Document Ranking with Quantum Probabilities

Guido Zuccon

School of Computing Science
College of Science and Engineering
University of Glasgow

A thesis submitted for the degree of
Doctor of Philosophy

To Andrea

Abstract

In this thesis we investigate the use of quantum probability theory for ranking documents. Quantum probability theory is used to estimate the probability of relevance of a document given a user's query. We posit that quantum probability theory can lead to a better estimation of the probability of a document being relevant to a user's query than the common approach, i. e. the Probability Ranking Principle (PRP), which is based upon Kolmogorovian probability theory. Following our hypothesis, we formulate an analogy between the document retrieval scenario and a physical scenario, that of the double slit experiment. Through the analogy, we propose a novel ranking approach, the quantum probability ranking principle (qPRP). Key to our proposal is the presence of quantum interference. Mathematically, this is the statistical deviation between empirical observations and expected values predicted by the Kolmogorovian rule of additivity of probabilities of disjoint events in configurations such that of the double slit experiment. We propose an interpretation of quantum interference in the document ranking scenario, and examine how quantum interference can be effectively estimated for document retrieval.

To validate our proposal and to gain more insights about approaches for document ranking, we (1) analyse PRP, qPRP and other ranking approaches, exposing the assumptions underlying their ranking criteria and formulating the conditions for the optimality of the two ranking principles, (2) empirically compare three ranking principles (i. e. PRP, interactive PRP, and qPRP) and two state-of-the-art ranking strategies in two retrieval scenarios, those of ad-hoc retrieval and diversity retrieval, (3) analytically contrast the ranking criteria of the examined approaches, exposing similarities and differences, (4) study the ranking behaviours of approaches alternative to PRP in terms of the kinematics they impose on relevant documents, i. e. by considering the extent and direction of the movements of relevant documents across the ranking recorded when comparing PRP against its alternatives.

Our findings show that the effectiveness of the examined ranking approaches strongly depends upon the evaluation context. In the traditional evaluation context of ad-hoc retrieval, PRP is empirically shown to be better or comparable to alternative ranking approaches. However, when we turn to examine evaluation contexts that account for interdependent document relevance (i. e. when the relevance of a document is assessed also with respect to other retrieved documents, as it is the case in the diversity retrieval scenario) then the use of quantum probability theory and thus of qPRP is shown to improve retrieval and ranking effectiveness over the traditional PRP and alternative ranking strategies, such as Maximal Marginal Relevance, Portfolio theory, and Interactive PRP.

This work represents a significant step forward regarding the use of quantum theory in information retrieval. It demonstrates in fact that the application of quantum theory to problems within information retrieval can lead to improvements both in modelling power and retrieval effectiveness, allowing the constructions of models that capture the complexity of information retrieval situations.

Furthermore, the thesis opens up a number of lines for future research. These include (1) investigating estimations and approximations of quantum interference in qPRP, (2) exploiting complex numbers for the representation of documents and queries, and (3) applying the concepts underlying qPRP to tasks other than document ranking.

Acknowledgements

The support and encouragement of my family, friends, and colleagues had been essential during my Ph.D. In particular, I am most grateful to my parents, Anna and Carlo, for their unconditional support and to Magdalena for her love and understanding.

I wish to thanks my supervisors, Leif Azzopardi and Keith van Rijsbergen. Leif has given me invaluable insights, constant encouragement, and most of all friendship. Keith has given me constructive feedbacks and intellectual stimuli. I could not be prouder of my academic roots and hope that I can in turn pass on the research values and skills that they have given to me.

I am grateful to my Ph.D. examiners, Norbert Fuhr and Iadh Ounis: their feedbacks largely contributed in improving this dissertation.

A special thanks to all members of the Glasgow Information Retrieval Group, for creating a stimulating environment. I am truly indebted and thankful to Teerapong Leelanupab for sharing and discussing common research, ideas, feelings, and most of all for his friendship. I am grateful to Benjamin Piwowarski and Alvaro F. Huertas Rosero for the research conducted together and to Stewart Whiting and Jesus A. Rodriguez Perez for their help and support in the last period of my Ph.D. I want to express my gratitude also to the numerous visitors of the Glasgow IR Group, and in particular to Claudia Hauff, Ronald T. Fernandez, and Joaquin Perez.

I have been fortunate to have excellent mentors during my summer research visits: Dawei Song at the Knowledge Management Institute of the Open University, and Peter Bruza at the Science and Engineering

Faculty of the Queensland University of Technology. I particularly enjoyed the discussions with Peter and the members of his group, Kirsty Kitto, Bevan Koopman, Mike Symonds and Lance De Vine, to which I am grateful.

I am also grateful to the Information Retrieval Facility (IRF) and in particular to Oliver Falk for providing the computational power and assistance for running the experiments conducted in this thesis and in many of my research.

Thanks to my current colleagues at CSIRO, and in particular to Anthony Nguyen, for their patience and support when finalising this dissertation while I was already part of CSIRO.

