Modeling Multi-query Retrieval Tasks Using Density Matrix Transformation

Qiuchi Li[1,2], Jingfei Li[1], Peng Zhang[1], Dawei Song [*1,3]
[1]School of Computer Science and Technology, Tianjin University, China
[2]Department of Electronic Engineering, Tsinghua University, China
[3]The Computing Department, The Open University, United Kingdom
liqiuchi2015@gmail.com, jingfl@foxmail.com,
{pzhang, dwsong}@tju.edu.cn

ABSTRACT

The quantum probabilistic framework has recently been applied to Information Retrieval (IR). A representative is the Quantum Language Model (QLM), which is developed for the ad-hoc retrieval with single queries and has achieved significant improvements over traditional language models. In QLM, a density matrix, defined on the quantum probabilistic space, is estimated as a representation of user's search intention with respect to a specific query. However, QLM is unable to capture the dynamics of user's information need in query history. This limitation restricts its further application on the dynamic search tasks, e.g., session search. In this paper, we propose a Session-based Quantum Language Model (SQLM) that deals with multi-query session search task. In SQLM, a transformation model of density matrices is proposed to model the evolution of user's information need in response to the user's interaction with search engine, by incorporating features extracted from both positive feedback (clicked documents) and negative feedback (skipped documents). Extensive experiments conducted on TREC 2013 and 2014 session track data demonstrate the effectiveness of SQLM in comparison with the classic QLM.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Relevance feedback, Retrieval Models

Keywords

Quantum Language Model, Session Search, Density Matrix Transformation

1. INTRODUCTION

Recently, various quantum theory (QT) based IR models are developed under the inspiration of the pioneering work of van Rijsbergen [8], which draws a clear connection between the QT and IR. Piwowarski et al. [5] proposed that queries and documents can be modeled as density operators and subspaces respectively, but the tensor space based representation method has not led to good performance. The advent of Quantum Language Model (QLM) [7] , a representative QT-based IR model, successfully solved this issue. In QLM, both single terms and compound term dependencies are represented as projectors in a vector space, while queries and documents are represented as density matrices defining a quantum probability distribution in the space. An EM-based training method for the estimation of density matrix is then devised [7]. The advantages of QLM over traditional language models have been demonstrated from both theoretical and experimental perspectives.

Despite its success in the ad-hoc retrieval, QLM (referred to as classical QLM in the rest of the paper) is solely targeted on single ad-hoc queries. It is insufficient to capture the dynamics of users' information need in response to the user's interaction with the search engine. As a result, it is difficult for the classical QLM to be applied in more complex search tasks, such as multi-query session search.

To address this challenge, we propose to integrate user' short-term interaction information into the estimation of QLM for the current query, and correspondingly a novel Session-based QLM (SQLM) is proposed. The evolution of the user's information need within a search session is modeled by the density matrix transformation, i.e., transforming the original density matrices (for single queries) by some principled rules based on user interactions (e.g., the click and dwell time). We also put forward the concepts of positive projectors and negative projectors extracted from the positive feedback documents (clicked documents) and negative feedback documents (skipped documents), respectively, to enhance the representation ability of the QLM. Specially, a novel training algorithm for QLM with different projectors is devised. Although there exists a body of related work [3][9] for integrating users' interaction information in IR models, they did not model term dependencies in queries and documents, compared with the SQLM proposed in this paper.

2. QLM PRELIMINARIES

2.1 Quantum Probability

In the field of IR, the quantum probability is defined on a real finite space $\mathbf{R}^n$ [7] for simplicity (originally, defined on the infinite Hilbert space). In this paper, we use

*Corresponding Author: Dawei Song, Email: dwsong@tju.edu.cn

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767819 .

the Dirac's notation to represent a unit column vector $\vec{u} \in \mathbf{R}^n$ as $|u\rangle$ and its transpose $\vec{u}^T$ as $\langle u|$, respectively. An *elementary quantum event* can be uniquely represented by a projector onto a 1-dimensional subspace of $\mathbf{R}^n$. For a unit vector $|u\rangle$, the corresponding elementary quantum event, or the projector, is denoted as $|u\rangle\langle u|$. Suppose $|e_1\rangle$,$|e_2\rangle$,...,$|e_n\rangle$ forms an orthonormal basis for $\mathbf{R}^n$, then each unit vector $|v\rangle$ can be uniquely expressed as the *superposition* of $|e_i\rangle$: $|v\rangle = \sum_i v_i|e_i\rangle$, where $\sum_i v_i^2 = 1$.

