量子干涉启发的神经匹配模型（QINM）

原创 张鹏、高珲 天大量子智能与语言理解团队 2020-10-22

点击蓝字，关注我们

导语

文本匹配是信息检索和自然语言处理等领域的基础研究问题。当前，常用的文本匹配模型的最终相关性得分源于局部匹配证据的累加，然而这种机械性的累加并不符合人类检索过程。为了建模人类检索过程中匹配单元之间的交互信息，本篇发表于SIGIR 2020的工作**首次提出量子干涉启发的神经匹配模型（Quantum Interference Inspired Neural Matching Model, QINM）并应用于ad-hoc检索任务中**。具体而言，该模型将查询（Query）及候选文档（Document）定义为量子子系统，构造查询-文档复合系统，进而通过约化密度矩阵编码文档的概率分布，从而建模信息检索过程中匹配单元（即单个查询词与文档的组合）之间的交互信息。实验结果表明，我们所提出模型在Robust-04和ClueWeb-09-cat-B两个数据集上均表现出最佳的性能。本论文是继Sordoni, Nie & Bengio合作发表SIGIR 2013年的量子语言模型之后，**时隔7年SIGIR 2020再次录用的关于量子信息检索的长文文章**。

A Quantum Interference Inspired Neural Matching Model for Ad-hoc Retrieval

Yongyu Jiang, Peng Zhang*, Hui Gao
College of Intelligence and Computing
Tianjin University
Tianjin, China
pzhang@tju.edu.cn

Dawei Song
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
dawei.song2010@gmail.com

GitHub: https://github.com/TJUIRLAB/SIGIR20_QINM

01 基于经典概率的检索思想

信息检索任务的核心是如何衡量查询与候选文档的相关程度，在本篇工作中我们将这一过程形式化为条件概率 $P(R_D|Q)$，其中 R_D 表示查询与文档相关这一事件。根据匹配单元的不同，当前主流的信息检索模型可大致分为经典概率模型、基于依赖关系的检索模型和神经匹配模型，这些基于经典概率的信息检索模型的检索思想大致相同：首先分别计算查询词与候选文档的相关性得分，最后将上述得分相累加获得文档与查询词的最终相关性得分。