Finally, I got into Information Retrieval in the first place by a fortuitous accident; for this I am grateful to Maristella Agosti who had given me the opportunity to work in this field of research and to visit Keith van Rijsbergen and the Glasgow IR Group during my Master degree at the University of Padua.

Contents

List of Figures

Chapter 1

Introduction

Information retrieval (IR) is the discipline concerned with searching for information in unstructured (or partially structured) document collections. The goal of IR systems is to retrieve all (and only) the documents that are likely to satisfy a user information need [van Rijsbergen, 1979], which is usually expressed by a query. Documents that satisfy a user's need are said to be relevant.

Early IR systems (such as those based on the Boolean model of IR) returned to users a set of documents as answer to their queries. It has been recognised, however, that retrieving an ordered list of documents rather than its unordered counterpart (that is, a set) enhances the effectiveness of IR tools [Frakes and Baeza-Yates, 1992], whether this is measured by user-centric measures as satisfaction, time the user is engaged with the system, abandonment rate [Radlinski et al., 2008], etc, or system-centric measures as precision, recall, cumulative gain, etc. [Manning et al., 2008].

Document ranking therefore plays a key role in the theoretical development of IR models and in shaping IR systems; and this is regardless of what we mean by documents, e.g. news articles, web pages, tweets, patents, people's profiles, etc.

The most popular ranking theory for IR is the *probability ranking principle* (PRP) [Robertson, 1977]. This principle states that documents should be ranked in decreasing probability of relevance to the user's information need. This statement is valid in both the following cases:

i) when the probability of relevance is introduced because of the continuous nature of relevance itself, i. e. documents are relevant to an information need with some extent or degree; or,

ii) when the probability of relevance is introduced because of the awareness that a retrieval system cannot discern with certainty the relevance of a document, i. e. only the user is able to judge the relevancy of a document (while the information retrieval system can only say that a document is likely to be relevant with a degree of confidence, i. e. a probability).

PRP has been shown to provide an[1] "optimal ranking" of documents from a theoretically perspective [Robertson, 1977]. Given two candidate documents, d_A and d_B, the ranking[2] $\langle d_A, d_B \rangle$ is optimal if d_A is more useful to the user than d_B. In this case, the optimality of PRP relates to measures such as precision and recall, and thus ultimately to the number of relevant documents retrieved amongst those examined by users.

PRP has shaped the field of IR, being central to many models (such as probabilistic models [Robertson and Sparck-Jones, 1976], language models [Hiemstra, 2001; Ponte and Croft, 1998], relevance models [Lavrenko and Croft, 2001]) and systems. However, this ranking principle is based upon a number of key assumptions. The most controversial assumption is that of independence between document relevance judgements. By following this assumption, the examination of a document and its assessment for relevance are carried out by users in complete isolation from other (retrieved) documents. This means that a relevance assessment for a document is not influenced by other documents. A number of empirical studies have suggested that this is often not the case and in general that PRP cannot be extended to all the retrieval scenarios that actually do happen in information retrieval applications [Carbonell and Goldstein, 1998; Chen and Karger, 2006; Eisenberg and Barry, 1988; Gordon and Lenk, 1991, 1992; Stirling, 1977; Wang and Zhu, 2009; Zhai and Lafferty, 2006]. Furthermore, it can be mathematically shown that PRP does not provide an optimal ranking when certain information retrieval measures are used to define the quality of a document ranking. Specifically, this is the case when the evaluation context accounts for the concept of interdependent document relevance, i. e. the relevance of a document

[1] Note that in general there can be more than one optimal ranking, i. e. there can be a family of document rankings that ideally provide the same performances to the user.

[2] Where $\langle d_A, d_B \rangle$ stands for document d_A followed by d_B.

might be influenced by that of other documents. The sub-optimality of PRP in specific search scenarios is not surprising, because PRP has not been designed for such search contexts, which have in fact been defined subsequently to the formalisation of PRP. So, if PRP does not hold or is ill-suited to such contexts, then is there an alternative ranking approach that is more suited?