A measure μ is introduced to define the quantum probability on $\mathbf{R}^n$. It satisfies two conditions: (I) for every projector $|v\rangle\langle v|$, $\mu(|v\rangle\langle v|) \in [0, 1]$ and (II) for any orthonormal basis $\{|u_i\rangle\}$ for $\mathbf{R}^n$, we have $\sum_{i=1}^{n} \mu(|u_i\rangle\langle u_i|) = 1$. The Gleason's Theorem [2] can prove the existence of a mapping function $\mu_\rho(|v\rangle\langle v|) = tr(\rho|v\rangle\langle v|)$ for any vector v given a density matrix $\rho \in S^n$ (S^n is the density matrix space containing all n-by-n positive semi-definite matrices with trace 1, i.e., $tr(\rho) = 1$). Formally, any density matrix ρ assigns a quantum probability for each quantum event in vector space $\mathbf{R}^n$, thereby uniquely determining a quantum probability distribution over the vector space.

2.2 Classical Quantum Language Model

The classical Quantum Language Model(QLM) aims at modeling term dependencies in the principled quantum theory formulation. Different from traditional language models, QLM extracts term dependencies in each document as projectors in the quantum probabilistic space. The single words correspond to projectors $|e_i\rangle\langle e_i|$, and the compound terms (with two or more words for each term) correspond to projectors $|v\rangle\langle v|$ (refer to Section 2.1). The projectors are used to estimate density matrices ρ_q and ρ_d for a query and each document by maximizing a likelihood function with the EM-based iterative approach, i.e., $R\rho R$ algorithm [7]. Then, the top retrieved documents in the initial search results returned by the traditional language model are re-ranked according to the negative VN-Divergence between ρ_q and ρ_d. For details of the classical QLM, please refer to [7].

3. SESSION QUANTUM LANGUAGE MODEL (SQLM)

3.1 Framework

In QLM, a single query can be represented by a density matrix ρ over a vector space for a certain vocabulary. The positive definite matrix with unitary trace can be decomposed as follows:

$$\rho = \sum_{i=1}^{n} \lambda_i(|u_i\rangle\langle u_i|) \tag{1}$$

where $|u_i\rangle$ is a eigenvector, and λ_i is the eigenvalue. Correspondingly, $\Pi_i = |u_i\rangle\langle u_i|$ can be interpreted as an elementary quantum event or projector, and λ_i is the corresponding quantum probability for the elementary event ($\sum_{i=1}^{n} \lambda_i = 1$). By obtaining a density matrix, we actually obtain a set of mutually orthogonal quantum elementary events along with the corresponding discrete probability distribution, and vice versa. In a real search scenario, a user often interacts with the search engine many times before achieving his/her actual information need. We propose a density matrix transformation framework to model the interaction process, which is mathematically formulated as a mapping function T in the density matrix space S^n.

$$T : S^n \to S^n \tag{2}$$

In session search, we assume that there exists an "ideal" transformation T for density matrices which can model dynamic query intention in the historical interactions. Specifically, T is a transformation that for any two consecutive queries q_{i-1} and q_i, the estimated density matrix $\hat{\rho}_i = T\rho_{i-1}$ represents the user's information need for q_i, where ρ_{i-1} is a representation of q_{i-1}. This implies we further make a 1^{st} order Markov assumption that a query is dependent solely on its last previous query. This assumption is reasonable because the dependency can continuously back-propagate in the session.

Theoretically, from Eq.(1), we can easily find that T can be divided into two separate transformation process: $T = T_1T_2$. T_1 is the transformation operator for quantum events $|u_i\rangle$. Since $|u_i\rangle$ forms an orthogonal basis, T_1 can be any standard transition matrix. T_2 changes the original probability distribution for the events (namely, change the values of λ_i), and it is be a diagonal matrix. In this sense, the transformation of density matrix is basically a transformation of main quantum events, and a reallocation of quantum probability for each event.

In practice, however, this consideration seems infeasible due to its high degree of freedom. Suppose a vocabulary $\mathcal{V}$ with $|\mathcal{V}|$ distinct words, T_1 will have a freedom of $O(|\mathcal{V}|^2)$, and T_2 will have a freedom of $O(|\mathcal{V}|)$. Thus the model is prone to be overfiting and computationally expensive. Moreover, it is hard to draw a clear and reasonable connection from the training of T_1 and T_2 to the extraction of projectors. Therefore, we propose an iterative training approach to represent the transformation process, inspired by the updating method of the classical QLM. Specially in this paper, we use the density matrix ρ_{i-1} for query q_{i-1} as the initial density matrix to train the density matrix ρ_i for query q_i.

