MAP	nDCG@20	P@20
	Topic titles			Topic descriptions		
QL	0.1	0.224	0.328	0.075	0.183	0.234
BM25	0.101	0.225	0.326	0.08	0.196	0.255
DRMM	0.113	0.258	0.365	0.087	0.235	0.31

•在基准数据集上，以**MAP**指标衡量，**DRMM**模型比**QL**模型绝对指标分别提升了**2.6%、2.9%**；**DRMM**模型比**BM25**模型绝对指标分别提升了**2.4%、3.4%**；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了**1.3%、1.2%**；**DRMM**模型比**BM25**模型绝对指标分别提升了**1.2%、0.7%**。

•混合信号量

表 3.5 HiNT 不同版本模型在 MQ2007 数据集上实验对比

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

表 3.6 模型在 Bing-Search-Unweight 数据集上的指标对比

Model	NDCG@1	NDCG@10
DSSM	34.3	64.4
CDSSM	34.3	64.0
Duet	37.8	66.4

•本节说明了混合信号量的重要性，即包括精确信号量和相似信号量的信号量。在**MQ2007**数据上，对比了**HiNT**中含有混合信号量**Mxor+Mcos+spatial GRU**和不含有混合信号量的其他模型相比，混合信号量占有较大的优势；在**Bing-Search_Unweight**上，含有混合信号量的模型**Duet**模型与不含混合信号量的其他模型**DSSM**和**CDSSM**。本节说明了混合信号量对模型指标提升起着一定的作用。

•查询词权重

表 3.8 NiHT 不同模型版本在 MQ2007 数据集上实验指标

Model	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	0.49	0.502

•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。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**Sxor+Scos+spatial GRU**为含有权重因子的模型，**Mxor+Mcos+spatial GRU** 则为不含有权重的模型。该模型在**MQ2007**数据集上**P@10,NDCG@10**和**MAP**指标提升了**1.3%，2%，3.7%**。

•临近量

•相关研究人员提出了临近量的不同计算方式。并给出了相关的统计数据，详见表**3.9**。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**3.10**来自**DeepRank**模型不同版本对比实验。**DeepRank-Const**为未加临近量的模型，而其他三种以不同的方式计算了临近量加入了模型，实验表明加入临近量的模型指标优于不加临近量的模型。

表 3.9 计算邻近度不同方法及指标

dataset	MinDist		Span	
	non-rel.	rel.	non-rel.	rel.
AP88-89	30.64	16.18	50.78	46.43
FR88-89	39.83	39.35	104.13	150.9
TREC8	31.77	19.15	56.25	57.43
WEB2g	67.91	61.2	108.38	153.48
DOE	11.68	7.66	108.38	153.48

表 3.10 DeepRank 不同版本在 MQ2007 数据集上指标对比

Model	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

•文本层次结构信息

表 3.11 模型在 Sogou-Log 数据集上的指标对比

Model	Sogou-Log			Bing-Log		
	NDCG@1	NDCG@10	MRR	NDCG@1	NDCG@10	MRR
K-NRM	0.264	0.428	0.338	0.208	0.334	0.265
Conv-KNRM	0.336	0.481	0.358	0.3	0.437	0.354

•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该任务提出了一大挑战。该实验来自于**Conv_KNRM**模型，在搜狗与**Bing**搜索的日志信息上进行。**Conv_KNRM**提取了文本词级别的信息，在数据集上，**Conv_KNRM**模型相对于**KNRM**模型**NDCG@1, MRR**指标分别提升了**27.2%, 33.5%**。

•不同匹配需求

表 3.12 由于不同匹配需求带给 HiNT 模型影响

Model	p@10	NDCG@10	MAP
HiNTb	38.9	40.5	41.8
HiNTa	44.6	47.2	49
HiNTe	46.4	48.3	50.2

•本节实验表明不同匹配需求在信息检索任务中的贡献。其中**HiNTID** 为模型只考虑局部信息的模型，**HiNTAD** 只考虑全局信息的版本， **HiNTHD** 为同时考虑两种信息的模型，实验表明同时考虑了局部信号量与全局信号量的模型的优越性。

论文结论

•本文在引用相关实验的基础商，研究探索发现在信息检索任务中，有七个建模因素对建模影响较大，即精确信号量、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，并且仔细分析了其重要性，其中精确信号量、查询词权重、不同分布假设对信息检索建模尤为重要，而如何很好结合以上七个建模因素探索出更优质的模型仍然有着一定的难度，仍然需要研究人员的进一步探索。

西安电子科技大学 XIDIAN UNIVERSITY