In this thesis we investigate the theoretical underpinnings of document ranking in information retrieval, and document ranking is examined from a novel perspective, inspired by quantum theory. The use of quantum theory for expressing and analysing problems of information retrieval was first proposed by van Rijsbergen [2004]. The rationale under that work is that the mathematical formalism and machinery that have been developed to model physical phenomena appearing at atomic and sub-atomic scale present numerous similarities with the models developed in IR. In quantum theory, systems are represented by state vectors over a Hilbert space (i. e. a particular vector space, see Appendix C) that can be projected into subspaces representing outcomes (i. e. physical quantities), through the use of density operators representing observables. This procedure produces a probability measure that assigns to a particular system the probability of being observed in a particular configuration. Logic relationships, and in particular conditionals in logic, can be represented as geometrical objects in vector spaces. Quantum probability theory develops from these underpinnings, and it presents numerous points of departure from the traditional Kolmogorovian probability theory. In particular, differences between the two theories arise when measuring incompatible observables. The incompatibility between two observables translates into the impossibility to simultaneously perform measurements on them: one measurement has to follow the other and measuring one observable can affect a subsequent measurement on the other observable. This gives rise to phenomena of distortion or *interference*. From a mathematical perspective, when observables are incompatible, the probability of an outcome is not necessarily equal to the sum across the joint probabilities of all outcome combinations, as opposed to what Kolmogorovian probability states. This can be exemplified considering the settings of the double slit experiment in Physics. This experiment consists of shooting a physical particle towards a screen, which acts as measuring device. The experimental setting is completed by putting an additional screen between the emitter of particles and the measurement screen; the interposed screen is characterised by the presence of two slits, i. e. two holes in the screen. The execution of the experiment consists in closing one of the slits, say A, while keeping the other slit, say B, open and shooting a number of particles while recording their arrival distribution on the measurement screen. These series of measurements are repeated in the opposite situation, i. e. when slit A is open and B is close. By applying the rules of Kolmogorovian probability theory, it is possible to state that the probability of a particle being measured in a particular location on the measuring screen when both slits A and B are open, is equal to the sum of the disjoint events of measuring a particle at that same location when only A is open and when only B is open. However, the value of the probability of detection when both slits are open predicted according to Kolmogorovian probability theory is not consistent with the probability distribution experimentally measured when the experiment is repeated leaving both slits open. However, if quantum probability is used in the predictions, obtained with respect to the event of a particle hitting the measurement screen when both slits are open, do differ from those given by Kolmogorovian's axioms. In particular, such predictions accurately reflect the distribution that is experimentally measured. This is because when both slits are open, the measurements related to "the particle passing through slit A and hitting the screen" and to "the particle passing through slit B and hitting the screen" are incompatible: the two measurements cannot be carried on at the same time, and the measurement of one disturbs (or distorts) the measurement of the other. Mathematically, this corresponds to the presence of an additional term, called the interference term, when considering the sum of the joint probabilities.

The importance of the double slit experiment in the context of this thesis will be clarified in the next paragraphs. In the following we instead discuss why quantum theory and quantum probability might be of interest outside Physics, and in particular in information retrieval.

A number of works outside Physics have adapted the mathematical framework of quantum theory to model problems in Cognitive Science, Economics, Politics:

see for example Gabora and Aerts [2002], Bruza and Cole [2005], Busemeyer et al. [2006], Franco [2009], Choustova [2007], Choustova [2009], and Dubois [2009].

The fact that quantum theory encompasses geometry, probability and logics within a unique framework opens up the opportunity for creating a unique and solid mathematical model that links together core IR models based respectively upon geometry [Salton et al., 1975], probability [Spärck-Jones et al., 1998] and logics [van Rijsbergen, 1997]. The works of van Rijsbergen [2004], Melucci [2008], Widdows [2004], and Zuccon et al. [2008] have shown how the mathematical machinery of quantum theory can lead to alternative interpretations of current information retrieval models.

This thesis concerns with the use of quantum probability theory in IR and aims to develop a novel approach for IR. The premise of this thesis is that quantum probability theory can provide better estimates of the probability of relevance of a document to a user's query than Kolmogorovian probability theory. Previous literature that attempted to use quantum theory to model IR problems (e.g. van Rijsbergen [2004], Widdows [2004], Zuccon et al. [2008]) has not shown strong empirical evidence that quantum-inspired models can provide advantages in terms of retrieval effectiveness. In this thesis, we focus on the problem of ranking documents and we investigate the following research questions:

RQ1 How can quantum theory be applied in information retrieval, and in particular to document ranking?

RQ2 How does a quantum probability view of document ranking differ from the traditional Kolmogorovian approach?

RQ3 What are the implications of using quantum probability theory for ranking documents?

RQ4 Does quantum probability theory lead to improvements in retrieval performances with respect to traditional and existing ranking approaches?

To this aim, we construct a parallel between a physical phenomena and document ranking. We therefore use the mathematical framework of quantum theory to derive a model of document ranking as suggested by the analogy itself. Specifically, we focus on the probabilistic characteristics of quantum theory, exploring the use of quantum probability in information retrieval in contrast with its traditional analog, i. e. Kolmogorovian probability theory. This is pursued through an analogy between the document ranking scenario and the double slit experiment, which has been briefly described in the previous paragraphs. The analogy brings forward the concept of quantum interference, for which we propose an interpretation in the document ranking scenario, and in information retrieval in general. Such interpretation is closely related to the notion of interdependent document relevance, suggesting the ideal evaluation context where the outcome of the analogy can be tested. The presence of interference phenomena when ranking strikes on the satisfaction documents generate for users, and ultimately on the probabilities of relevance associated with documents. The consequence of the presence of interference is the derivation of a novel ranking principle, the quantum probability ranking principle (qPRP), that encompasses the PRP, but extends its optimality (considered within the view of ranking proposed through the analogy) to cases where the optimality of the PRP does not hold.

