The Quantum Probability Ranking Principle for Information Retrieval

G. Zuccon[1⋆] and L.A. Azzopardi[1] and C.J. van Rijsbergen[1]
{guido, leif, keith}@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow
Scotland, UK

Abstract. While the Probability Ranking Principle for Information Retrieval provides the basis for formal models, it makes a very strong assumption regarding the dependence between documents. However, it has been observed that in real situations this assumption does not always hold. In this paper we propose a reformulation of the Probability Ranking Principle based on quantum theory. Quantum probability theory naturally includes interference effects between events. We posit that this interference captures the dependency between the judgement of document relevance. The outcome is a more sophisticated principle, the Quantum Probability Ranking Principle, that provides a more sensitive ranking which caters for interference/dependence between documents' relevance.

1 Introduction

The core task of Information Retrieval (IR) is to retrieve a set of documents satisfying a user's information need [6]. A key paradigm in IR [4] employs formal theories to estimate the probability of relevance of a document given a user's information need. In order to achieve an optimal retrieval performance, documents retrieved by the IR system are ranked in accordance to the Probability Ranking Principle (PRP) [5]. This posits that the system should rank documents in decreasing order of their probability of being relevant to the user's information need. Among others, one of the most controversial assumption made by the PRP is that the relevance of a document to an information need does not depend on other documents (*independent relevance* assumption). However, in real search situations the judgement of relevance made by the user about a document is influenced by the documents he previously examined through the search process [2]. Moreover, it has been shown that the utility of a document might become void if the user has already obtained the same information. This "interference" is due to several factors such as changes in information need, or information overlap among documents, or contrary information and is not accounted for by the PRP as relevance of a document judgements is assumed independent from other documents.

[⋆] Supported by EPSRC Grant number EP/F014384/. The authors are grateful to A. Huertas, D. Elliot, M. Lalmas, B. Piwowarski for their insightful comments.

In this paper, we model the PRP using quantum probability. The formulation of the PRP based on quantum probability naturally encodes quantum interference, which can be interpreted as modeling dependent relevance, thus overcoming the independent relevance assumption made by the original PRP.

The remainder of the paper follows. In Section 2 we present the double slit experiment, drawing a metaphor between IR and Physics. The classical PRP will be framed in the proposed analogy (Section 3), while arising of interferences in the experiment will be the stimulus towards a ranking principle which accounts for interference, the QPRP. In Section 4 we discuss a possible interpretation of the interference term in IR. The paper concludes stating the contribution of this study and lines of future work (Section 5).

2 The double slit experiment

In this section we introduce quantum probabilities and the quantum interference effect. Quantum interference is of major importance in our approach. To illustrate the difference between Kolmogorovian and quantum probabilities, we present a simple physical test, the double slit experiment [3], which describes arising the of interference between the probabilities of two events. The double slit experiment consists of shooting a physical particle (i.e. an electron, a photon, etc.) towards a screen with two slits, named A and B (Fig. 1(a)). Once the particle passes through one of the slits, it hits a detector panel, positioned behind the screen, in a particular location x with probability $p_{AB}(x)$.

By closing slit B, it is possible to measure the probability of the particle being detected in position x passing through A, namely $p_A(x)$. Similarly, by closing just slit A, we can measure $p_B(x)$. We call ϕ_A the (complex) probability amplitude associated to the events of passing through A when B is closed and being detected at x, and vice-versa for ϕ_B. The following equations state the relationship between probability and probability amplitudes: $p_A(x) = |\phi_A(x)|^2$; $p_B(x) = |\phi_B(x)|^2$. Intuitively[1], we would expect that the probability of the particle being detected at x when both slits are open is the sum of the probability of passing through A and being detected at x, $p_A(x)$, and the probability of passing through B and hit the detector panel in x, $p_B(x)$. Formally,

$$p_{AB}(x) = p_A(x) + p_B(x) = |\phi_A(x)|^2 + |\phi_B(x)|^2 \tag{1}$$

We refer to this case with the adjective *classical* meaning that no quantum phenomena would be observed. However, experimentally it has been noted that $p_{AB}(x) \neq p_A(x) + p_B(x)$, i.e. the probability of the particle being detected at x when both slits are open *is not* the sum of the probability with just slit A open plus that with just slit B open. Actually, the probability distribution that can be obtained measuring $p_{AB}(x)$ across the whole detection panel presents an interference pattern akin to waves that would pass through both slits and hit the detector panel. Thus, representing with $\phi_{AB}(x)$ the (complex) probability amplitude of a particle being measured at position x after passing through either

