The Pseudo-relevance Feedback Model Based on Quantum Probability Theory

Yueheng Sun, Chenjun Zou
School of Computer Science and Technology, Tianjin University
Tianjin, China
yhs@tju.edu.cn

***Abstract*—Relevance Model (RM) is one of typical and generally stable methods for the query expansion in information retrieval (IR). This paper presents a novel information retrieval model based on the quantum probability theory, and makes a preliminary exploration on the application of quantum model in pseudo-relevance feedback. In the document weight allocation framework, we propose two re-ranking approaches based on linear weight allocation and quantum interference weight allocation, respectively. The experimental results on standard TREC datasets show that the proposed model is effective and potential in information retrieval tasks.**

Keywords- information retrieval; pseudo relevance feedback; quantum probability theory; document weight allocation

I. Introduction

Information retrieval is the task of seeking information that meets the user needs from a large number of unstructured data collections [1] (such as computer-stored text messages). In most situations, the users are reluctant to provide explicit information, which makes explicit feedback become difficult to implement. Although the information system can analyze the user click records to obtain the implicit feedback information, but some click records are unable to accurately reflect users' intention.

Pseudo relevance feedback (PRF) [6] is a simple and effective method for dealing with the difficulties mentioned above. The method assumes that the top *n* documents in the initial retrieval are relevant documents, and uses them to refine the retrieval results. Relevance model is a stable and efficient PRF model and it is often used to build query expansion system. However, in practice, relevance model has two drawbacks; the first is document scores integrate the language-based query likelihood. But experiments show that query similarity score is exponential distributed, which can cause severe problems in query expansion. The second is relevance model ignores document dependency. The similarity between documents reflects the cluster assumption of feedback documents. Although certain document may rank low after the initial retrieval while it is clustered with high-ranking documents, on the cluster assumption, it is more confident to believe it is a relevant document. Therefore, in the context of PRF, document weight allocation becomes the key to solving the above problems.

In this paper, quantum-based document weight allocation is proposed, and it is based on quantum information retrieval model [7]. Unlike document ranking work in [7], it extends to the smoothing of ranking scores. This paper presents a preliminary application of quantum theory in the document relevance feedback, and shows its potential research value in the field of information retrieval.

The rest of this paper is organized as follows. Section 2 introduces the document relevance model and its existing problems in weight allocation. In section 3, document weight allocation framework is proposed based on linear weight allocation and quantum inference weight allocation, respectively. Experimental results and analysis are presented in section 4. Finally, we give concluding remarks in section 5.

II. Background and Related Works

A. Pseudo-relevance Feedback and Relevance Model

In ad-hoc document retrieval, pseudo-relevance feedback is often used in estimating the model of a certain topic. Relevance model is the typically stable and effective method among PRF methods, and is widely used in the field of query expansion. Ponte & Croft introduced the language model in the field of speech recognition to IR [3]. The aim of the language model is to establish a probability distribution for every word sequence appeared in a document. In ad-hoc retrieval, every document has a statistical language model and is regarded as a sample of the relevance model.

A Query can be regarded as the process of generating query sequence from corresponding relevance model. Then documents can be ranked based on the probability that language model of each document generates the given query. Lavrenko & Croft combined traditional probability models with probabilistic language model [2], and proposed a new method to estimate relevance model (RM).

More specifically, the relevance probability is estimated by equation (1) for a given query.

$$p(w\mid R)=\sum_{D\in M}p\left(w|D\right)\frac{p(q\mid D)p(D)}{\sum_{D'\in M}p(q\mid D')p(D')}\tag{1}$$

Where *M* denotes the set of feedback documents, *p*(*D*) is the prior probability, and *p*(*q*|*D*) is the query likelihood of document *D*, defined as equation (2):

$$p\left(q\mid D\right)=\prod_{i}^{m}p(q_i\mid D)\tag{2}$$

B. Document Weight Allocation

As mentioned above, the relevance model has two drawbacks: first, the document weights are exponentially distributed; second, the document similarities are ignored in the ranking process. In this section, we analyze the distribution of the document weight.

In retrieval systems, the prior probability $p(D)$ is usually assumed uniform. Combined with Equation (1), document weight is given as follows:

$$f\left(D,q\right)=p\left(q\mid D\right)/\sum_{D^{'}\in M}p(q\mid D') \tag{3}$$

Experimental results show that $f(D,q)$ is not properly smoothed. Firstly, the document weights decrease rapidly in the top k documents ($k<<n$). For instance, in table 1, for all three TREC queries (Topic 151, 152 and 153), the weight of doc 1 is 2 times of doc 2 and 3 times of doc 4. However, all three top ranked documents are irrelevant (all with very large weight) which could hurt the retrieval performance. Moreover, this would cause topic drifting problem.

