Quantum Language Model-based Query Expansion

Qiuchi Li
University of Padova
qiuchili@dei.unipd.it

Massimo Melucci
University of Padova
melo@dei.unipd.it

Prayag Tiwari
University of Padova
tiwari@dei.unipd.it

ABSTRACT

The analogy between words, documents and queries and the Quantum Mechanics (QM) concepts gives rise to various quantum-inspired Information Retrieval (IR) models. As one of the most successful applications among them, Quantum Language Model (QLM) achieves superior performances compared to various classical models on ad-hoc retrieval tasks. However, the EM-based estimation strategy for QLM is limited in that it cannot efficiently converge to global optimum. As a result, subsequent QLM-based models are more or less restricted to a limited vocabulary. In order to ease this limitation, this study investigates a query expansion framework on the QLM basis. Essentially, the additional terms are selected from the constructed QLM of top-K returned documents in the initial ranking, and a re-ranking is conducted on the expanded query to generate the final ranks. Experiments on TREC 2013 and 2014 session track datasets demonstrate the effectiveness of our model.

KEYWORDS

Quantum Language Model, Query Expansion, Relevance Feedback

ACM Reference Format:
Qiuchi Li, Massimo Melucci, and Prayag Tiwari. 2018. Quantum Language Model-based Query Expansion. In *2018 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18), September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3234944.3234970

1 INTRODUCTION

Quantum-inspired information retrieval (IR) modeling has been an emerging topic ever since the pioneering work by Van Rijsbergen [7]. Motivated by this work, researchers have been trying to get inspiration from various quantum concepts for addressing IR tasks, such as quantum interference [14], quantum entanglement [9], quantum measurement [13], quantum detection [5] and quantum probability theory [4, 6].

Among them, one of the most successful and well-known models is the Quantum Language Model (QLM) [6]. Inspired by quantum probability theory, QLM incorporates term dependency information into the language modeling of a document. Experiments on ad-hoc retrieval tasks show that QLM has achieved better performances than classical models [6].

However, the training algorithm for the original QLM is not globally convergent [3, 4]. Seeking to minimize the influence of the imperfect algorithm on the final ranking result, later QLM-based works only take a limited vocabulary (mostly query terms) into modeling, which undermines the representation capability of QLM.

Recently, researchers in the IR field have found an earlier work [2] on a globally convergent way of estimating the model. The application of the new training strategy in both IR [8] and sentiment analysis [12] leads to a performance superior to the original QLM, and also shows a very fast speed of convergence for the training algorithm in practice.

Unfortunately, no effort has been made to take into account a larger vocabulary on the basis of this advanced training algorithm, to the author's knowledge. Even though Zhang et al. [11] proposed a quantum-inspired query expansion model called RM-HS, it essentially combined the relevance scores of the original query and expansion terms in a quantum manner. In the quantum-inspired relevance feedback model, Melucci [5] conducted a re-weighting of query terms instead of adding additional terms to the original query. Li et al. [4] attempted to extend QLM for the session search task, but the model only involved query terms in the whole session.

Therefore, we are motivated to explore a query expansion framework for QLM in order to take a larger number of terms into modeling. We first utilize QLM with the advanced training algorithm to generate first-round ranking, and then build a single QLM for top returned results in order to identify the expansion terms. Finally, a QLM with the new training algorithm is applied to the expanded query to generate the final ranking result. The proposed Quantum Language Model-based Query Expansion (QLM-QE) framework achieves better performances than both QLM and other existing quantum-inspired models on two benchmarking datasets.

2 BACKGROUND

2.1 Preliminary of quantum probability theory

Quantum probability theory provides a mathematical formalism for interpreting concepts in quantum mechanics and explaining quantum phenomena. It can also be viewed as a generalization of classical probability theory, with the measurable space being the set of subspaces of an infinite Hilbert space.

In this paper, we use Dirac notation to represent a unit vector $\vec{\mu}$ and its transpose $\vec{\mu}^T$ as a ket $|u\rangle$ and a bra $\langle u|$ respectively. On this infinite Hilbert space $\mathbb{H}$, an *event* Π uniquely corresponds to a projector onto a subspace, which is formulated by a projection matrix (also denoted as Π for convenience). In particular, an *elementary event* corresponds to a 1-dimensional projector $\Pi = |u\rangle\langle u|$ or equivalently a unit vector $|u\rangle$.