[1] And applying the Kolmogorovian law of total probability.

slit A or B, it is possible to state that $\phi_{AB}(x)$ is the sum of the probability *amplitude* associated to the event of opening just slit A plus the counterpart event of having open just slit B. In other words, $\phi_{AB}(x) = \phi_A(x) + \phi_B(x)$, and the probability of such event is $p_{AB}(x) = |\phi_{AB}(x)|^2$. The application of the previous relationships involving probabilities amplitudes results in

$$\begin{aligned} p_{AB}(x) &= |\phi_A(x)|^2 + |\phi_B(x)|^2 + (\phi_A(x)^*\phi_B(x) + \phi_A(x)\phi_B(x)^*) \\ &= p_A(x) + p_B(x) + I_{AB}(x) \end{aligned} \tag{2}$$

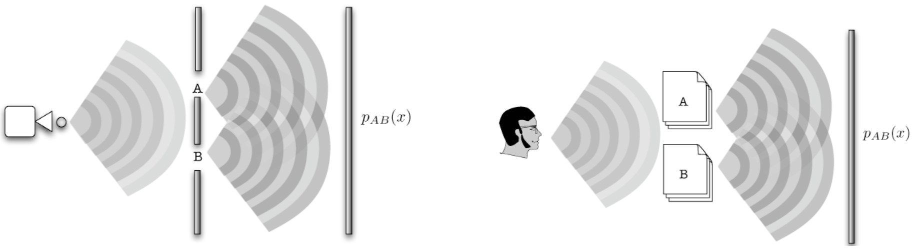
The term $I_{AB}(x)$ in Eq. 2 represents quantum *interference* between the events associated to $p_A(x)$ and $p_B(x)$ and is modulated by the phase difference between the correspondent amplitudes.

In summary, the conventional Kolmogorovian rule for addition of probabilities of alternatives, Eq. 1, is violated in the double slit experiment. When the event can occur in several alternative ways, the probability amplitude of the event, $\phi_{AB}(x)$, is the sum of the probability amplitude (the absolute square of a complex quantity) for each alternative considered separately. In the case of quantum probabilities, Eq. 1 is re-written with the addition of a perturbation term (shown in Eq. 2). The interpretation and the behavior of the interference term will be discussed later (Section 4); in the following we devise an analogy between the double slit experiment and the IR ranking process.

3 The analogy

In the following, we discuss (i) the classical PRP in terms of its decision theory derivation, adopting the analogy of the double slit experiment without interference effects, (ii) the case in which interference effects arise, and (iii) the derivation from the analogy of the new ranking principle.

We propose an analogy between the double slit experiment and the IR situation. In our analogy, the particle is associated with the user and his information need, while each slit represents a document. The event of passing from the left of the screen to the right (through a slit) is seen as the action of examining the ranking of documents, e.g. read the associated snippets or the documents themselves. Measuring at x means assessing the satisfaction of the user given the



(a) The double slit experiment

(b) The IR view of the double slit experiment.

Fig. 1. Schematic representation of the analogy between the double slit experiment and the IR ranking problem.

presented ranking of documents, or more concretely the decision of the user to stop his search (event x, the user is fully satisfied) or continue searching ($\bar{x}$, he is not completely satisfied by the documents presented). Thus, being detected with probability $p_{AB}(x)$ at position x on the panel means choosing to stop the search with probability $p_{AB}(x)$ after being presented with documents A and B. This scenario is represented in Fig 1(b). The user is presented with two documents, A and B, and he has to decide whether to stop the search (event x, associated probability $p_{AB}(x)$) or to continue (event $\bar{x}$, probability $p_{AB}(\bar{x}) = 1 - p_{AB}(x)$).

Probability $p_{AB}(x)$ is influenced by the characteristics of slits (documents) A and B. Consider the case several experiments are ran varying the screen among a set of them, all having the same slit A but each of them being characterized by a different slit B: e.g. B_h is narrow while B_j is wide, B_k is close to A while B_l is farer apart from A. The set of all different slits B_i is identified by $\mathfrak{B}$ and in our analogy it represents the set of candidate documents to be ranked immediately after document A.