The analysis of document ranking is complemented by considering approaches alternative to the PRP and the qPRP. Within this work we show how the alternative approaches relate to or differ from the qPRP. Specifically we expose analytically and empirically their ranking behaviours. Ultimately, we also suggest how different approaches can be expressed as instantiations of the qPRP, under particular circumstances.

During the course of our investigation on document ranking, we also encompass issues related to pseudo-relevance feedback, similarity measures, evaluation, and fundamental theoretical aspects in information retrieval, such as the use of complex numbers.

1.1 Contributions

The main contributions of this work can be summarised in the following points:

- an alternative view of document ranking, inspired by quantum theory and realised through an analogy with the double slit experiment (Chapter 4);

- a theoretically sound ranking principle, quantum probability ranking principle, that extends PRP to situations where interdependent document relevance is admitted, and proposes to rank documents according to quantum probabilities (Chapter 4);
- the proposal of an interpretation of the phenomena of quantum interference that appears in document ranking when the analogy with the double slit experiment is considered and is central to qPRP (Chapter 5);
- a deep analytical and empirical understanding of PRP and alternative ranking approaches (such as qPRP) for information retrieval in a number of evaluation contexts (Chapter 6);
- an approach to bootstrap the parameter instantiation of a strategy for document ranking called portfolio theory (PT), through a mathematical relationship between the qPRP and PT (Chapter 6);
- a demonstration that the application of quantum theory to problems within information retrieval can lead to improvements in retrieval effectiveness.

1.1.1 Further contributions

Throughout this thesis, the following minor contribution can be identified:

- the proposal of a number of empirical instantiations of qPRP to rank documents in different scenarios, i. e. ad-hoc and diversity retrieval (Chapters 4, 5 and 6);
- the first empirical instantiation of the interactive PRP Fuhr [2008] in the case of first passage retrieval (Chapter 3);
- new insights on the proposal of using complex numbers in information retrieval, which has been first put forward in van Rijsbergen [2004] (Chapter 5).

1.2 Overview of the thesis

The thesis is structured as follows.

- Basic concepts in information retrieval are revisited in Chapter 2, where we describe those tasks, methodologies and evaluation practices we use later throughout this thesis to test the qPRP and other ranking approaches.

- In Chapter 3, we present an overview of ranking principles and strategies for information retrieval, with particular attention to ranking approaches that go beyond the Probability Ranking Principle.

- Chapter 4 contains the main theoretical contribution of the thesis. In this chapter, we propose a new ranking principle for information retrieval based on quantum probabilities, the quantum Probability Ranking Principle (qPRP).

- The notion of quantum interference plays a key role within qPRP. Quantum interference, its meaning in IR and methods for effectively estimating it when ranking documents are discussed in Chapter 5

- qPRP is analytically contrasted against PRP and alternative ranking strategies in Chapter 6, where also similarity and differences in ranking behaviours are exposed. Moreover, we empirically challenge the ranking approaches in two evaluation contexts: ad-hoc document retrieval and novelty and diversity document retrieval.

- Chapter 7 concludes the thesis discussing the implications of our proposal and experiments, and proving directions for future investigations.

The thesis is completed by a number of appendices.

- Appendix A summarises the notation and conventions used throughout this thesis.

- Appendix C presents the Dirac notation and the fundamental definitions related to Hilbert spaces: then it frames the double slit experiment in terms of Hilbert spaces.

- Finally, Appendix D reports the proofs of some relationships used in Chapter 4 for deriving and examining qPRP.

1.3 Publications

The following publications have arisen as part of this thesis:

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, "A Formalization of Logical Imaging for Information Retrieval using Quantum Theory", in DEXA Workshop on Textual Information Retrieval (TIR'08), pages 3–8, IEEE Computer Society, 2008 [Zuccon et al., 2008].

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, "Semantic Spaces: Measuring the Distance between Different Subspaces", in Proceedings of the Third Inter- national Quantum Interaction Symposium (QI'2009), volume 5494 of Lecture Notes in Computer Science, pages 225–236, Springer, 2009 [Zuccon et al., 2009a].

- Zuccon, G., "An Analogy between the Double Slit Experiment and Document Ranking", in the 3rd IRSG Symposium: Future Directions in Information Access 2009 (FDIA'09), 2009 [Zuccon, 2009]; (Chapter 4).

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, "Revisiting Logical Imaging for Information Retrieval", in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09), pages 766–767, ACM, 2009 [Zuccon et al., 2009b].

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, "The Quantum Probability Ranking Principle for Information Retrieval", in Advances in Information Retrieval Theory (ICTIR'09), volume 5766 of Lecture Notes in Computer Science, pages 232–240, Springer, 2009 [Zuccon et al., 2009c]; (Chapter 4).

- G. Zuccon, and L. Azzopardi, "Developing the Quantum Probability Ranking Principle to Rank Interdependent Documents", in Proceedings of the First Italian Information Retrieval Workshop (IIR'10), volume 560, pages

21–22, CEUR-WS.org, 2010 [Zuccon and Azzopardi, 2010b]; (Chapters 4 and 6).