Given the definition of events, the probability measure can also be defined on $\mathbb{H}$. Gleason's Theorem [1] proves that a quantum probability measure on $\mathbb{H}$ can be uniquely formulated as a *density matrix* ρ with the following two properties:

I) ρ is square and positive semi-definite, i.e. $\rho \geq 0$

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09.
https://doi.org/10.1145/3234944.3234970

II) ρ has unit trace, i.e. $tr(\rho) = 1$

As a probability measure, the density matrix ρ assigns a probability to every quantum event on $\mathbb{H}$: $pr_\rho(\Pi) = tr(\rho\Pi)$.

2.2 Quantum Language Model

Quantum language model (QLM) is a quantum-like approach to construct the term distribution of a document. In the original QLM, the space is simplified to n-dimensional real space $\mathbb{R}^n$. For a document d, a density matrix ρ_d is estimated as a representation of the document. The estimation procedure has the following steps:

I) **Extracting quantum events.** Elementary quantum events are extracted from both terms and term co-occurrences. For an occurrence of a query term in the document, we extract the one-hot vector as a quantum event; for a co-occurrence of query terms within a window of given length L, we normalize the inverse document frequencies (IDF) of these terms to construct a unit vector for the quantum event. After this step, a set of projectors (i.e. events) $\{\Pi_i\}_{i=1}^{M}$ are collected, where M is the size of the set.

II) **Training density matrix.** Based on the collected quantum events, we try to find the density matrix ρ_d that maximizes the overall log likelihood function $F(\rho)$ of the occurrences of all collected quantum events. In this way, the estimation of ρ_d is converted to a likelihood maximization problem:

$$\begin{aligned}\rho_d &= argmax_{\rho \in S} F(\rho) \\ &= argmax_{\rho \in S} \sum_{i=1}^{M} log(tr(\rho\Pi_i)),\end{aligned} \tag{1}$$

where S is the set of n-by-n density matrices. An iterative algorithm called $R\rho R$ is introduced to find the optimal density matrix ρ_d for the target document d.

III) **Smoothing density matrix.** The trained density matrix ρ_d is smoothed by the diagonal matrix of the collection language model ρ_c in the same way as Dirichlet smoothing [10].

Based on the density matrix of a document ρ_d and the query ρ_q, the query relevance score is given by the negative query-to-document *Von-Neumann* (VN) *divergence*:

$$score(q, d) = -tr(\rho_q(log\rho_q - log\rho_d)), \tag{2}$$

where the logarithm of a matrix is obtained by taking the logarithm of its eigenvalues, whilst keeping its eigenvectors the same.

The proposed QLM achieves better performance than various classical retrieval models [6]. However, it is limited in that it only manages to estimate the quantum probability distribution of query terms, but fails to consider non-query terms which might also be crucial to the task. This is largely due to the problem that the $R\rho R$ training algorithm does not guarantee global convergence [2–4].

Goncalves et al. [2] proposed a globally convergent algorithm for training the density matrix to address this theoretical limitation. At the k-th iteration of this iterative algorithm, it first computes the R operator, which is the derivative of the target log likelihood function (1) with respect to the density matrix ρ:

$$R(\rho^k) = \sum_i \frac{\Pi_i}{tr(\rho^k\Pi_i)} \tag{3}$$

Then, it computes the two possible ascent directions $\bar{D}$ and $\tilde{D}$ of the target function $F(\rho)$:

$$\bar{D}(\rho^k) = \frac{R(\rho^k)\rho^k + \rho^k R(\rho^k)}{2} - \rho^k \tag{4}$$

$$\tilde{D}(\rho^k) = \frac{R(\rho^k)\rho^k R(\rho^k)}{tr(R(\rho^k)\rho^k R(\rho^k))} - \rho^k \tag{5}$$

Now the search direction D can be expressed as a combination of $\bar{D}$ and $\tilde{D}$:

$$D(\rho^k) = \frac{2}{q(t_k)}\bar{D}(\rho^k) + \frac{t_k tr(R(\rho^k)\rho^k R(\rho^k))}{q(t_k)}\tilde{D}(\rho^k), \tag{6}$$

where $q(t_k) = 1 + 2t_k + t_k^2 tr(R(\rho^k)\rho^k R(\rho^k))$, and $t_k \in [0, 1]$ is the steplength controlling how much the density matrix moves towards the search direction. It is chosen such that $F(\rho^k + t_k D(\rho^k)) > F(\rho^k) + \gamma t_k tr(R(\rho^k)D(\rho^k))$, i.e. to guarantee the updating scheme does increase the value target function. Here the learning rate γ determines the minimum leap of the target value at each iteration, and controls the speed of learning by directly influencing the step length t_k. We fix $\gamma = 0.0001$ in this study.