Following the analogy, maximizing the expected utility of the ranking of documents is seen as maximizing the probability $p_{AB_i}(x)$, i.e. the probability of stopping the search having seeing A and B_i and, in the case of the physical experiment, maximizing the probability of the particle hitting the detector panel at position x (stop the search) passing though one of the slits. The problem then concretizes in determine which configuration of slits AB_i with $B_i \in \mathfrak{B}$ exhibits maximal $p_{AB_i}(x)$.

The classical case. If the double slit experiment is modeled assuming no interference, i.e. the "classical" case, the maximum $p_{AB_i}(x)$ is obtained by the configuration of slits with maximal $p_{B_i}(x)$. In fact, the probability of being detected at x being passed through either A or B_i is given by Eq. 1, and thus imposing maximal $p_{AB_i}(x)$ is equivalent to maximize $p_A(x) + p_{B_i}(x)$. However, since $p_A(x)$ is constant among all screen's configuration, we obtain

$$\underset{x}{\operatorname{argmax}}\big(p_{AB_i}(x)\big) = \underset{x}{\operatorname{argmax}}\big(p_A(x) + p_{B_i}(x)\big) = \underset{x}{\operatorname{argmax}}\big(p_{B_i}(x)\big) \tag{3}$$

In IR terms, given a fixed A (the document at first position of the ranking), the best document B_i to select among all the candidates $\mathfrak{B}$ is given by the document which maximizes $p_{AB_i}(x)$. In the classical case, $p_{AB_i}(x)$ is given by Eq. 1, and then maximizing it means choosing the document B_i with maximal $p_{B_i}(x)$, the document among the candidates $\mathfrak{B}$ with maximal probability of inducing the user to stop the search, i.e. probability of relevance.

In summary, maximizing the outcome of the measurement of a particles system passing through slits by choosing which pair of slits to use is analogous to choose which document to rank next, given a set of possible documents to rank. In absence of interference, the optimal rank suggested by the analogy with the double slit experiment is in accordance with the PRP: the slit B_i that should be used in order to maximize $p_{AB_i}(x)$ is the one for which $p_{B_i}(x)$ is maximal.

The quantum case. In the following we examine the situation where quantum phenomena appears in the double slit experiment and from this we abstract and derive a formulation of the PRP based on quantum probabilities. Maintaining the same analogy exploited previously, in presence of interference the probability $p_{AB}(x)$ is governed by Eq. 2. The probability of the particle being measured at position x in the detector panel is given by the sum of the probability of the particle being measured at x and passing either through A (term $p_A(x)$) or B (term $p_B(x)$), and a third term, the interference between the phases of the probability amplitudes associated to the mutually exclusive events of passing through Y $(Y = A, B)$ and being measured at x.

We suppose to have at our disposal a set of screens with a fixed slit A and different implementation of a second slit B_i. We aim to select the configuration of slits A and $B_i \in \mathfrak{B}$ which maximize probability $p_{AB_i}(x)$, representing the probability of finding a particle at position x on the detection panel after it passed either by slit A or B_i, analogous in the instituted metaphor to the probability of a user deciding to stop his search (because satisfied of the results obtained) after having examined either document A or B_i.

In presence of interference, $p_{AB_i}(x) = p_A(x) + p_{B_i}(x) + I_{AB_i}(x)$ leading to

$$\begin{aligned} \underset{x}{\operatorname{argmax}}\big(p_{AB_i}(x)\big) &= \underset{x}{\operatorname{argmax}}\big(p_A(x) + p_{B_i}(x) + I_{AB_i}(x)\big) \\ &= \underset{x}{\operatorname{argmax}}\big(p_{B_i}(x) + I_{AB_i}(x)\big) \end{aligned} \tag{4}$$

since $p_A(x)$ is constant among all the available screens. Allowing quantum interference, the maximum $p_{AB_i}(x)$ is reached when the sum $p_{B_i}(x) + I_{AB_i}(x)$ is maximal. The choice of the optimal screen among the possible screens with pairs of slits (A, B_i), $B_i \in \mathfrak{B}$ is not the same as in the classical case (the pair for which $p_{B_i}(x)$ is maximal) but depends upon $p_{B_i}(x)$ and the interference between A and B_i, $I_{AB_i}(x)$.