TABLE I. TOP 4 DOCUMENT WEIGHT DISTRIBUTION

Query	$f(D,q)/r$	$f(D,q)/r$	$f(D,q)/r$	$f(D,q)/r$
#151	0.190/0	0.155/1	0.097/1	0.058/0
#152	0.131/0	0.084/1	0.073/0	0.064/1
#153	0.226/0	0.181/1	0.101/1	0.087/1

Secondly, the top k documents account for most of the total weights, which makes low-ranking documents insignificant based on equation (1). For instance, in table II, top 5 documents occupy nearly half of the document weights; top 10 documents occupy nearly 60% of the document weights.

TABLE II. THE TOP *N* DOCUMENTS WEIGHT PERCENTAGE

Data set	Query terms	Top5 sum(*f*)	Top10 sum(*f*)
AP8889	151-200	0.466	0.610
WSJ8792	151-200	0.529	0.668
ROBUST2004	601-700	0.525	0.665

The above empirical analysis illustrate that document weight distribution is not well smoothed (an exponential-like distribution).

The exponential-like distribution reduces retrieval performance, especially for the query expansion task. In this paper, we propose to re-allocate document weights in the framework of PRF via quantum theory. More specifically, we propose the quantum interference weight allocation (QIWA) model to smooth the document weights by considering the dependencies among all feedback documents.

III. PREPARE YOUR PAPER BEFORE STYLING

A. An Brief Introduction of Quantum Interference

In this paper, we propose the quantum interference method to simulate the PRF process. We analyses the classical physical experiment: Young's double-slit experiment. In figure 1, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen. The detection of individual photons is observed to be inherently probabilistic, so we define $p_1(x)$ to stand for the probability of a photon observed in x when slit 1 is shaded, $p_2(x)$ is defined almost the same as $p_1(x)$ except slit 2 is shaded, and $p_{12}(x)$ when neither slit is shaded, experiment shows that $p_{12}(x)\neq p_1(x)+p_2(x)$, which is inexplicable using classical mechanics.

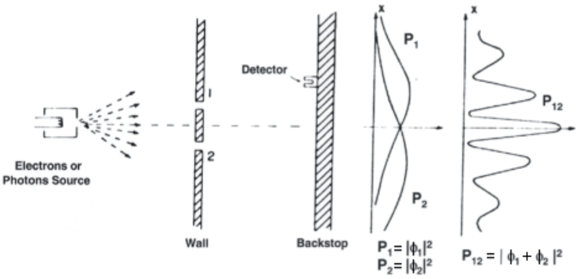

Figure 1. The double-slit interference experiment

Apart from classical mechanics, quantum probabilistic theory regards certain quantum status as the parallelization of bases; all the bases can interfere with each other. To describe the quantum behavior, the probability in classical events is replaced by complex probability amplitude. In practice, quantum events is the linear combination of the bases defined as {1,2,...,D}. So any quantum status can be represented as $\left\{\alpha_1,...,\alpha_D\right\}\in C^D$ and the probability of base i is observed as $|\alpha_i|^2$. Specifically, φ_1 stands for the complex probability amplitude when slit 2 is shaded, while φ_2 is alike except slit 1 is shaded, and φ_{12} stands for the amplitude when neither is closed. We have the following equation:

$$\begin{aligned}p_{12}(x)&=\left|\varphi_1(x)+\varphi_2(x)\right|^2\\&=\left|\varphi_1(x)\right|^2+\left|\varphi_2(x)\right|^2+(\varphi_1(x)\varphi_2(x)^*+\varphi_1(x)^*\varphi_2(x))\end{aligned} \tag{4}$$

$p_{12}(x)$ = $p_1(x)$ + p_2(x) + l_{12}(x). The right part of the formula is the sum of classical probability theory and the quantum interference incremental l_{12}(x). In summary, the classical probability theory probability formula does not apply to the double-slit interference experiment. In quantum probability model, the joint event is equal to the sum of complex probability amplitude of all possible single events.

B. Quantum Interference and IR

Quantum theory has an intrinsic connection with IR, because users' information needs are often unclear. They know their real information needs after certain relevant webpages are listed. In norm of quantum physics, the process is called "collapse". We regard the document collection as the light source, and the process of the initial retrieval as light source passing through *N* slits. Relevance model [2] is the retrieval model based on the conventional probability theory. As aforementioned, relevance model ignores the document dependencies, and the total weight is equal to the sum of all the document weights.

Taking document similarities into consideration, in the initial retrieval, documents passing through *N* slits of plate exhibit the phenomenon of quantum interference, i.e. there exists certain dependencies among documents (based on the cluster assumption in [1]), which makes the joint probability not equal to the sum of complex probability amplitude of all possible single event. In quantum nature, the formal definition is given below:

Definition 1: Feedback document can be represented as $\varphi = \alpha * |1> + \beta * |0>$, where $\alpha, \beta \in C$, |1> and |0> represents relevance base status, and $p(d \mid R) = |\alpha|^2, 1 - p(d \mid R) = |\beta|^2$. $p(d \mid R)$ denotes the probability that doc *d* is relevant.