Finally the new density matrix is updated by the formula below:

$$\rho^{k+1} = \rho^k + t_k D^k \tag{7}$$

Given a random legal initial density matrix ρ^0, the algorithm iteratively updates the density matrix using equation 3-7 until it reaches a steady point ρ^*, which is the optimal density matrix in this algorithm.

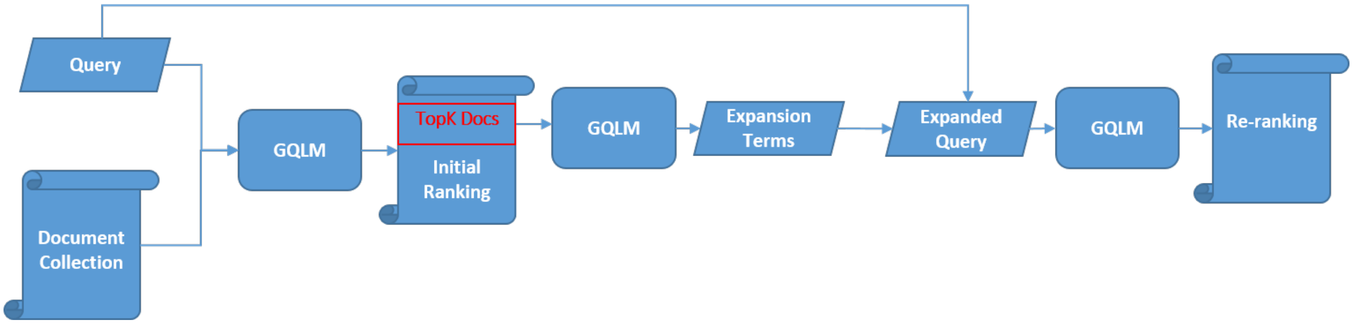
Replacing the $R\rho R$ algorithm with the globally convergent training strategy in theory leads to a better way of constructing the QLM. In practice, recent works [8, 12] have demonstrated that the new training algorithm does increase the performance on an application context. Furthermore, Zhang et al. [12] found that the new algorithm converges within 17 iterations. In this paper, we follow [12] to call it Global-convergence based Quantum Language Model (GQLM).

3 A QUANTUM LANGUAGE MODEL-BASED QUERY EXPANSION FRAMEWORK

This section introduces our proposed Quantum Language Model-based Query Expansion (QLM-QE) framework. Existing QLM-based models [4, 8, 12] are more or less limited to modeling only query terms, and a quantum-inspired approach for automatically enlarging the vocabulary is absent. The rapid convergence and superior performance of the globally convergent training algorithm provides a feasible way to involve non-query terms into the quantum language modeling process. Therefore, in this study, exploratory research is conducted into a query expansion framework on the basis of GQLM.

Figure 1 shows the diagram of the QLM-QE framework. It consists of the following three main steps:

Figure 1: Our QLM-QE Framework.



I) **Initial ranking.** For a given query, a GQLM is employed to generate the initial ranking, in which negative VN-divergence is used to compute the query relevance score. Here only the query terms are taken into consideration, so the density matrix for both the query and documents is of size Q^*Q, where Q denotes the vocabulary size of the query.

II) **Identifying expansion terms.** A pseudo relevance feedback is conducted to identify the additional terms that should be added to the original query. Assuming the top K returned documents are relevant to the information need of the user, we treat them as a whole and apply GQLM to build a single density matrix ρ_{top}. Then the top V non-query terms with the highest quantum probabilities, or the greatest corresponding diagonal values, are selected from ρ_{top} as the expansion terms. This expansion term selection method differs from taking the most frequent terms directly from the top documents in that quantum probabilities are computed from both single term occurrence and co-occurrences among the terms. In this way, a term with a high frequency and low co-occurrence with other terms may have a lower quantum probability than a term with a low frequency but a higher co-occurrence frequency with other terms in the vocabulary. In order to avoid unnecessary computations in estimating the document density matrix, we pre-select top W terms with the highest TF-IDF values, making ρ_{top} to be of size $W * W$.

III) **Re-ranking with expanded query.** We finally re-rank the documents with the expanded query by using another GQLM. There is only one difference from the normal GQLM procedure: the density matrix for the expanded query ρ_{exp} is estimated from the projector set $\{\Pi_{exp}\}$, which is the combination of the projectors extracted from the original query $\{\Pi_q\}$ and the ones extracted from top K returned documents $\{\Pi_{top}\}$: $\{\Pi_{exp}\} = \lambda\{\Pi_q\} \bigcup(1-\lambda)\{\Pi_{top}\}, \lambda \in [0, 1]$.

