Investigating Query-Drift Problem from a Novel Perspective of Photon Polarization

Peng Zhang[1], Dawei Song[1], Xiaozhao Zhao[2], and Yuexian Hou[2]

[1] School of Computing, The Robert Gordon University, United Kingdom
[2] School of Computer Sci & Tec, Tianjin University, China
{p.zhang1,d.song}@rgu.ac.uk, {0.25eye,krete1941}@gmail.com

Abstract. Query expansion, while generally effective in improving retrieval performance, may lead to the query-drift problem. Following the recent development of applying Quantum Mechanics (QM) to IR, we investigate the problem from a novel theoretical perspective inspired by photon polarization (a key QM experiment).

1 Introduction

Query expansion usually improves overall retrieval performance [1]. However, some expanded query may shift from the underlying intent of the original query, leading to the query-drift problem [6]. As a result, for some individual queries, the performance of the expanded query can be inferior to that of the original one. Motivated by the emerging research in applying Quantum Mechanics (QM) as a new IR formalism [3], we investigate the query-drift problem from a novel perspective of photon polarization [4], which has recently inspired a new model [5] to re-rank the top n (e.g. 50) documents obtained from the first-round retrieval. In this paper, our focus is on the query-drift problem with the expanded query.

The photon polarization experiment [4] involves the probability measurement of photons that can pass through a polarization filter. We can view documents as photons, and the retrieval process as measuring the probability of each document that can pass through the query's retrieval filter (as polarization filter). Then, the measured probability can be regarded as the estimated probability of relevance of each document. This QM experiment usually inserts an additional filter between the original filter and the photon receiver (e.g. a screen). Similarly, in query expansion, the expanded query is constructed for the second-round retrieval.

In QM, the probability that a photon can pass through an additional filter is the combined effect of probability measurement on both filters (i.e., the original and the additional ones). This inspires us, in IR, to fuse (i.e. combine) the retrieved results from the original query and the expanded one. Indeed, such fusion-based method has been shown to be an effective approach to tackling the query-drift problem [6]. Photon polarization provides a new perspective and a novel mathematical framework to look at the problem by considering the representation of the additional filter under the same basis as the original filter. This means that the expanded query can be implicitly observed with respect to the

G. Amati and F. Crestani (Eds.): ICTIR 2011, LNCS 6931, pp. 332–336, 2011.

original one. In this paper, we formulate the query expansion under the QM and derive a novel fusion approach to alleviating the query-drift problem.

2 Quantum-Inspired Approach

2.1 Photon Polarization

We first briefly introduce the idea of photon polarization [4]. A photon's state can be modeled by a unit vector $\varphi = a\,|{\rightarrow}\rangle + b\,|{\uparrow}\rangle$, which is a linear combination of two orthogonal basis vectors $|{\rightarrow}\rangle$ (horizontal polarization) and $|{\uparrow}\rangle$ (vertical polarization). The amplitudes a and b are complex numbers such that $|a|^2 + |b|^2 = 1$. Suppose the original filter is a horizontal polarization filter. Each photon will be measured by the basis $|{\rightarrow}\rangle$ and the probability is $|a|^2$, i.e., the squared norm of corresponding amplitude a in the horizontal direction. After the measurement, the photon's state will collapse to the original basis vector $|{\rightarrow}\rangle$. If we now insert an additional filter (e.g. with direction $\nearrow$ of 45-degree angle), then the new basis vectors become $|{\nearrow}\rangle$ and its orthogonal counterpart $|{\nwarrow}\rangle$.

2.2 QM-Inspired Fusion Approach

In the first-round retrieval, under the QM formulation, a document d's state can be formulated as:

$$|\varphi_d\rangle = a_d\,|q\rangle + b_d\,|\neg q\rangle \tag{1}$$

where q is the original query, $|q\rangle$ denotes the basis vector for relevance, $|\neg q\rangle$ denotes the basis for irrelevance which is orthogonal to $|q\rangle$, and $|a_d|^2 + |b_d|^2 = 1$. $|a_d|^2$ can denote the estimated relevance probability of the document d with respect to q. If we do not consider the state collapse after the first-round retrieval, d's state with respect to the expanded query q^e can be represented as

$$|\varphi_d^e\rangle = a_d^e\,|q^e\rangle + b_d^e\,|\neg q^e\rangle \tag{2}$$

where $|a_d^e|^2 + |b_d^e|^2 = 1$ and $|a_d^e|^2$ denotes the estimated relevance probability of document d with respect to q^e.