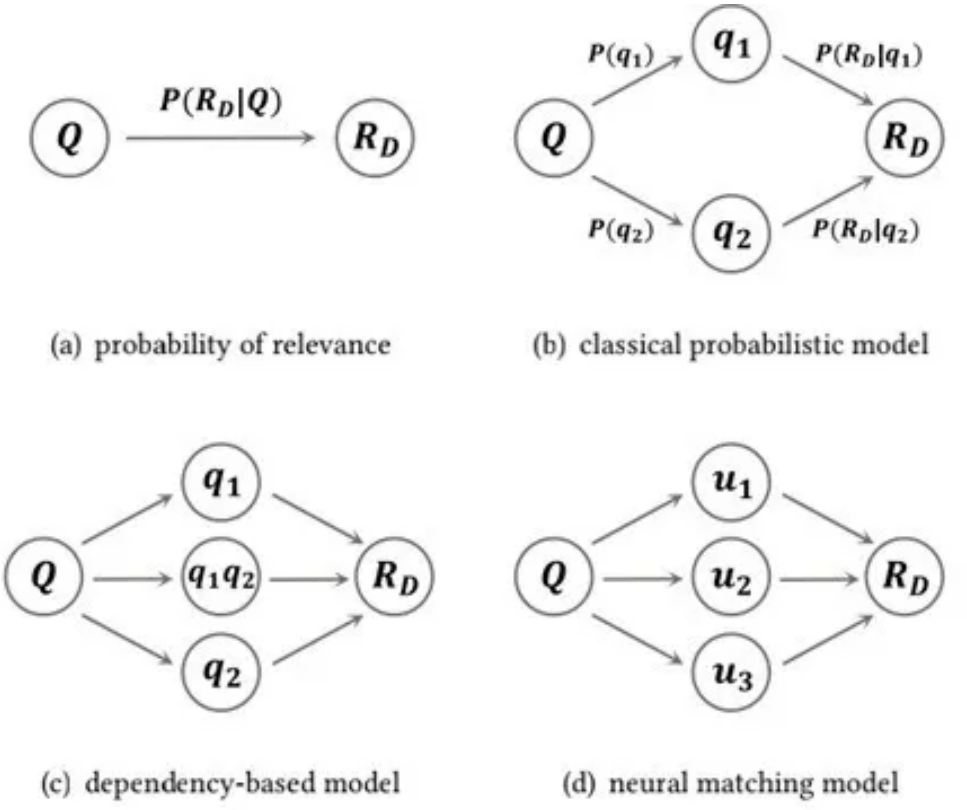

图1 经典信息检索模型

以DRMM为例（图2），将模型进行概率形式化之后可以发现该模型符合经典概率（全概率公式）。具体而言，首先将由匹配直方图所获得局部匹配得分进行归一化处理，然后将该结果与TGN所获得的权值进行加权求和获得最终的相关性得分。我们可以将这一过程表示为下面的公式：

$$P(R_D|q_i) = exp(f(q_i, D))/\sum_{j=1}^{n} exp(f(q_j, D))$$

$$\begin{aligned} P(R_D) &= P(q_1, R_D) + P(q_2, R_D) \\ &= P(q_1)P(R_D|q_1) + P(q_2)P(R_D|q_2) \end{aligned}$$