Deriving the Quantum PRP. The analogy suggests that the best choice for the document to rank after A is not the one for which $p_{B_i}(x)$ is maximal, i.e. the probability of relevance is maximal among the possible candidates $\mathfrak{B}$. Optimal rank would be produced when taking into account also the interference term. The probability of a document Y inducing the user to stop his search because his information need has been satisfied by the document is proportional to the probability of relevance to the information need of the document itself: $p_Y(x) \propto P(R|Y, q)$. We define $u(x)$ and $u(\bar{x})$ as the utility of retrieving a document which induces the user to stop his search and the utility of retrieving a document which does not induce the user to stop his search, respectively. We can safely assume $u(x) > u(\bar{x})$, setting for convenience $\big(u(x) - u(\bar{x})\big) = U$. Then, the expected utility in presence of interference can be written as:

$$\mathfrak{U} = p_A(x)U + p_Y(x)U + I_{AY}(x)U + u(\bar{x}) \tag{5}$$

The maximum value of expected utility is reached for the configuration which exhibits the maximum $p_Y(x) + I_{AY}(x)$, in fact $\operatorname{argmax}(\mathfrak{U}) = \operatorname{argmax}\big(p_Y(x) + I_{AY}(x)\big)$. When evaluating which is the optimal document to rank after A not only probability $p_Y(x)$ has to be taken into account, but also the probability of interference between the two documents affects the expected utility. Thus if dealing with quantum probabilities, document B should be ranked immediately after A and before any other document C if and only if

$$\begin{aligned} u(x)p_{AB}(x) + u(\bar{x})p_{AB}(\bar{x}) &\geq u(x)p_{AC}(x) + u(\bar{x})p_{AC}(\bar{x}) \\ \Leftrightarrow \boxed{p_B(x) + I_{AB}} &\geq \boxed{p_C(x) + I_{AC}} \end{aligned} \tag{6}$$

that is, B is the document belonging to $\mathfrak{B} = \mathfrak{Y} \setminus \{A\}$ for which $p_B(x) + I_{AB}$ is maximal. The statement of the Quantum PRP follows:

> *The quantum probability ranking principle (QPRP):* in order to maximize the effectiveness of an IR system, document B should be ranked after the set $\mathfrak{A}$ of documents already ranked and before any other document C in the list returned to the user who submitted the query if and only if $p_B(x) + I_{\mathfrak{A}B} \geq p_C(x) + I_{\mathfrak{A}C}$, where $I_{\mathfrak{A}Y}$ is the sum of all the interference terms associated to each pair of documents Y and $X \in \mathfrak{A}$.

Note that both the classical PRP and its quantum counterpart posit that the document at the first position of the ranking is the one with highest probability of relevance given the information need, since this is the document associated with the highest expected utility.

4 Discussion

In the quantum version of the PRP, the interference probability has a major role; but, **what is its interpretation?** We hypothesize that in IR interference occurs in the ranking between documents (or representations of them) at the relevance level. For example, [1] and [7] showed that the user is more likely to be satisfied by documents addressing his information need in different aspects than documents with the same content. Then, it might be sensible to model documents expressing diverse information as having higher degree of interference than documents that are similar. For the same reason, documents containing novel information might highly interfere with documents ranked in previous positions. Even contrary information might be captured by the interference term: documents containing content contrary to the one presented at the previous rank position might trigger a revision of user's beliefs about the topic. In summary, interference might model dependencies in documents' relevance judgements: the QPRP suggests that documents ranked until position $n-1$ interfere with the degree of relevance of the document ranked at position n. The classical PRP does not take into account dependent relevance of documents. Conversely, due to the presence of the interference term, the quantum ranking principle models dependent relevance and might be suited to address novelty/diversity in the documents ranking.

In what ways does the QPRP differ from the PRP? Both the classic PRP and its quantum counterpart posit that the document at the first position of the ranking is the one with highest probability of relevance given the information need, e.g. document A. The PRP ranks the documents that are left in decreasing order of relevance, while the QPRP postulates interference has to be taken into account. In the PRP the decision to rank a document in a particular position is not determine by the documents retrieved at previous ranks but only upon the relevance score assigned to other documents candidate to be ranked (i.e. independent relevance). Conversely, the interference term in the QPRP depends upon the documents ranked at previous positions. This means, the optimal order of documents under the PRP is different to that of the QPRP, and such difference is influenced by the interference term. **How does the interference term influence ranking of documents?** Consider Table 1. Assume $p_B(x)$ is greater than $p_C(x)$; then the PRP ranks B before C. However, from Eq. 6 the quantum PRP behaves in the same way (rank B before C) if and only if the difference between the probabilities associated to the single documents $(p_B(x) - p_C(x))$ is greater than the difference between their interference terms $(I_{AC}(x) - I_{AB}(x))$. Conversely, if this is not the case (i.e. $p_B(x) - p_C(x) < I_{AC}(x) - I_{AB}(x)$), the QPRP imposes to rank C before B. Then document C is promoted above B because its interference with the document ranked at the previous position (A) is so high that it fills the gap given by $p_B(x) + I_{AB}(x) - p_C(x)$. We interpret then document C as a document carrying diverse and novel information related to the query with respect to document A, while document B's content is less novel or possibly not novel at all with respect to document A. Moreover, when B and C are equally probable to be relevant ($p_B(x) = p_C(x)$), the PRP ranks first either one of them. However, in the same situation, the QPRP favors B above C if and only if the probability of B interfering with A is greater than the one of the pair (A, C). It is a matter of empirical investigation to determine how many times the rankings provided by the classical PRP and by its quantum counterpart differ.

