A Neural Matching Model based on Quantum Interference and Quantum Many-body System

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

Abstract

In information retrieval (IR), quantum theory (QT) is introduced to model the decision-making mechanism based on human cognition, and try to explain the cognitive bias in human matching process. Meanwhile, quantum many-body language modeling approach provides a novel theoretical basis and mathematical framework for text representation and semantic modeling. In this paper, we introduce a Quantum Interference inspired Neural Matching model (QINM), which is based on the quantum many-body language modeling approach and constructs the interference effect of human cognition in the retrieval process. Experimental results on two benchmark collections demonstrate that our approach outperforms the quantum-inspired retrieval models, and some well-known neural retrieval models in the ad-hoc retrieval task.

1 Introduction

IR system calculates the relevance score of candidate documents and queries to find the optimal retrieval mechanism [9]. Classical probabilistic models [10, 13], dependency-based models [7, 2, 4, 11] and neural matching models [8, 3, 5, 1] are based on the matching idea of law of total probability (LTP) : firstly, they calculate the local relevance matching evidence of each matching unit and documents, and then accumulate these evidences as the final relevance probability prediction. This matching idea ignores the influence of additional evidence generated by the interaction between matching units on the retrieval results. The additional evidence similar to quantum interference based on human cognition cannot be modeled by classical probability, so we introduce quantum interference to model interference information in IR.

The cross research of quantum theory (QT) and IR has made progress in representation optimization and user cognitive interaction [12]. On the one hand, some works [17, 19, 18] focus on the language representation and modeling method based on quantum-many body and tensor network. They propose a general method of construct text representation and explore the essential relationship between tensor network and neural network. Compared with the application of QT in image processing and computer vision, introducing QT to represent text is more consistent with its theoretical properties (e.g., polysemy can be modeled using superposition states). On the other hand, some works focus on the matching and decision-making mechanism based on human cognition in IR. For example, Zuccon and Azzopardi [20] propose a quantum probability ranking principle (QPRP), which encodes quantum interference effects. Wang et al. [14] aim to explore and model the quantum interference effects in users' relevance judgment caused by the presentation order of documents. However, existing work has not modeled interference effects in neural matching models.

*The corresponding auther.

First Workshop on Quantum Tensor Networks in Machine Learning, 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

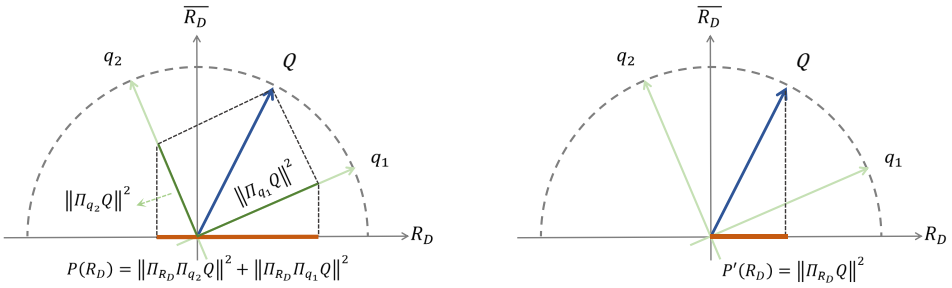
In order to model the interference effects in neural matching models, this paper introduces a Quantum Interference inspired Neural Matching model (QINM) published by SIGIR 2020 [6], which can effectively construct additional matching features provided by interference between matching units. QINM regards a query and its candidate document as a quantum subsystem defined in the vector space, constructs a query-document composite system, and then encodes the probability distribution of a document into the reduced density operator, which is a key step in modeling interference effects. Through an N-gram Window Convolution Network and Query Attention mechanism, we select the effective matching features in the operator. Finally, the ranking score is calculated by the Multi-Layer Perceptron (MLP).