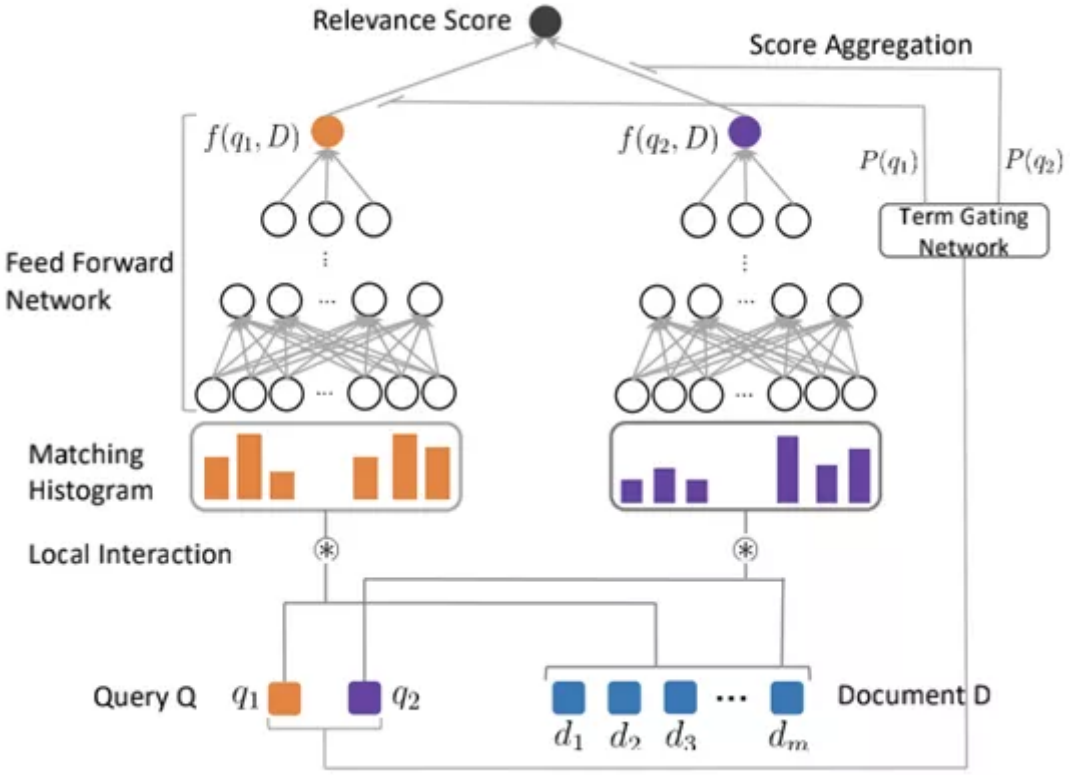

图2 DRMM

02 基于量子概率的检索思想

基于量子概率的检索主要是将检索过程中的概率事件 Q, q_1, q_2 分别表示为相应的状态向量，通过投影测量来探讨文档相关性判断过程中的干扰效应。若将查询表示为 $Q = \alpha q_1 \otimes \beta q_2$ ，则概率 $P(q_1)$ 可以通过图3(a)的投影测量进行表示。进一步，局部相关性得分可以通过条件概率 $P(R_D|q_1)$ 进行表示，这在图3(b)中可表示为一个连续的投影过程。

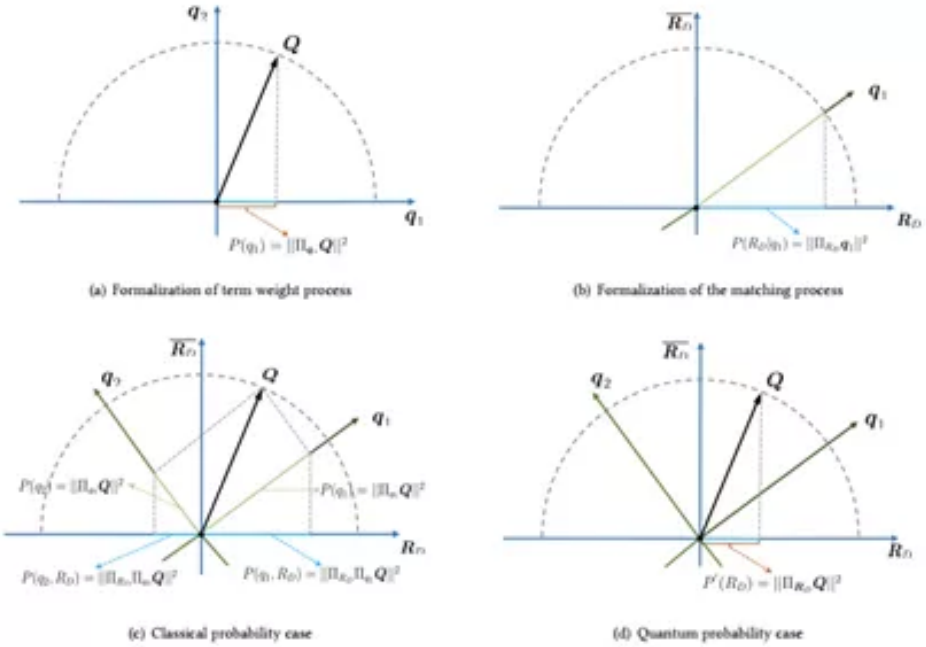

图3 基于量子概率的检索过程

在此基础上，可以通过量子概率分析经典信息检索模型所存在的不足。当假设查询单元的检索过程独立时(如图3(c))，我们将DRMM中的匹配过程形式化为下面的公式：

$$\begin{aligned} P(R_D) &= P(q_1)P(R_D|q_1) + P(q_2)P(R_D|q_2) \\ &= ||\Pi_{R_D}\Pi_{q_1}Q||^2 + ||\Pi_{R_D}\Pi_{q_2}Q||^2 \end{aligned}$$

然而，在文档相关性判断过程中，用户通常会在文档相关性判断过程中考虑文本匹配单元之间的交互作用。与现有的神经匹配模型不同的是，如果查询首先表示为一个状态向量 Q，这意味着它可以在匹配过程中将查询作为一个整体来考虑（如图3(d)）。因此，文档 D 与查询相关的概率计算如下：

$$\begin{aligned} P'(R_D) &= ||\Pi_{R_D}Q||^2 = |\Pi_{R_D}(\Pi_{q_1}Q + \Pi_{q_2}Q)||^2 \\ &= |\Pi_{R_D}\Pi_{q_1}Q + \Pi_{R_D}\Pi_{q_2}Q)||^2 \\ &= |\Pi_{R_D}\Pi_{q_1}Q||^2 + |\Pi_{R_D}\Pi_{q_2}Q||^2 + 2|q_1R_D^T||q_1Q^T||q_2R_D^T||q_2Q^T| \\ &= P(R_D) + I(Q, R_D, q_1, q_2) \end{aligned}$$