- G. Zuccon, and L. Azzopardi, “Using the Quantum Probability Ranking Principle to Rank Interdependent Documents”, in Advances in Information Retrieval (ECIR’10), volume 5993 of Lecture Notes in Computer Science, pages 357–369, Springer, 2010 [Zuccon and Azzopardi, 2010a]; (Chapter 6).

- G. Zuccon, L. Azzopardi, C. Hauff, and C. J. van Rijsbergen, “Estimating Interference in the QPRP for Subtopic Retrieval”, in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR’10), pages 741–742, ACM, 2010 [Zuccon et al., 2010a]; (Chapter 5).

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, “Has Portfolio Theory got any Principles?”, in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR’10), pages 755–756, ACM, 2010 [Zuccon et al., 2010b]; (Chapter 6).

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, “The Interactive PRP for Diversifying Document Rankings”, in Proceeding of the 34th international ACM SIGIR conference on Research and development in information retrieval (SIGIR’11), pages 1227–1228, ACM, 2011 [Zuccon et al., 2011a]; (Chapters 3 and 6).

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, “Back to the Roots: Mean-Variance Analysis of Relevance Estimations”, in Advances in Information Retrieval (ECIR11), volume 6611 of Lecture Notes in Computer Science, pages 716–720, Springer, 2011 [Zuccon et al., 2011b]; (Chapter 3).

- G. Zuccon, B. Piwowarski, and L. Azzopardi, “On the use of Complex Numbers in Quantum Models for Information Retrieval”, in Proceedings of the Third international conference on Advances in information retrieval theory (ICTIR’11), pages 346–350, Springer, 2011 [Zuccon et al., 2011d]; (Chapter 5).

- G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen, "An Analysis of Ranking Principles and Retrieval Strategies", in Proceedings of the Third international conference on Advances in information retrieval theory (ICTIR'11), pages 151–163, Springer, 2011 [Zuccon et al., 2011c]; (Chapter 6).

- G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang, "Top-k Retrieval using Facility Location Analysis", in Advances in Information Retrieval (ECIR12), in Advances in Information Retrieval (ECIR'12), volume 7224 of Lecture Notes in Computer Science, pages 305–316, Springer, 2012 [Zuccon et al., 2012] (best paper award); (Chapter 3).

Chapter 2

Retrieval Models, Tasks and Evaluation

2.1 Introduction

In this chapter we provide a brief overview of IR concepts and techniques. Our presentation is not intended to report a complete account of techniques developed by IR researchers; instead it aims to introduce notions, measures, models and approaches that shall be used throughout the remainder of this thesis. Topics covered in this chapter comprise retrieval models, approaches to diversity, and evaluation frameworks. While, the notion of document ranking and approaches to ranking are presented in Chapter 3.

2.2 Models for Representing Information and Retrieving Documents

IR models specify how documents and queries are represented, and provide algorithms and criteria used to assess or estimate the relevance of documents to queries and hence retrieve documents. Models do not necessarily provide implementation details, but they act as artefacts that abstractly describe the information contained in documents and queries, and how these are matched. This idea of IR models, where a uniform representation of documents and queries is given together with the rules for matching these representations, is rooted in the work of Luhn [1957].

2.2.1 Boolean Model

Early approaches to document retrieval were based on the Boolean model for IR [Baeza-Yates and Ribeiro-Neto, 1999]. In this model, documents are indexed with respect to representative keywords or terms. Users requests (i. e. queries) are composed of keywords organised so as to form an expression conform to Boolean Algebra. In both documents and queries keywords are connected by Boolean Algebra operators. The basic operators of Boolean Algebra are the logical product, indicated by AND, the logical sum, called OR, and the logical difference, represented by the negation NOT[1]. The selection of keywords can be used to identify sets of documents: applying the logic operators to keywords modifies the sets identified by the keywords themselves. Combining keywords with the operator AND produces a document set that is smaller than or equal to the document sets formed by any of the single keywords. On the other hand, the combination of terms conjuncted by the operator OR, defines a document set which results larger than or equal to the document sets of any of the single keywords. To exclude the document set in which unwanted keywords are present, the operator NOT can be applied. However, nothing can be inferred about the dimension of the set corresponding to the negation of a keyword.

An IR system based on the Boolean model retrieves *all and only* the documents that satisfy the Boolean expression formed by the query with respect to the presence/absence of keywords in documents. For example, the query "information AND retrieval" would retrieve all and only documents that contain both keywords "information" and "retrieval". While, the expression "(information AND retrieval) NOT ranking", would retrieve only a subset of the documents retrieved previously, i. e. those documents that are indexed with both "information" and "retrieval" keywords, but not with the keyword "ranking".

In the Boolean model, retrieved documents are treated as a set: no formal order is imposed on the elements of the set. This is a key factor when deciding how to display the retrieved documents to a users: which document should be returned first? How should subsequent documents be ordered? This issue is further discussed in Chapter 3.