Definition 2: Quantum interference incremental of feedback document φ_1 and φ_2:

$$QII(\varphi_1, \varphi_2) = \alpha_1(x) * \alpha_1(x)^* + \alpha_2(x) * \alpha_2(x)^*$$
$$= 2|\alpha_1(x)||\alpha_2(x)|\cos(\theta_1 - \theta_2)$$

Definition 3: Total weight of the relevance probabilities of feedback documents is:

$$SumWeight = \left|\sum \alpha_i\right|^2$$
$$= p_1 + p_2 + \ldots + p_N + \sum_{i,j<N} QII(\alpha_i, \alpha_j)$$

C. Document Weight Allocation Framework

Compared with relevance model, this paper introduces the quantum interference incremental into relevance model, and extends the total document weight in accordance with the similarities between the documents. In initial retrieval, relevance model is a good estimation for user's information need. However, if the document score is directly used for query expansion, it will hurt the performance to some extent.

Based on above observations, we propose a document weight allocation framework as follows:

- Initial retrieval based on language model in [3]. The top *N* documents are used as pseudo-relevant document set, denoted as *M*.
- Estimate the language model using the TF-IDF weight for each feedback document *D* in the document set *M*.
- Calculate the document similarity matrix *S*. Cosine similarity is adopted as document similarity measurement. There are two considerations: On one hand, language model is in the form of vector space representation, using the cosine is straightforward and generally achieves good performance. On the other hand, the cosine value can be used as an estimation of the weight of quantum interference (definition 2), which is necessary in QIWA model (see section *D*).
- Take top *k* document in pseudo relevant document collection *M* as real relevant document collection M_t.
- Allocate *scores* using weight allocation methods (LWA and QIWA).
- Calculate the *k* nearest neighbors for each documents in M_t based on document similarity matrix *S*. According to the cluster assumption in [1], if a feedback document is not in *k*-nearest neighbor documents set, it will be treated as irrelevant and is neglected in the Re-ranking process.
- Re-rank all the feedback documents based on the new document weights. Mean Average Precision (MAP) is adopted as evaluation metric.

D. Evaluation Methods for the Final Retrieval Results

1) Linear weight allocation (LWA)

For the top *k* documents, i.e. the real relevant document collection M_t, their weight occupies most of the total weight. Based on this observation, we propose a linear document weight allocation, i.e. we re-assign the weight of a document based on its neighbors (the documents with high similarity) in M_t. The new document weight is calculated as follows:

$$\tilde{p}(d \mid R) = \frac{1}{z} \sum_{dt \in M_t} (1 - sim(d, d_t)) * p(d \mid R) + sim(d, d_t) * p(d_t \mid R) \quad (5)$$

Where *sim*(*d*,d_t) represents the document similarity, $p(d \mid R)$ and $p(d_t|R)$ denotes the original weight of documents, and 1/*z* is a normalization factor.

2) Quantum interference weight allocation (QIWA)

Pseudo relevance feedback can be simulated by the phenomenon of quantum interference. We regard the document collection as light source, and the process of the initial retrieval as light source passing through *N* slits. Due to the inherent similarities between documents, the interference between documents has practical significance. Definition 3 gives the formal description of the total weight of Pseudo-relevant document sets. Now the key question is how to allocate the weight of quantum interference incremental into each document.

In order to simplify the calculation in practice, we have ignored quantum interference between two low ranking documents because the weight is insignificant. For high-ranking documents, the scores are usually high and decrease exponentially. Therefore, only the quantum interferences between the given document and documents in M_t are taken into consideration. Another simplification is that the cosine value of two document vectors is used to estimate the cosine of difference of amplitude angles between two documents.

After the simplification, quantum interference model is formalized as follows:

$$\tilde{p}(d|R) = \frac{1}{z} \sum_{d_t \in M} \sqrt{p(d|R)} \sqrt{p(d_t|R)} CosSim(d, d_t) * p(d_t \mid R) \quad (6)$$

CosSim(*d*,*dt*) denotes the cosine similarity between document *d* and d_t.

IV. Experimental Design and implementation

A. Experimental Data

In this paper, we evaluate the effectiveness of the proposed document weight allocation methods described above on standard TREC dataset. Specifically, three TREC datasets are used in this experiment: WSJ (Topic 151 to 200, 173252 docs), AP8889 (Topic 151-200, 164597 docs), and ROBUST2004 (Topic 601-700, 528155 docs). The Lemur toolkits are used to do indexing and stemming. For Language model, Dirichlet smooth Method in [5] is adopted with μ=700. The retrieval scores are calculated using KL model in [1].