4 EXPERIMENT

The experiments were conducted on TREC 2013 [1] and 2014 [2] Session Track datasets because they have gradually been becoming the benchmarking datasets for Quantum-inspired models in recently years. Table 1 shows the statistics for the datasets.

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

The document collection for the two datasets is the Clueweb12 Category B collection (Clueweb-B) [3] which contains more than 50 million English webpages collected from the Internet. We removed the documents that do not belong to Clueweb-B from both the query logs and the ground truth. The Clueweb-B collection was indexed by using Indri [4] 5.11. The queries and documents were preprocessed following the same standard procedure. In particular, we usec Porter Stemmer to stem words.

In order to test the performance of QLM-QE, we compared the MAP@10 and nDCG@10 metrics of the following models in the experiments:

I) Unigram, which is the baseline model.
II) RM-HS [11], which is a query expansion approach inspired by quantum interference.
III) QLM [6].
IV) QMT [8], which is a quantum-like session search model inspired by the Two-State Vector Formalism (TSVF).
V) QLM-QE proposed in our study. In order to fairly compare with other models, we employ a simple strategy to combine the current query with the historical queries to form the original query: the projectors are obtained by combining the historical query projectors $\{\Pi_{hist}\}$ and current query projectors $\{\Pi_{cur}\}$ controlled by a linear parameter $\alpha \in [0, 1]$: $\{\Pi_{orig}\} = \alpha\{\Pi_{cur}\} \bigcup(1-\alpha)\{\Pi_{hist}\}$. For the selection of parameters, we take top $K = 10$ returned documents in the first-round query and construct ρ_{top} with top $W = 50$ terms to identify top $V = 10$ additional non-query terms as expansion terms. A grid search on the window length L and

[1] https://trec.nist.gov/data/session2013.html
[2] https://trec.nist.gov/data/session2014.html
[3] https://lemurproject.org/clueweb12/
[4] https://www.lemurproject.org/indri/

two linear interpolation weights α, λ is conducted to find the optimal parameter combination.

All the experiments were implemented in Python 3.6.5 on a Desktop with Intel Xeon CPU E5-1650 v3 @ 3.50GHz, 16GB RAM and Window 10 as the OS. All the codes are open sourced[5].

5 RESULTS AND DISCUSSION

Table 2: Performance on TREC 2013 and 2014 session track datasets in percentages (%). The values in the parentheses (in percentage(%)) are the improvements over the baseline Unigram model. The MAP@10 for RM-HS is missing because it is not reported in the original paper [11].

Model	TREC 2013		TREC 2014	
	nDCG@10	MAP@10	nDCG@10	MAP@10
Unigram	6.05(0.00)	4.91(0.00)	14.22(0.00)	14.52(0.00)
RM-HS	6.00(-0.82)	-	13.93(-0.29)	-
QLM	6.70(10.77)	6.14(24.65)	14.27(0.38)	14.52(0.36)
QMT	8.88(46.77)	8.39(70.83)	14.89(4.71)	13.09(-9.76)
QLM-QE	**10.37(71.40)**	**8.94(89.81)**	**15.19(6.82)**	**14.79(1.86)**

Table 2 shows the performances of all five models on TREC 2013 and 2014 session track datasets. The values in bold denote the best-performed values out of the five models on a single dataset. For significant test, we conducted a paired t-test with 0.05 as the acceptance threshold on a pair of performance values. The optimal parameters for QLM-QE are $\{\alpha = 0.4, \lambda = 0.8, L = 8\}$ for session track 2013 and $\{\alpha = 0.6, \lambda = 0.8, L = 8\}$ for session track 2014.

First of all, QLM-QE occupies all of the four best-performing values among the five models. This means that our proposed framework is more effective than the classical Unigram model as well as various existing quantum-based IR models.

Furthermore, QLM consistently outperformed the original QLM on both datasets, which indicates that the query expansion mechanism in our proposed framework brings benefits to the QLM, and expansion terms have positive influences on the final ranking.

Last, it is worth noting that quantum-inspired models (except RM-HS) perform much better than classical models on session track 2013, but they have similar performances on session track 2014. The authors conjecture that it is likely to be because session track 2013 is more "quantum" in nature. However, if one can measure how much "quantumness" there is in the data in the first place, then he may determine which type of models are more suitable for this task. We believe the detection of quantum phenomena among data could be a very crucial point of study for future works in the relevant research fields.