[1] $\wedge$, $\vee$, and $\neg$ are often used instead of AND, OR, and NOT, respectively.

2.2.2 Vector Space Model

In the vector space model for IR documents and queries are represented as points on an Euclidian vector space. To each document d and query q correspond a vector, e.g. $\mathbf{d} = \{\underline{d}_1, \ldots, \underline{d}_n\}$ for the document and $\mathbf{q} = \{\underline{q}_1, \ldots, \underline{q}_n\}$ for the query. The elements of the vectors, e.g. $\underline{d}_1$ and $\underline{q}_1$, are usually associated with terms appearing in the document collection [Salton and McGill, 1986; Salton et al., 1975] (e.g. $\underline{d}_1$ may be associated with term t_1), but other forms are possible, e.g. stemmed versions of the terms [Manning et al., 2008]. In the first case, the dimension of the Euclidean vector space that spans the entire document collection is equivalent to the number of terms contained in that collection.

The values of the elements of each vector depend on the weighting schema that is employed. If a binary weighting schema is used, each element of a vector representing an information object (i. e. a document or a query) is either one or zero, respectively corresponding to the presence or absence of the associated term. Alternative weighting schemas may encode the importance of terms into the value of the vectors elements. For example, the popular TF-IDF weighting schema assigns more importance (and thus higher values) to terms that appear frequently in a document but not in the collection. This is achieved by weighting vectors components according to:

$$\underline{d}_i = tf_{t_i,d} \log_2 \frac{N}{N_{t_i}} \tag{2.1}$$

where $tf_{t_i,d}$ represents the frequency of term t_i (associated with the component $\underline{d}_i$) in document d (i. e. the TF component), N is the total number of documents in the collection and N_{t_i} is the number of documents that contain term t_i at least once (and $\log_2 \frac{N}{N_{t_i}}$ is referred to as the IDF component). Weighting schemas that effectively capture the importance of terms within documents and queries are not a trivial problem [Salton, 1971] and have been extensively investigated in IR: a review of classical weighting schemas is provided by Salton and Buckley [1988].

Similarity between documents and queries is assessed by measuring the cosine of the angle θ that is spanned by the vectors, i. e. their inner product; thus the similarity between $\mathbf{d}$ and $\mathbf{q}$ is:

$$sim(\mathbf{d}, \mathbf{q}) = \cos \theta_{\mathbf{d},\mathbf{q}} = \frac{\sum_{i=1}^{n} \underline{d}_i \cdot \underline{q}_i}{\sqrt{\sum_{i=1}^{n} (\underline{d}_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (\underline{q}_i)^2}}$$

where n is the dimension of the Euclidean vector space into which the vectors are defined. The Boolean model retrieve documents that *exactly* match the Boolean statement expressed by the query. On the contrary, the use of cosine similarity in the vector space model allows the retrieval of documents that only *partially* match the query representation. Furthermore, while the output of the Boolean matching process is an unordered set of documents, cosine similarity in the vector space model provides a natural strategy for ordering documents: i. e. in decreasing order of their similarity with the query.

2.2.3 Probabilistic Model

The classical probabilistic retrieval model [Manning et al., 2008] treats document retrieval as a classification problem, where the goal is to distinguish the relevant documents from the non-relevant ones. In particular, assuming that $\mathcal{R} \in \{R, \bar{R}\}$ is a binary variable with value either R, corresponding to relevant, and $\bar{R}$, corresponding to non-relevant, we are interested to measure the probability of relevance given a document d, i. e. $P(\mathcal{R} = R|d)$, or in short $P(R|d)$. The following derivation is valid, where $\propto$ indicates rank equivalence:

$$P(R|d) \propto \frac{P(R|d)}{P(\bar{R}|d)} = \frac{P(d|R)P(R)}{P(d|\bar{R})P(\bar{R})} \propto \frac{P(d|R)}{P(d|\bar{R})} \tag{2.2}$$

Assume that the terms contained in document d, i. e. $t_1, \ldots, t_i, \ldots, t_n$ (where n is the number of terms contained in document d), are conditionally independent. Then equation 2.2 can be rewritten as:

$$\frac{P(d|R)}{P(d|\bar{R})} \approx \prod_{i=1}^{n} \frac{P(t_i|R)}{P(t_i|\bar{R})} \propto \sum_{i=1}^{n} \log \frac{P(t_i|R)}{P(t_i|\bar{R})} \tag{2.3}$$

Thus, probabilities are assigned to documents indicating their likelihood of being relevant to a user's request or information need: these are indeed determined by the probability of drawing the terms that compose each document from the classes of relevant and non-relevant documents. The estimation of such probability has been, and currently is, a major area of research in IR. Several approaches have been proposed to estimate this probability; among others we highlight: the 2-Poisson model [Bookstein and Swanson, 1974], the Binary Independence model [Robertson and Sparck-Jones, 1976], the BM25 model [Robertson and Walker, 1994].