B. The Comparison of LWA & QIWA

The relevance model is used as baseline method. For comparison, two weight allocation methods (LWA and QIWA) are tested in the re-ranking framework proposed in section III-C, and the feedback document numbers are set to 30, 50, 70 and 90, respectively. The test results are summarized as follows:

TABLE III. RELEVANCE MODEL WITH 30 FEEDBACK DOCUMENTS

MAP%(chg%)	Baseline	LWA	QIWA
AP	14.92	15.95(+6.85)	16.09(+7.85)
ROBUST	20.69	21.67(+4.73)	22.11(+6.84)
WSJ	17.68	18.76(+6.09)	19.13(+8.22)

TABLE IV. RELEVANCE MODEL WITH 50 FEEDBACK DOCUMENTS

MAP%(chg%)	Baseline	LWA	QIWA
AP	18.49	20.04(+8.42)	20.32(+9.94)
ROBUST	22.78	24.08(+5.72)	24.61(+8.04)
WSJ	21.57	23.34(+8.21)	23.60(+9.40)

TABLE V. RELEVANCE MODEL WITH 70 FEEDBACK DOCUMENTS

MAP%(chg%)	Baseline	LWA	QIWA
AP	20.61	22.55(+9.40)	22.82(+10.72)
ROBUST	24.28	26.05(+7.30)	26.65(+9.75)
WSJ	23.71	25.72(+8.48)	26.26(+10.78)

TABLE VI. RELEVANCE MODEL WITH 90 FEEDBACK DOCUMENTS

MAP%(chg%)	Baseline	LWA	QIWA
AP	22.39	24.31(+10.23)	24.50(+11.59)
ROBUST	25.27	27.38(+8.33)	27.99(+10.75)
WSJ	25.03	27.00(+7.87)	28.04(+12.04)

C. Result Analysis

As shown in Table III-VI, both two allocation methods significantly outperform the baseline (RM). This indicates that the refined document weights after re-allocation are more reasonable.

Compare with LWA, QIWA shows superior performance in all cases. Although LWA can also give more smoothing document weights, QIWA has more expressive power to characterize the dependencies between feedback documents within a more disciplined mathematical foundation. The preliminary results show the advantages of the quantum model in re-ranking task.

V. conclusion & Future work

In this paper, we simulate the retrieval process of PRF by *N*-slit interference based on quantum theory and propose QIWA to re-allocate the document weights of feedback documents. The experimental results on three TREC datasets show that the proposed models are effective in the re-ranking task. This also demonstrates that the proposed model based on quantum interference have the potential of in-depth research.

As mentioned before, quantum models are still at the conceptual level and need more formalization to exploit new applications in IR tasks. In practical experiment, we make several strong assumptions. For instance, it is assumed that the cosine similarity between documents is still valid in the quantum model.

There are two future research directions. First, we will investigate the effectiveness of the document weight allocation framework in query expansion task; Second, more formalization are needed for the quantum application in PRF to improve existing QIWA and exploit other novel model.

Acknowledgment

This work is sponsored by the Ministry of Education Humanities and Social Sciences Research Fund (13YJC870023) and Independent Innovation Fund of Tianjin University (06060306004).

References

[1] Christophe D. Manning, Prabhakar Raghavan et al. An Introduction to Information Retrieval. Cambridge University Press UK, 2009. pp. 40-58.

[2] V. Lavrenko, W. B. Croft, “Relevance-based language models”. SIGIR, New Orleans, Louisiana, USA, 2001, pp. 120–127.

[3] J. M. Ponte, W. B. Croft, “A Language Modeling Approach to information Retrieval”. SIGIR, Melbourne, Australia, 1998, pp. 275–281.

[4] Fei Song, W. B. Croft. “A General Language Model for Information Retrieval”. CIKM, Kansas City, Missouri, USA, 1999, pp. 316–321.

[5] ChengXiang Zhai, John D. Lafferty. “A study of smoothing methods for language models applied to ad hoc information retrieval”. SIGIR, New Orleans, Louisiana, USA, 2001, pp. 334–342.

[6] Efthimis N., Efthimiadis, “Query Expansion”. Annual Review of Information Systems and Technology. Vol. 31(1), 1996, pp. 121–187.

[7] Zuccon, G., Azzopardi, L., van Rijsbergen, K. “The quantum probability ranking principle for information retrieval”. ICTIR, Cambridge, UK, 2009, pp. 232–240.

[8] Xiaozhao Zhao, Peng Zhang, Dawei Song, Yuexian Hou, “A Novel Re-Ranking Approach Inspired by Quantum Measurement”, ECIR, Dublin, Ireland, 2011, pp. 721-724.