To facilitate subsequent discussions, we define notations for the session search. In a search session, we have a set of historical interaction units $\{Q_i, R_i, C_i\}_{i=1}^{N-1}$, where Q_i, R_i and C_i represent the query, returned documents and clicked documents for the i^{th} interaction unit respectively. We need to use the historical interaction information to retrieve documents for the current query Q_N. To this end, we first obtain the top N retrieved documents returned by the traditional language model (LM), denoted as R_N. $\{\rho_i\}_{i=1}^{N-1}$ denote a set of $|\mathcal{V}|$-order density matrices representing user's information need for each historical query, where $|\mathcal{V}|$ is the size of the vocabulary containing all distinct terms in the historical queries and the current query.

3.2 Modeling a Single Query

For a historical interaction of a search session, the first clicked document of the query is not always the first one in the search results list. In other words, users often skip some irrelevant results before clicking the first assumingly relevant document. Therefore, we assume that the "skip" behavior is a strong negative feedback signal of users, since the user would have otherwise clicked them. In our assumption, some extreme cases are neglected. For example, the user may gain the right information only by reading the snippets without detecting any click behaviors. Based on this point, we form a positive documents set $R_{pos}(i)$ with all clicked documents as well as a negative set $R_{neg}(i)$ with all skipped documents

in R_i for each query q_i. Note that, $R_{pos}(i)$ is equivalent to C_i, and $R_{neg}(i)$ is null for the queries whose first returned document is clicked.

From $R_{pos}(i)$ and $R_{neg}(i)$, positive projectors $\mathcal{P}_{pos} = \{\Pi_i\}_{i=1}^{M_{pos}}$ and negative projectors $\mathcal{P}_{neg} = \{\Pi_j\}_{j=1}^{M_{neg}}$ are extracted using the method discussed in Section 2.2, where M_{pos} and M_{neg} is the number of positive and negative projectors respectively. Note that some details when extracting projectors: i) only single words, bi-grams and tri-grams are considered as possible compound dependencies, since otherwise the computational complexity will be exponential to the vocabulary size; ii) we use TFIDF to assign the superposition weight v_i rather than the IDF or UNIFORM weights introduced in the original paper [7], since TFIDF is a document specific measure and has better distinguishability across documents. In order to maximize the probability that all positive events happen while all negative events not happen with respect to the quantum probability distribution (i.e., the density matrix ρ), the Maximum Likelihood Estimation (MLE) problem can be formulated as

$$\hat{\rho} = argmax(\sum_{i=1}^{M_{pos}} \log tr(\rho\Pi_i) \sum_{j=1}^{M_{neg}} \log(1 - tr(\rho\Pi_j))) \quad (3)$$

where Π_i and Π_j denote a positive projector and a negative projector. Since

$$\begin{aligned} 1 - tr(\rho\Pi_j) &= tr(\rho(I - \Pi_j)) \\ &= (|\mathcal{V}| - 1) * tr(\rho\widehat{\Pi_j})) \end{aligned} \quad (4)$$

where $\widehat{\Pi_j} = \frac{I - \Pi_j}{|\mathcal{V}| - 1}$ is also a legal density matrix and $|\mathcal{V}| - 1$ is a constant. Then Eq.(3) can be rewritten as

$$\hat{\rho} = argmax(\sum_{i=1}^{M_{pos}} \log tr(\rho\Pi_i) \sum_{j=1}^{M_{neg}} \log tr(\rho\widehat{\Pi_j})) \quad (5)$$

Eq.5 is similar to the objective function in classical QLM. Thus we can apply the similar updating method used in [7] to update the density matrix ρ. Since the $R\rho R$ algorithm in [7] dose not guarantee convergence, we revise it by utilizing the updating method in [4]:

$$\tilde{\rho}_{(m+1)} = (1 - \zeta)\hat{\rho}_{(m)} + \zeta \frac{\hat{\rho}_{(m)} R(\hat{\rho_{(m)}}) + R(\hat{\rho_{(m)}})\hat{\rho}_{(m)}}{2} \quad (6)$$

It can be strictly proved in [4] that for a sufficiently small value of ζ, Eq.(6) guarantees global convergence. Although this updating method guarantees the global convergence theoretically, it requires a sufficiently small value of parameter ζ, resulting in a slow training speed. Therefore, in this paper we do not target on training the density matrix to its convergence, but control an appropriate iterative steps (will be discussed in Section 3.3).