2-Poisson Model

In this model, probabilities $P(d_i|R)$ and $P(t_i|\bar{R})$ are approximated by the number of occurrences (i. e. term frequency TF) of the index term t_i in the class of relevant (R) and non-relevant ($\bar{R}$) documents, respectively. In particular, the model assumes that these probabilities can be approximated by two distinguished Poisson distributions, characterised by different means μ_1 and μ_2 ($\mu_1 > \mu_2$, when μ_1 is associated with class R). If $\mathcal{X}$ is a random variable for the number of occurrence (i. e. term frequency tf) of a term, then:

$$P(\mathcal{X} = tf) = \lambda \frac{e^{-\mu_1}(\mu_1)^{tf}}{tf!} + (1-\lambda)\frac{e^{-\mu_2}(\mu_2)^{tf}}{tf!} \tag{2.4}$$

where λ is the proportion of documents that belong to R.

Given the 2-Poisson model, Harter [1975] suggested that the ratio $\frac{\mu_1-\mu_2}{\sqrt{\mu_1+\mu_2}}$ is proportional to $P(R|d)$ and can then be used to retrieve and rank documents with respect to a query. A more complete review of the 2-Poisson model has been given by Fuhr [1992].

A generalisation of the 2-Poisson model is Amati and van Rijsbergen [2002]'s Divergence from Randomness (DFR) model. The intuition behind this model is similar to that underneath the 2-Poisson: informative words can be represented by an elite set of documents where they occur more frequently than in the rest of the documents. On the other hand, not elite words are likely to follow a random distribution. The selection of the correct basic model of the random distribution is key to the DFR model. Proposed basic DFR models include the traditional concepts of inverse document frequency and inverse term frequency, as well as the Bose-Einstein distribution and the geometric Bose-Einstein distribution. A further characteristic of the DFR model is the presence of a two steps term frequency normalisation process. In the first step, documents are assumed of uniform length, while in the second step the actual document lengths are used to obtain the weighting formula.

Binary Independence Model

This approach considers the presence or absence (in a binary fashion, i. e. not recording the frequencies) of term t_i in document d; the Bernoulli distribution is used to approximate the conditionals $P(t_i|R)$ and $P(t_i|\bar{R})$, with parameters $\pi_{t_i,R}$ and $\pi_{t_i,\bar{R}}$, for classes R and $\bar{R}$ respectively. By assuming that terms not in the query q provide a constant contribution to the approximation of $P(R|d)$, the following equation can be derived:

$$\sum_{i=1}^{n} \log \frac{P(t_i|R)}{P(t_i|\bar{R})} \approx \sum_{i=1 \wedge t_i \in q}^{n} \log \frac{\pi_{t_i,R}(1-\pi_{t_i,\bar{R}})}{\pi_{t_i,\bar{R}}(1-\pi_{t_i,R})} \tag{2.5}$$

BM25 Model

Robertson and Sparck-Jones [1976] (re-examined later by Robertson et al. [1981]) proposed variations of the 2-Poisson model within the probabilistic model, so as to form a series of best match weighting schemas. In their approach, the weight w of a term t in document d is computed as[1]

$$w = \log \frac{(R_t + 0.5)/(R_q - R_t + 0.5)}{(N_t - R_t + 0.5)/(N - N_t - R_q + R_t + 0.5)} \tag{2.6}$$

where:

- N is the number of documents in the collection
- N_t is the number of documents that contain term t
- R_t is the number of relevant documents (i. e. belonging to class R) that contain term t
- R_q is the number of relevant documents for query q

Croft and Harper [1979] noted that when no relevance information is provided (i. e. when R_q is unknown or cannot be determined a priori) Equation 2.6 can be simplified so as to obtain:

$$w' = \log \frac{N - N_t + 0.5}{N_t + 0.5} \tag{2.7}$$

Note that in the previous equations, term frequencies within documents are not used to produce estimates of the probability of relevance. However, the previous model inspired the creation of variations that include term frequencies in

[1] Equation 2.6 is often referred to as the RSJ weighting schema.

the probability estimation. Of these, the most successful is the BM25 model (also called BM25 weighting schema in the IR literature and in this thesis), proposed by Robertson and Walker [1994]. In this model term frequencies, inverse document frequencies, and document lengths (of a document and the average document lengths of documents in the collection) are combined together to provide an effective and robust retrieval approach. In particular, in BM25 $P(R|d)$ is approximated as:

$$P(R|d) \propto \sum_{t \in q} w' \frac{(k_1+1)ntf_{t,d}}{k_1+1+ntf_{t,d}} \times \frac{(k_3+1)qtf_t}{k_3+qtf_t} \tag{2.8}$$

where w' is given by Equation 2.7 (and represented the inverse document frequency component), $ntf_{t,d}$ is the normalised term frequency of term t, qtf_t is the term frequency of t in query q, and k_1, k_3 are parameters. The normalised term frequency of t in a document is given by the following equation:

$$ntf_{t,d} = \frac{tf_{t,d}}{(1+b)+b\frac{l}{avg_l}} \tag{2.9}$$

where b is a parameter that ranges between 0 and 1, l and avg_l are respectively the length of document d and the average length of the documents in the collection. Note that when terms have a low inverse document frequency (i. e. $N_t > N/2$), BM25 may produce negative term weightings.