To prevent query-drift, the existing fusion models in [6] directly combine two probabilities $|a_d|^2$ and $|a_d^e|^2$. This direct combination ignores the theoretical fact that the two probabilities are under different basis, i.e. $|q\rangle$ and $|q^e\rangle$, respectively.

In this paper, we propose to fuse $|a_d|^2$ and $|a_d^e|^2$ on the same basis. First, to connect different basis $|q\rangle$ and $|q^e\rangle$, let $|q^e\rangle = a_{q^e}\,|q\rangle + b_{q^e}\,|\neg q\rangle$, where $|a_{q^e}|^2 + |b_{q^e}|^2 = 1$. Assuming that the amplitudes in Eq. 1 and Eq. 2 have been estimated, a_{q^e} can be estimated by solving the equation $|\varphi_d\rangle = |\varphi_d^e\rangle$ (see Eq. 1 and 2). If we consider the collapse of $|\varphi_d\rangle$ to $|q\rangle$ after the first-round retrieval, another equation $|q\rangle = a_d^f\,|q^e\rangle + b_d^f\,|\neg q^e\rangle$ needs to be solved too, using the estimate of a_{q^e}. The a_d^f here denotes the fused amplitude on the basis $|q^e\rangle$. The process of solving the above equations is omitted due to the space limit. The solution is that $a_d^f = a_d a_d^e + b_d b_d^e$. The amplitudes b_d and b_d^e correspond to the irrelevance basis and often lead to unstable performance in our experiments.

For the purpose of this paper, we drop the term $b_d b^e_d$ in a^f_d. Nevertheless, we will investigate its effect in more detail in the future. Then, we have

$$a^f_d = a_d a^e_d \tag{3}$$

Let $|a^f_d|^2 = |a_d|^2 \cdot |a^e_d|^2$ denote the fused relevance probability, which considers both $|a_d|^2$ (see Eq. 1) and $|a^e_d|^2$ (see Eq. 2), on the same basis $|q^e\rangle$. For each document d, $|a_d|^2$ and $|a^e_d|^2$ can be estimated as the normalized scores by a retrieval model for the original query q and the expanded query q^e, respectively.

It is also necessary [6] to define two functions $\delta_q(d)$ and $\delta_{q^e}(d)$, the value of which is 1 if d is in the result list of the corresponding query, and 0 otherwise. Then, based on Eq. 3, we propose two QM-inspired Fusion Models (namely QFM1 and QMF2), as formulated in Tab. 1. Two existing fusion models in [6], namely combMNZ and interpolation, are re-formulated in Tab. 1 for comparison. The combMNZ and interpolation are additive (i.e. adding up two scores $|a_d|^2$ and $|a^e_d|^2$), while the QM-based models are multiplicative. In QMF2, the smaller η can make scores of different documents retrieved for q^e more separated from each other, leading to more distinctive scores. In interpolation model, the smaller λ, the more the fused score is biased to the second-round score (i.e. $|a^e_d|^2$).

Table 1. Summary of Fusion Models

Model	**Fused Score for each** d
combMNZ	$(\delta_q(d) + \delta_{q^e}(d)) \cdot (\delta_q(d)\lvert a_d\rvert^2 + \delta_{q^e}(d)\lvert a^e_d\rvert^2)$
interpolation	$\lambda\delta_q(d)\lvert a_d\rvert^2 + (1-\lambda)\delta_{q^e}(d)\lvert a^e_d\rvert^2 \quad (0 \le \lambda \le 1)$
QFM1	$(\delta_q(d)\lvert a_d\rvert^2) \cdot (\delta_{q^e}(d)\lvert a^e_d\rvert^2)$
QFM2	$(\delta_q(d)\lvert a_d\rvert^2) \cdot (\delta_{q^e}(d)\lvert a^e_d\rvert^2)^{1/\eta} \quad (\eta > 0)$

3 Empirical Evaluation

Experimental Setup. Our experiments are constructed on four TREC collections (see Tab. 2). The title field of TREC topics is used as the original query q. Lemur 4.7 is used for indexing and retrieval [2]. The Dirichlet prior for smoothing language model is set as default 1000. Top 50 documents from the first-round retrieval are used for constructing the expanded query q^e (with 100 terms) by the Relevance Model (RM) [1]. For both q and q^e, the negative KL-divergence model [2] is adopted as the retrieval model and 1000 documents are retrieved. In both cases, the normalized score is computed by $\exp\{-D\}/Z$, where $\exp\{-D\}$ is to transform the negative KL-Divergence $(-D)$ into the interval $(0, 1)$, and Z as a normalization factor is the sum over all the transformed scores.