In SQLM, we also model the dwelling time and click sequence for each clicked document. The assumption is that a longer dwelling time and an earlier click mean that the document is more likely to be relevant. Specifically, the weight for a clicked document d is calculated as

$$W_d = e^{\frac{t_d}{t_{all}}} * c^{Seq_d - 1} \quad (7)$$

where t_d is the dwelling time for the document d, t_{all} is the lasting time of the whole interaction, Seq_d denotes the rank of d in the returned document list, and c is a decaying parameter in [0,1], which we will further discuss in Section 4.2. The objective function (3) can therefore be updated as

$$\mathcal{L}_{\mathcal{P}}(\rho) = \sum_{i=1}^{M_{pos}} W_{D(i)} \log tr(\rho\Pi_i) \sum_{j=1}^{M_{neg}} \log tr(\rho\widehat{\Pi_j}) \quad (8)$$

where $D(i)$ is the document containing the projector Π_i. The new objective function is similar to Eq.5, and the only difference is that the new one multiplies each projector in clicked documents by a weight $W_{D(i)}$. Thus the updating methods discussed for Eq.6 can still be applied to the new objective function in Eq.8.

3.3 Density Matrix Transformation

In this paper, we do not train the quantum events transformation operator T_1 and the quantum probability change operator T_2 for density matrix transformation operator T, because of the high freedom. Instead, we propose an iterative training algorithm to approximate the process of density matrix transformation between two subsequent queries:

Algorithm 1 : Density Matrix Transformation.

1: $\rho_0 \leftarrow diag(LM)$; // Initiate the density matrix ρ_0 with the traditional unigram language model.
2: **for** $k = 1$; $k \leq N - 1$; k += 1 **do**
3: Extract projectors from ${R_k}^{pos}, {R_k}^{neg}$ (Section 3.2);
4: Estimate ρ_k with initial density matrix ρ_{k-1} with $TrainingSteps(k) = \mathcal{S}\gamma^{k-1}$ iterative steps;
5: **end for**
6: Return the desired density matrix for interactions ρ_{N-1}.

The training steps are different for different queries, since we believe nearer queries to the current query will have stronger influence on the estimation of current query. The initial training steps $\mathcal{S}$ and the discount factor γ are free parameters which need to be further discussed. The more steps the density matrix is trained, the closer it moves towards the current query density matrix and away from the initial matrix. Thus, gaining an appropriate training steps can achieve a balance between the current query information and historical interaction information.

3.4 Ranking

We use the top K (we set $K = 50$ in this paper) retrieved documents (pseudo feedback documents) returned by traditional LM to train a pseudo feedback QLM ρ_N^p for current query. The representation of user's search intention can be formulated as the linear combination of ρ_{N-1} and ρ_N^p:

$$\hat{\rho} = \alpha\rho_N^p + (1 - \alpha)\rho_{N-1} \quad (9)$$

where α controls the extent to which the history influence on the query representation. After obtaining $\hat{\rho}$, it can be used to re-rank the retrieved documents following the same method in [7].

4. EMPIRICAL EVALUATION

4.1 Experimental Setup

Empirical evaluations are conducted on the TREC 2013 and 2014 session track data shown in Table 1. The corpus

Table 1: Statistics For TREC 2013 and 2014 Datasets (TREC 2014's official ground truth only contains the first 100 sessions).

Items	TREC 2013	TREC 2014
#Sessions	87	100
#Queries	442	453
#Avg. session length	5.08	4.53
#Max. session length	21	11

Table 2: Performance on TREC 2013 and 2014.

TREC2013	NDCG@10	chg%	MAP	chg%
QLM	0.0763	+0.00	0.01708	+0.00
SQLM	0.0847	+11.01	0.01799	+5.32
SQLM+LM	0.0967	+26.74	0.01994	+16.74
TREC2014	NDCG@10	chg%	MAP	chg%
QLM	0.0909	+0.00	0.0164	+0.00
SQLM	0.0950	+4.51	0.0170	+3.66
SQLM+LM	0.1033	+13.64	0.0180	+9.76

used in our experiments is the ClueWeb12 full corpus[1] which consists of 733,019,372 English webpages collected from the Internet. We index the corpus with Indri search engine[2]. In the indexing process, we filtered out all documents with Waterloo's spam scores [1] less than 70, removed the stop words and stemmed all words with Porter stemmer [6].

To verify the effectiveness of the proposed model, we compared the following models: (i) **QLM**, the classic quantum language model which is regarded as the baseline model; (ii) **SQLM**, the proposed session-based quantum language model; (iii) **SQLM+LM**, the combination model of SQLM and traditional language model (LM), which takes the feature of LM into consideration (the linear combination parameter is β). We employ the official evaluation metrics MAP and NDCG@10 to evaluate the models.