2 Interference Effects via Projection Measurement

Assuming that Query can be represented by a state vector $\boldsymbol{Q} = \alpha \boldsymbol{q}_1 + \beta \boldsymbol{q}_2$, the weight of query term q_1 can be calculated as $P(q_1) = ||\Pi_{\boldsymbol{q}_1} \boldsymbol{Q}||^2$, where projection $(\Pi_{\boldsymbol{q}_1} \boldsymbol{Q})$ means the vector $\boldsymbol{Q}$ projects to basis $\boldsymbol{q}_1$. The conditional probability $P(R_D|q_1) = ||\Pi_{\boldsymbol{R}_D} \boldsymbol{q}_1||^2$, which represents the local interacting between query unit q_1 and document D. The matching pattern of the query unit independent can be expressed as:

$$P(R_D) = P(q_1)P(R_D|q_1) + P(q_2)P(R_D|q_2) = ||\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_1} \boldsymbol{Q}||^2 + ||\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_2} \boldsymbol{Q}||^2 \tag{1}$$

where $P(R_D)$ represents relevance probability between current query and document D. As shown in Figure 1 (a), the projection $(\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_1} \boldsymbol{Q})$ means the process that the query vector $\boldsymbol{Q}$ firstly projects to basis $\boldsymbol{q}_1$, and then projects to basis $\boldsymbol{R}_D$.



(a) Classical probability case

(b) Quantum probability case

Figure 1: Analogy of two neural matching processes in quantum probability and analysis on the existence of interference effect in the document relevance judgment.

However, in the process of document relevance judgment, users usually consider the interaction between text matching units. If the query is considered as a whole, the relevance probability should be calculated as:

$$\begin{aligned} P^{'}(R_D) &= ||\Pi_{\boldsymbol{R}_D} \boldsymbol{Q}||^2 = ||\Pi_{\boldsymbol{R}_D}(\Pi_{\boldsymbol{q}_1} \boldsymbol{Q} + \Pi_{\boldsymbol{q}_2} \boldsymbol{Q})||^2 = ||\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_1} \boldsymbol{Q} + \Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_2} \boldsymbol{Q}||^2 \\ &= ||\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_1} \boldsymbol{Q}||^2 + ||\Pi_{\boldsymbol{R}_D} \Pi_{\boldsymbol{q}_2} \boldsymbol{Q}||^2 + 2|\boldsymbol{q}_1 \boldsymbol{R}_D^T||\boldsymbol{q}_1 \boldsymbol{Q}^T||\boldsymbol{q}_2 \boldsymbol{R}_D^T||\boldsymbol{q}_2 \boldsymbol{Q}^T| \\ &= P(R_D) + I(Q, R_D, q_1, q_2) \end{aligned} \tag{2}$$

where the projection $(\Pi_{\boldsymbol{R}_D} \boldsymbol{Q})$ represents that Q directly project onto basis $\boldsymbol{R}_D$. This process is shown in Figure 1 (b). The $I(Q, R_D, q_1, q_2)$ is named after the interference term, which based on quantum probability to explain the violation of LTP in relevance judgment.

3 The Quantum Interference Inspired Neural Matching Model

This section describes how to model quantum interference terms, and describes Quantum Interference Inspired Neural Matching Model (QINM), as shown in Figure 2. Next, it mainly introduces the core part of the model, namely Document Probability Distribution Representation.

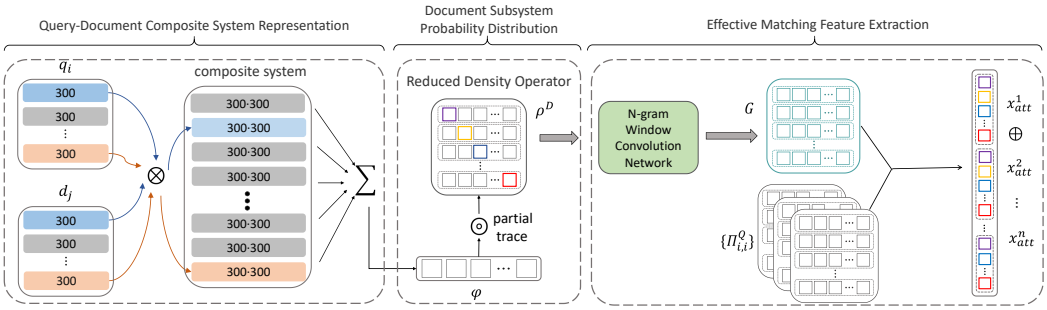


Figure 2: The Document Probability Distribution Representation of the QINM model.

A query is represented as a set of query term vectors denoted by $Q = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_n\}$ and a document is represented as a set of document term vectors, which is denoted by $D = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_m\}$, where $\boldsymbol{q}_i$ and $\boldsymbol{d}_j$ represent the i^{th} query term vector and the j^{th} document term vector, respectively.