The Mean Average Precision (MAP) is used as the effectiveness measure, and the Wilcoxon significance test is used to compute the statistical significance. A robustness measure, i.e. <*Init* as used in [6], is adopted to test the percentage of queries for which the (M)AP drops after the query expansion.

Experimental Results. From Tab. 2, we can observe that, on ROBUST2004 and WT10G collections, for almost 50% queries (see <*Init.*), the performance

Table 2. Experimental Results. The smaller <*Init* generally means more robust performance. Statistical MAP improvements (at significance level 0.05) over Init.Rank. and RM are marked with α and β, respectively.

Collections	**WSJ8792**		**AP8889**		**ROBUST2004**		**WT10G**	
Topics	Topics 151-200		Topics 151-200		Topics 601-700		Topics 501-550	
Metrics	MAP(%)	<*Init*(%)	MAP(%)	<*Init*(%)	MAP(%)	<*Init*(%)	MAP(%)	<*Init*(%)
Init. Rank.	31.27	–	30.58	–	28.80	–	20.22	–
RM	37.75^{α}	22	**39.74**$^{\alpha}$	28	32.82^{α}	44	21.72	46
combMNZ	35.76^{α}	10	35.42^{α}	**12**	32.60^{α}	**19**	$23.31^{\alpha\beta}$	26
QFM1	36.87^{α}	**8**	36.12^{α}	14	32.81^{α}	21	$23.69^{\alpha\beta}$	**24**
interpolation	$38.84^{\alpha\beta}$	14	39.53^{α}	16	$34.47^{\alpha\beta}$	29	$24.38^{\alpha\beta}$	30
QFM2	**39.01**$^{\alpha\beta}$	16	39.20^{α}	18	**34.90**$^{\alpha\beta}$	30	**24.58**$^{\alpha\beta}$	30

of query expansion by RM is inferior to that of initial retrieval. All the fusion-based models can improve the robustness of query expansion. Two parameter-free models, i.e., combMNZ and QFM1, performs better than other two models in terms of robustness. QFM1 outperforms combMNZ in terms of MAP. On the other hand, QFM2 achieves a competitive performance in comparison with the interpolation model in terms of both MAP and robustness. For the interpolation model, we selected the best performing λ in the interval [0.1, 0.9] with step 0.1 for each collection, since we find that the model is sensitive to λ. For QFM2, we fix the η value as 0.1 and the performance is stable on all collections.

4 Conclusion

In this paper, we propose to investigate query expansion from a novel theoretical perspective inspired by the photon polarization in QM, and accordingly we have developed a novel fusion approach to alleviating the query-drift problem. The proposed models have been shown to largely improve the effectiveness and the robustness of a standard query expansion model. The performance is also comparable to two state-of-the-art fusion-based methods [6], shedding light on a promising new angle and mathematical formalism for further investigation.

Acknowledgments. This research is funded in part by the UK's EPSRC (EP/F014708/2), the China's NSFC (61070044) and the EU's Marie Curie Actions-IRSES (247590).

References

1. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: SIGIR 2001, pp. 120–127 (2001)
2. Ogilvie, P., Callan, J.: Experiments using the lemur toolkit. In: TREC 2001, pp. 103–108 (2001)
3. Piwowarski, B., Frommholz, I., Lalmas, M., van Rijsbergen, C.J.: What can quantum theory bring to information retrieval. In: CIKM, pp. 59–68 (2010)

4. Rieffel, E.G., Polak, W.: An introduction to quantum computing for non-physicists. ACM Comput. Surveys 32, 300–335 (2000)
5. Zhao, X., Zhang, P., Song, D., Hou, Y.: A novel re-ranking approach inspired by quantum measurement. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 721–724. Springer, Heidelberg (2011)
6. Zighelnic, L., Kurland, O.: Query-drift prevention for robust query expansion. In: SIGIR, pp. 825–826 (2008)