6 CONCLUSION

This paper successfully addresses the problem of limited vocabulary size for a well-known quantum-inspired IR model called Quantum Language Model (QLM), by developing a novel query expansion framework on the basis of an improved training algorithm for QLM.

Despite the observed performance improvements over the other models under experiment, additional experiments could be conducted to reach a more concrete conclusion in the future. On the one hand, we plan to compare our proposed model with a QLM plus classical query expansion approach, which can better demonstrate the effectiveness of the quantum query expansion framework. On the other hand, the results would be more convincing by experimenting on web track or ad-hoc retrieval datasets, where the influence of historical interactions on the current query are completely eliminated.

ACKNOWLEDGEMENT

This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie SkÅĆodowska-Curie grant agreement No. 721321.

REFERENCES

[1] ANDREW M. GLEASON. 1957. Measures on the Closed Subspaces of a Hilbert Space. *Journal of Mathematics and Mechanics* 6, 6 (1957), 885–893. http://www.jstor.org/stable/24900629

[2] D. S. Goncalves, M. A. Gomes-Ruggiero, and C. Lavor. 2013. Global convergence of diluted iterations in maximum-likelihood quantum tomography. *arXiv:1306.3057 [math-ph]* (June 2013). http://arxiv.org/abs/1306.3057 arXiv: 1306.3057.

[3] Zdenek Hradil, Jaroslav Rehacek, Jaromir Fiurasek, and Miroslav Jezek. [n. d.]. 3 Maximum-Likelihood Methodsin Quantum Mechanics. In *Quantum State Estimation*. Springer, Berlin, Heidelberg, 59–112. https://doi.org/10.1007/978-3-540-44481-7_3

[4] Qiuchi Li, Jingfei Li, Peng Zhang, and Dawei Song. 2015. Modeling Multi-query Retrieval Tasks Using Density Matrix Transformation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 871–874. https://doi.org/10.1145/2766462.2767819

[5] M. Melucci. 2016. Relevance Feedback Algorithms Inspired By Quantum Detection. *IEEE Transactions on Knowledge and Data Engineering* 28, 4 (April 2016), 1022–1034. https://doi.org/10.1109/TKDE.2015.2507132

[6] Alessandro Sordoni, Jing He, and Jian-Yun Nie. 2013. Modeling Latent Topic Interactions Using Quantum Interference for Information Retrieval. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 1197–1200. https://doi.org/10.1145/2505515.2507854

[7] Cornelis Joost Van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.

[8] Panpan Wang, Yuexian Hou, Jingfei Li, Yazhou Zhang, Dawei Song, and Wenjie Li. 2017. A quasi-current representation for information needs inspired by Two-State Vector Formalism. *Physica A: Statistical Mechanics and its Applications* 482 (Sept. 2017), 627–637. https://doi.org/10.1016/j.physa.2017.04.145

[9] Mengjiao Xie, Yuexian Hou, Peng Zhang, Jingfei Li, Wenjie Li, and Dawei Song. 2015. Modeling Quantum Entanglements in Quantum Language Models. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, Buenos Aires, Argentina, 1362–1368. http://dl.acm.org/citation.cfm?id=2832415.2832439

[10] ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies* 1, 1 (2008), 1–141.

[11] Peng Zhang, Jingfei Li, Benyou Wang, Xiaozhao Zhao, Dawei Song, Yuexian Hou, and Massimo Melucci. 2016. A Quantum Query Expansion Approach for Session Search. *Entropy* 18, 4 (April 2016), 146. https://doi.org/10.3390/e18040146

[12] Yazhou Zhang, Dawei Song, Xiang Li, and Peng Zhang. 2018. Unsupervised Sentiment Analysis of Twitter Posts Using Density Matrix Representation. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Vol. 10772. Springer International Publishing, Cham, 316–329. https://doi.org/10.1007/978-3-319-76941-7_24

[13] Xiaozhao Zhao, Peng Zhang, Dawei Song, and Yuexian Hou. 2011. A novel re-ranking approach inspired by quantum measurement. In *Advances in Information Retrieval*. Springer, 721–724.

[14] Guido Zuccon and Leif Azzopardi. 2010. Using the quantum probability ranking principle to rank interdependent documents. In *Advances in information retrieval*. Springer, 357–369.

[5] https://github.com/https://github.com/qiuchili/QLM-QE