Query-Document Composite System Representation Query and document are regarded as two quantum subsystems, and a query-document composite system is constructed by their tensor product. Its state vector is defined as:

$$\boldsymbol{\varphi} = \sum_{i,j=1}^{n,m} (g_i^Q \boldsymbol{q}_i) \otimes (g_j^D \boldsymbol{d}_j) \tag{3}$$

where the state vector $\boldsymbol{\varphi}$ is obtained by tensor product operation $\otimes$. The coefficient g_j^D is the tf-idf value of the j^{th} document term in its query candidate document set, and g_i^Q is a trainable parameter about the i^{th} query term, both of which satisfy the normalization.

Document Subsystem Probability Distribution In composite system, this work applies partial trace operation to calculate the reduced density operator of document subsystem:

$$\begin{aligned} \boldsymbol{\rho}^D &= tr_Q(\boldsymbol{\varphi}\boldsymbol{\varphi}^{\boldsymbol{T}}) \\ &= C^Q(\sum_{i=1}^{m}(g_i^D)^2 \Pi_{i,i}^D + \sum_{j,k=1}^{m,m}(g_j^D g_k^D)\Pi_{j,k}^D) \\ &= M_S + M_I, (j \neq k) \end{aligned} \tag{4}$$

where $tr_Q(\cdot)$ represents the partial trace operation used to obtain the probability distribution of document subsystem in composite system. The coefficient is $C^Q = \sum_{i,j}^{n,n}(g_i^Q g_j^Q) tr(\Pi_{i,j}^Q)$, which indicates the overall interaction between query terms. $\boldsymbol{\rho}^D$ consists of two parts: M_S (similarity feature matrix) and M_I (interference feature matrix) , the former can be used to calculate the similarity matching features in some neural matching models, and the latter is obtained by outer product of any two different document terms, which can be applied to matching features generated by the interaction between document terms.

Further, QINM model calculates the probability of document D related to query Q (i.e., $P^{'}(R_D)$) by applying the document probability distribution $\boldsymbol{\rho}^D$:

$$\begin{aligned} P^{'}(R_D|q_i) &= (\boldsymbol{q}_i)^T \boldsymbol{\rho}^D \boldsymbol{q}_i = tr(\boldsymbol{\rho}^D \Pi_{i,i}^Q) = P(R_D|q_i) + I(Q, D, q_i) \\ P^{'}(R_D) &= P(q_1)P^{'}(R_D|q_1) + P(q_2)P^{'}(R_D|q_2) = P(R_D) + I(Q, D, q_1, q_2) \end{aligned} \tag{5}$$

The specific process is to calculate the relevance probability $P^{'}(R_D|q_i)$ provided by the document probability distribution and each query projection operator separately, and then accumulate the final document relevance probability $P^{'}(R_D)$. $P(q_1) = (g_1^Q)^2$ denotes the importance of the first query term. Compared with Eq. 1, the joint probability $P^{'}(q_i, R_D)$ calculated by Eq. 5 has an extra interference term that can be applied to explain some non-classical phenomena. Meanwhile, compared with Eq. 2, the interference term in the probability $P^{'}(R_D)$ calculated by Eq. 5 is related to the interaction between all query matching units.

Effective Matching Feature Extraction Based on Eq.5 , Query Attention mechanism is proposed for generating matching features:

$$\begin{aligned} \boldsymbol{x}_{att}^{i} &= (g_i^Q)^2 diag(CNN(\boldsymbol{\rho}^D)\Pi_{i,i}^Q) = (g_i^Q)^2 diag(\boldsymbol{G}\Pi_{i,i}^Q), \\ \boldsymbol{x}_{att} &= \boldsymbol{x}_{att}^1 \oplus ... \oplus \boldsymbol{x}_{att}^n \end{aligned} \tag{6}$$

where $\boldsymbol{x}_{att}^i$ denotes the matching feature provided by i^{th} query term in candidate document D, and all the matching features are combined into the final matching tensor $\boldsymbol{x}_{att}$ by concat operation $\oplus$. The $\boldsymbol{G} = CNN(\boldsymbol{\rho}^D)$, where $CNN(\cdot)$ represents the N-gram Window Convolution Network (likely [3] to extract effective features.