我们将这种文本匹配单元的交互通过**量子干涉项** $I(Q, R_D, q_1, q_2)$ 表示，并在QINM中建模这种信息，以达到更好的实验结果。

03 量子干涉启发的神经匹配模型（QINM）

在量子概率的基础上，我们提出了量子干涉启发的神经匹配模型（QINM），用于自组织检索。具体而言，模型可分为三部分：复合系统构建、文档概率分布表示和相关性预测，如图4所示。

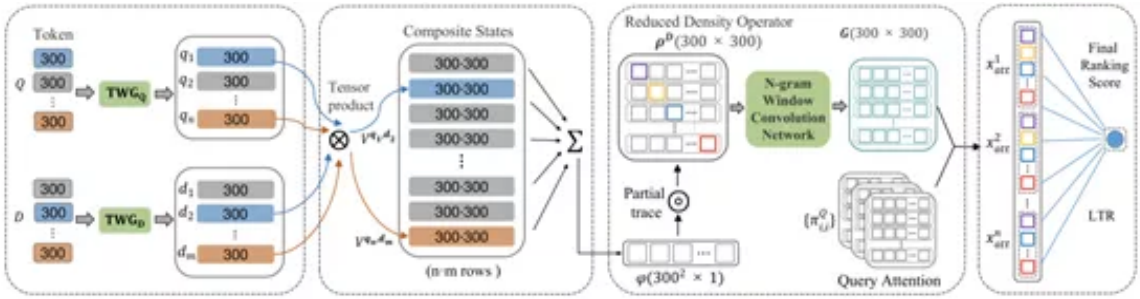

图4 QINM模型结构图

复合系统构建

将查询和文档分别看作向量空间中的两个量子子系统，并使用输入的查询-文档对构造查询-文档复合系统，该系统可以表示为：

$$\varphi = \sum_{i,j}^{n,m}(g_i^Q q_i) \otimes (g_j^D d_j)$$

其中 g_j^D 是查询候选文档集中第 j 个文档项的tf-idf值，可通过图4中的 TWG_d 网络计算； g_i^Q 是可训练参数，通过图4中的 TWG_q 获得，这些权重系可以看作全局匹配信息且满足归一化。

文档概率分布

为了节约计算资源，我们选择使用文档-查询复合系统计算文档子系统的约化密度矩阵，以此建模文档子系统的概率分布。其中，文档子系统的约化密度矩阵 ρ^D 可以表示为：

$$\begin{aligned}\rho^D &= tr_Q(\varphi\varphi^T) \\ &= C^Q(\sum_{i=1}^{m}(g_j^D)^2\Pi_{i,i}^D + \sum_{j,k=1}^{m,m}(g_j^D g_k^D)^2\Pi_{j,k}^D) \\ &= M_S + M_I, (j \neq k)\end{aligned}$$

其中 $tr_Q(\cdot)$ 表示对查询 Q 求偏迹，此操作可以排除复合系统中查询子系统的影响，从而获得文档子系统的概率分布。其中 $C_Q = \sum_{i,j}^{n,n}(g_i^Q g_j^Q)tr(\Pi_{i,j}^Q)$ ，表示查询中匹配单元的交互。 M_S 和 M_I 可以看作约化密度矩阵的两部分： M_S 称为相似特征矩阵，可用于计算某些神经匹配模型（如MP和KNRM）中常用的相似匹配特征； M_I 称为**干涉特征矩阵**，由任意两个不同的文档术语的外积计算，可以应用于文档术语之间交互产生的匹配特征。