What governs the interference term? Recall that the probability associated to the interference is given by $I_{AB}(x) = 2\,|\phi_A(x)|\,|\phi_B(x)| \cos\theta_{AB} = 2\sqrt{p_A(x)p_B(x)}\cos\theta_{AB}$, where θ is the difference of the phases of $\phi_A(x)$ and $\phi_B(x)$. When $\cos\theta_{AB} > 0$, $I_{AB}(x)$ is called constructive interference; conversely, destructive interference is obtained when $\cos\theta_{AB} < 0$. The behavior of the probability of the interference is governed by the phase θ.

How does interference behave by varying θ? The phase actively affects the documents ranking. For example, when $p_B(x) = p_C(x)$, document B would be ranked above document C when $\cos\theta_{AB} > \cos\theta_{AC}$. In general, when $p_B(x) \geq p_C(x)$ the interference term is able to subvert the ordering suggested by the classical PRP (i.e. "rank B above C") if

$$\frac{p_B(x) - p_C(x)}{2\sqrt{p_A(x)}} < \sqrt{p_C(x)}\cos\theta_{AC} - \sqrt{p_B(x)}\cos\theta_{AB} \tag{7}$$

How is θ computed in IR? While $p_A(x)$, $p_B(x)$, etc., are estimated from statistical feature of the document collection, the computation of the phase θ is still an open question and will be subject of further investigation. However, we suggest that θ could be approximated using the cosine similarity between documents. In particular, $\theta_{AB} \approx \arccos(\mathrm{sim}\,(A, B)) + \pi$. Alternative strategies might relate θ to the information gain or cross entropy between documents.

In summary, interference occurs between documents at relevance level. While the classical version of the PRP does not provide optimal ranking in presence of interference, the quantum PRP copes with this situation, promoting documents that positively interfere at relevance level.

Table 1. When does B have to be ranked above C? A comparison between classical PRP and its quantum counterpart (QPRP).

	$p_B(x) > p_C(x)$	$p_B(x) = p_C(x)$
PRP	B before C	either
QPRP	B before C iff $p_B(x) - p_C(x) > I_{AC}(x) - I_{AB}(x)$	B before C iff $I_{AB}(x) > I_{AC}(x)$

5 Conclusions

In this paper we exploit an analogy between the ranking problem in IR and the double slit experiment. The analogy introduces the presence of quantum interference between events. Taking into account the probability of interference, a new version of the Probability Ranking Principle, namely the Quantum PRP, has been proposed. We showed that the quantum version of the principle is a generalization of the classical PRP, and that it leads to optimal ranking solutions in presence of interference. In particular, it has been proposed that the interference term models the relationships between documents at the relevance level. Then, the document independency assumption needed for the classical PRP can be dropped in its quantum counterpart. In practice, the interference term is governed by the phase θ. The estimation of the phase in an effective way for IR is still an open issue; however, we have suggested possible avenues of research. To the best of our knowledge, our approach is the only that models dependent relevance in a principled way. It is interesting to investigate if other strategies which might violate the classical PRP, e.g. [1, 7], uphold for the QPRP.

References

1. H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06*, pages 429–436, NY, USA, 2006. ACM.
2. M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *JASIS*, 39(5):293–300, 1988.
3. R. P. Feynman. The concept of probability in quantum mechanics. In *Proc. 2nd Berkeley Symp. on Math. Statist. and Prob.*, pages 533–541. Univ. of Calif. Press, 1951.
4. M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, 1960.
5. S. E. Robertson. *The probability ranking principle in IR*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
6. C.J. van Rijsbergen. *Information Retrieval.* Butterworths, 1975.
7. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, NY, USA, 2003. ACM.