4 Experiment

In the experiment, two TREC collections, ClueWeb-09-Cat-B and Robust-04 are selected as datasets, and three types of retrieval models were used as baselines:

• Classical retrieval models : **QL** [16] and **BM25** [10].

• Neural IR models : **MP** [8] , **DRMM** [5] , **K-NRM** [15] **Conv-KNRM** [3] and **MIX** [1].

• Retrieval models inspired by QT : **QLM** [11], **NNQLM** [18] and **QMWF-LM** [19].

Table 1: Comparison of different retrieval models over the ClueWeb-09-Cat-B and Robust-04 collections. ($*$, ¶, §, † and ‡ mean a significant improvement over BM25*, DRMM¶, Conv-KNRM§, NNQLM-II† and QMWF-LM‡ using Wilcoxon signed-rank test p$<$0.05.)

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***¶§†‡

Table 1 presents the performance results of different retrieval models over the two benchmark datasets. We can see that QINM is better than all retrieval models inspired by QT as well as most of the existing neural matching models. For example, on ClueWeb-09-Cat-B topic titles, the relative improvement of our model over the Conv-KNRM is about 1.3%, 5.3%, 0.8% and 9.0%in terms of MAP, NDCG@20, P@20 and ERR@20, respectively. Meanwhile, we find that compared with KNRM and Conv-KNRM, QINM could improve by about 2% on average in the four evaluation indicators, indicating that matching features constructed by QINM could still play a certain role during the process of document relevance judgment in the case of relatively short candidate documents.

5 Conclusion and Future Work

Inspired by language modeling approach based on quantum many-body[19], QINM model is the first to model quantum interference information for neural matching model. The experimental results in ClueWeb-09-Cat-B and Robust-04 show that QINM model achieves significant improvement compared with quantum languages model and neural matching models, which indicates that interference information modeled by QINM can effectively improve the retrieval performance.

In the future, we expect more work to focus on the application of quantum interference in IR and NLP fields, such as modeling quantum interference effects as BERT components. Meanwhile, QT and tensor network have broader application scenarios, such as the application of quantum entanglement in text modeling, using tensor network to the optimization of neural network.

References

[1] Haolan Chen, Fred X Han, Di Niu, Dong Liu, Kunfeng Lai, Chenglin Wu, and Yu Xu. Mix: Multi-channel information crossing for text matching. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 110–119. ACM, 2018.

[2] W Bruce Croft, Howard R Turtle, and David D Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 32–45, 1991.

[3] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134. ACM, 2018.

[4] Joel L Fagan. Automatic p hrase indexing for document retrieval: an examination of syntactic and non-syntactic methods. In *ACM SIGIR Forum*, volume 51, pages 51–61. ACM New York, NY, USA, 2017.

[5] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM, 2016.

[6] Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. A quantum interference inspired neural matching model for ad-hoc retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 19–28, New York, NY, USA, 2020. Association for Computing Machinery.

[7] Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[8] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. A study of matchpyramid models on ad-hoc retrieval. corr abs/1606.04648 (2016). hp. *arxiv. org/abs/1606.04648*, 2016.

[9] Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[10] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer, 1994.

[11] Alessandro Sordoni, Jian Yun Nie, and Yoshua Bengio. Modeling term dependencies with quantum language models for ir. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 653–662, 2013.

[12] Sagar Uprety, Dimitris Gkoumas, and Dawei Song. A survey of quantum theory inspired approaches to information retrieval. 2020.

[13] Cornelis Joost Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 1977.

[14] Benyou Wang, Peng Zhang, Jingfei Li, Dawei Song, Yuexian Hou, and Zhenguo Shang. Exploration of quantum interference in document relevance judgement discrepancy. *Entropy*, 18(4):144, 2016.

[15] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64. ACM, 2017.

[16] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, 2001.

[17] Lipeng Zhang, Peng Zhang, Xindian Ma, Shuqin Gu, and Dawei Song. A generalized language model in tensor space. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[18] Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. End-to-end quantum-like language models with application to question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. A quantum many-body wave function inspired language modeling approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1303–1312. ACM, 2018.

[20] Guido Zuccon, Leif A. Azzopardi, and Keith Van Rijsbergen. The quantum probability ranking principle for information retrieval. In *International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, 2009.