A number of parameters are involved in the proposed models, and they are summarized as follows: c in Eq.7, $\mathcal{S}$ and γ in Algorithm 1, α in Eq.9, and β in model $SQLM+LM$. For the global setup, we select $\zeta = 0.01$ in Eq.6. The selection of best parameters will be discussed in next section.

4.2 Results and Discussion

Table 2 reports the experimental results for TREC 2013 and 2014 datasets respectively. In the tables, "chg%" means the improvement percentage over the baseline, i.e., QLM.

Since the modeling process of SQLM only involves matrix addition and multiplication, the computing complexity is low, allowing us to conduct a grid search to find the best parameter configuration. For TREC 2013, the best parameter configuration is $\{c = 0.95, \mathcal{S} = 10, \gamma = 1.05, \alpha = 0.7\}$ for SQLM; and $\{c = 0.95, \mathcal{S} = 30, \gamma = 1.05, \alpha = 0.6 , \beta = 0.9\}$ for SQLM+LM. For TREC 2014, the best parameter configuration is $\{c = 0.95, \mathcal{S} = 30, \gamma = 1.0, \alpha = 0.7\}$ for SQLM, and $\{c = 0.95, \mathcal{S} = 30, \gamma = 1.15, \alpha = 0.9, \beta = 0.9\}$ for SQLM+LM.

The results indicate that the proposed SQLM achieves improvements over the classical QLM, on both TREC 2013 (11.01% improvement for NDCG and 5.33% for MAP), and TREC 2014 Session data (4.51% relative improvements for NDCG and 3.66% for MAP). Moreover, a linear combination of SQLM and LM can further enhance the performance of SQLM, suggesting that SQLM is adaptive to other features such as the traditional LM. It also indicates that SQLM has a large potential for further improvements.

[1]http://www.lemurproject.org/clueweb12/index.php
[2]http://www.lemurproject.org

5. CONCLUSION AND FUTURE WORK

In this paper, we present a novel quantum theory based probabilistic framework for multi-query retrieval task, i.e., session search. By extending the classical Quantum Language Model (QLM), our proposed Session-based Quantum Language Model (SQLM) incorporates the sound mechanism of the density matrix transformation to approximate the dynamics of information need entailed in historical interactions, for re-ranking the initial results generated by the search engine. At the operational level, we utilise the information from both clicked documents and top unclicked documents, and devise a new training algorithm. Extensive experiments on both TREC 2013 and 2014 Session track datasets demonstrate that SQLM does perform better than classical QLM for multi-query retrieval systems, and also show its potential of being further improved for session search.

Therefore, it is safe and reasonable to conclude that the proposed Session-based Quantum Language Model(SQLM) is a feasible expansion of classical Quantum Language Model(QLM) on the multi-query session search tasks. As for future work, we believe that a better retrieval result could be achieved if one can find a better realization of density matrix transformation based on the quantum inference, and incorporate more features into the framework. We will also apply the model to data closer to real-time retrieval systems.

6. ACKNOWLEDGMENTS

This work is supported in part by them Chinese National Program on Key Basic Research Project (973 Program, grant No.2013CB329304, 2014CB744604), the Chinese 863 Program (grant No. 2015AA015403), the Natural Science Foundation of China (grant No. 61402324, 61272265), and the Research Fund for the Doctoral Program of Higher Education of China (grant no. 20130032120044). Any comments from anonymous reviewers are appreciated.

7. REFERENCES

[1] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465, 2011.

[2] A. M. Gleason. Measures on the closed subspaces of a hilbert space. *J. Math. Mech*, 6(6):885–893, 1957.

[3] D. Guan, S. Zhang, and H. Yang. Utilizing query change for session search. In *SIGIR*, pages 453–462. ACM, 2013.

[4] M. G. A. Paris and J. Rehĺćcek. Quantum State Estimation. 649, 2004.

[5] B. Piwowarski, I. Frommholz, M. Lalmas, and K. Van Rijsbergen. What can quantum theory bring to information retrieval. In *CIKM*, pages 59–68. ACM, 2010.

[6] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[7] A. Sordoni, J.-Y. Nie, and Y. Bengio. Modeling term dependencies with quantum language models for ir. In *SIGIR*, pages 653–662. ACM, 2013.

[8] C. J. Van Rijsbergen. *The geometry of information retrieval.* Cambridge University Press, 2004.

[9] S. Zhang, D. Guan, and H. Yang. Query change as relevance feedback in session search. In *SIGIR*, pages 821–824. ACM, 2013.