相关性预测

使用上一步所获得 ρ^D 的计算文档D的相关性概率 $P'(R_D)$ ：

$$
\begin{aligned}
P'(R_D|q_i) &= (q_i)^T \rho^D q_i = tr(\rho^D \Pi^Q_{i,i}) \\
&= P(R_D|q_i) + I(Q, D, q_i) \\
P'(R_D) &= P(q_1)P'(R_D|q_1) + P(q_2)P'(R_D|q_2) \\
&= P(R_D) + I(Q, D, q_1, q_2)
\end{aligned}
$$

与 $P(R_D)$ 相比， $P'(R_D)$ 所包含的 $I(Q, D, q_1, q_2)$ 可以建模检索过程中的**干涉信息**，该信息包含查询匹配单元之间、查询与文档匹配单元之间的交互信息。

为了提升模型性能，在模型架构中我们使用N-gram窗口卷积网络提取的特征，同时与查询注意力矩阵相乘获得最终的相关性预测：

$$
\begin{aligned}
x^i_{att} &= (g^Q_i)^2 diag(CNN(\rho^D)\Pi^Q_{i,i}) \\
&= (g^Q_i)^2 diag(G\Pi^Q_{i,i}) \\
x_{att} &= x^1_{att} \oplus \ldots \oplus x^n_{att}
\end{aligned}
$$

其中， x^i_{att} 表示与文档 D 的匹配过程中第i个查询单元的匹配特征信息，将所有的 x^i_{att} 进行concat操作获得最后的匹配向量 x_{att} 。 $diag(\cdot)$ 表示对角线元素操作， $G = CNN(\cdot)$ 表示N-gram窗口卷积网络。

最后，使用 MLP 获得最后的相关性得分：

$$
f(x_{att}) = 2 \cdot tanh(W_T \cdot x_{att} + b)
$$

其中， W_T 和 b 是可训练的参数， $tan(\cdot)$ 表示激活函数，根据实验中使用数据集的标签范围，将排序分数限制在-2到2之间。

04 实验结果

实验数据集选取两个TREC数据集：ClueWeb-09-Cat-B和Robust-04。数据集的详细信息如下表所示：

	Robust-04	ClueWeb-09-cat-B
Vocabulary	0.6M	38M
Document Count	0.5M	34M
Collection Length	252M	26B
Query Count	250	150

图5 数据集

我们在实验过程中选取三类不同的baseline与QINM进行对比：以QL和BM25为代表的经典信息检索模型；以MP、DRMM、K-NRM、Conv-KNRM和MIX为代表的神经信息检索模型；以QLM、NNQLM和QMWF-LM为代表的量子信息检索模型。与三类模型相比我们的QINM都取得了比较好的实验结果，如下表所示：

Model Name	ClueWeb-09-Cat-B				Robust-04			
	MAP	NDCG@20	P@20	ERR@20	MAP	NDCG@20	P@20	ERR@20
QL	0.100†	0.224†	0.328†‡	0.139	0.253†‡	0.415†‡	0.369†‡	0.213
BM25	0.101†	0.225†	0.326†‡	0.141	0.255†‡	0.418†‡	0.370†‡	0.220
QLM	0.082	0.164	0.167	0.112	0.103	0.247	0.208	0.193
NNQLM-I	0.089	0.181	0.169	0.128	0.134	0.278	0.237	0.210
NNQLM-II	0.091	0.203	0.216	0.132	0.150	0.290	0.249	0.236
QMWF-LM	0.103†	0.223†	0.237†	0.151†	0.164†	0.314†	0.257†	0.243†
CDSSM	0.064	0.153	0.214	0.117	0.067	0.146	0.125	0.185
MP	0.066	0.158	0.222	0.124	0.189†‡	0.330†‡	0.290†‡	0.207
DRMM	0.113*†‡	0.258*†‡	0.365*†‡	0.142†	0.279*†‡	0.431*†‡	0.382*†‡	0.342*†‡
K-NRM	0.109†	0.273*¶†‡	0.361*†‡	0.153*¶†‡	0.262*†‡	0.407*†‡	0.364*†‡	0.353*†‡
Conv-KNRM	0.121*¶†‡	0.285*¶†‡	0.367*†‡	0.177*¶†‡	0.274*¶†	0.432*†	0.376*†	0.367*¶†
MIX-weight	0.119*¶†	0.297*¶†	0.349*†	0.215*¶†	0.281*¶†	0.438*†	0.383*†	0.372*¶†
QINM	**0.134***¶†‡	**0.338***¶†‡	**0.375***¶†‡	**0.267***¶†‡	**0.294***¶†‡	**0.453***¶†‡	**0.408***¶†‡	**0.396***¶†‡

图6 实验结果

后记

量子信息检索的研究方向自2008年被英国EPSRC重点项目资助以来，国际上一些著名大学例如英国格拉斯哥大学、加拿大蒙特利尔大学、以及天津大学的研究团队，十余年间在此方向上持续耕耘。本文的通讯作者天津大学的张鹏副教授，自2008年远赴英国读博士，就一直致力于探索量子力学应用到信息检索和自然语言处理的可行之路和突破点。

从最开始简单地将量子力学的一些经典实验（例如光子极化实验）类比到信息检索（该工作获得ECIR 2011最佳短论文奖），到后来看到量子语言模型在SIGIR 2013的发表，张鹏带领他的研究生们开始集中研究量子力学在语言建模中的突破点，并与团队侯越先教授一起，在2015年合作发表了IJCAI长文，成功的将量子纠缠建模在语言模型中，该文是量子认知和量子人工智能发表于IJCAI的首篇长文。

近年来，随着深度学习和神经网络的兴起，量子信息检索和量子语言模型也迎来了重要的研究机会。张鹏副教授敏感意识到这一点，在量子力学、神经网络、语言学习三个领域的交叉点上，先后提出了量子语言模型的神经网络版本，即端到端的量子语言模型（AAAI 2018），基于量子力学和神经网络本质联系的量子多体语言模型（CIKM 2018），张量空间的语言模型（AAAI 2019），张量化的多线性注意力模型（NeurIPS 2019），复数域词向量神经网络（ICLR2020）等工作，且均是长文发表或亮点论文。

随着量子力学启发的语言模型逐渐被上述人工智能和机器学习的顶级会议接受，张鹏一直以来思考的信息检索领域的科学问题是，虽然量子干涉等理论观察在人类认知和人类的相关性判断过程中已经表现出一些证据，我们怎么样通过科学的手段，将量子干涉的理论通过数学公式建模在信息检索的文本匹配模型中。针对这一研究问题，张鹏带领两个学生蒋永余和高珲持

续开展研究，虽然先后被三次拒稿，但张鹏一直鼓励两位同学：“这篇论文的创新点很明显，结果也很好，现在的关键问题是审稿人不熟悉量子力学的原理和数学符号，所以咱们不妨从已有的经典模型开始写起，逐步阐述清楚经典模型的问题，明确量子模型的必要性和优势，如果这些都做到了，相信论文总有一天会发表”。然后，张鹏带领两位同学，在2019年除夕夜来临之前，改了很多遍论文，北京理工大学的宋大为教授也帮助改论文，最终将这篇论文投出，且被SIGIR 2020接受。

作为SIGIR时隔七年再次录用的量子信息检索长文，以量子干涉为代表的量子信息检索工作展现出崭新的生命力。今后，我们将在此工作的基础上进一步探讨量子干涉在信息检索、智能问答与对话、视觉对话、多媒体信息处理等领域的实际应用，推动量子人工智能的发展。

本论文由国家重点研发计划、国家自然基金重点/面上项目、欧盟地平线2020计划支持。

往期回顾

01 自然语言处理中的复数词向量

02 Transformer压缩模型——参数少一半同时效果还更好

03 一种量子多体波函数启发的语言建模方法

04 张量空间语言模型

05 基于神经网络的类量子语言模型（NNQLM）

06 量子因果发现—休谟怀疑论的终结

版权说明