For more details, the reader is referred to Robertson and Zaragoza [2009], who provide a thorough review of the probabilistic model and in particular of BM25.

2.2.4 Language Model

Statistical language models for IR have been first independently proposed by Ponte and Croft [1998] and Hiemstra [1998](see also [Hiemstra, 2001]). The intuition behind this approach is that a probabilistic language model is built for each document d in the collection, and documents are retrieved with respect to the likelihood of a document model generating the user query q, i. e. $P(d|q)$. Applying Bayes rule, $P(d|q)$ can be rewritten as:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d) \tag{2.10}$$

where $P(q)$ can be ignored for rank equivalence motivations. $P(d)$ is the prior probability that document d is relevant and is used to encode query independent prior knowledge (such as the PageRank score of a document [Brin and Page, 1998], etc [Kraaij et al., 2002]). However, the prior is often considered uniform: in this case, $P(q|d)P(d)$ becomes rank equivalent to considering $P(q|d)$ alone. By assuming term independence[1], $P(q|d)$ can be computed by

$$P(q|d) = \prod_{t \in q} P(t|d)^{qtf_t} \tag{2.11}$$

where qtf_t is the frequency of term t in query q and is used to give more weight to frequent terms in long queries.

Different approaches have been proposed to calculate $P(t|d)$, aiming at addressing the sparseness problem of query terms not present in a document model. These approaches are known as smoothing techniques.

Jelinek-Mercer smoothing [Hiemstra, 2001] considers a linear combination of $P(t|d)$ and the background language model (i. e. $P(t)$, the probability of the term in the collection[2]):

$$P(t|d) = \prod_{t \in q} [\lambda P(t|d) + (1-\lambda)P(t)] = \prod_{t \in q} [\lambda ntf_{t,d} + (1-\lambda) ntf_{t,C}] \tag{2.12}$$

where $0 \leq \lambda \leq 1$ is an unknown parameter that has to be set, tuned or learned empirically; and as usual $ntf_{t,d}$ is the normalised term frequency of t in document d, while $ntf_{t,C}$ is the normalised term frequency of t in the collection C.

Dirichlet smoothing [Zhai and Lafferty, 2001] considers that documents contain too small evidence to reliably derive a language model, and therefore suggests to calculate $P(t|d)$ as:

$$P(t|d) = \prod_{t \in q} \frac{tf_{t,d} + \mu P(t)}{(\sum_{t_i \in d} tf_{t_i,d}) + \mu} \tag{2.13}$$

where $\mu \geq 0$ is a parameter of the model.

[1] This assumption is common in the binary independence model, [Robertson and Sparck-Jones, 1976], and BM25 [Robertson and Walker, 1994].

[2] That is, its frequency in the collection divided by the total amount of terms in the collection.

2.3 IR Tasks and Evaluation

2.3.1 Experiments and Evaluations in IR

The ultimate goal of evaluating IR systems is to assess whether users are satisfied with the documents returned in answer to their queries, i. e. to measure the *user satisfaction* with a particular system. IR theories and models develop techniques for maximising user satisfaction given their queries [van Rijsbergen, 1979]. To assess which approach performs best in addressing this problem, IR evaluation principally relies on experimental methodologies that are reliably repeatable, such as the Cranfield evaluation paradigm [Cleverdon, 1991], although other forms of evaluation are also used (e.g. user studies [Kelly, 2009]).

The Cranfield evaluation paradigm, which has been the corner stone of evaluation campaigns such as TREC [Voorhees and Harman, 2005], is based "on the abstraction of a test collection" [Voorhees, 2005]: a set of documents (also called corpus), a set of topics (which are usually a collection of queries, sometimes with a brief description of the associated information need) and a set of relevance assessments from topic experts[1]. In conjunction, these three elements form a test collection.

Note that in TREC, relevance assessments may be incomplete[2], i. e. not all the documents in the collection may have been judged with respect to their relevance to all the queries. This is the case when *pooling* techniques are used so as to allow evaluation of IR systems using incomplete relevance assessments. Pooling is usually performed by selecting the top k documents returned in response to each query by participating systems at TREC [Spärck-Jones and van Rijsbergen, 1975; van Rijsbergen, 1979]. These documents are then merged into a set, i. e. the pool: TREC assessors provide relevance assessments only for those documents contained in the pool. If many different IR systems have contributed to the formation of the pool, the relevance assessments are unlikely to be biased towards any particular system or model. Moreover the document corpus completed with queries and relevance judgements to form a test collection can be employed to reliably test other IR systems and techniques that did not contribute to the pool.

[1] Note that users are abstracted from the evaluation paradigm.

[2] On the contrary, the original Cranfield experiments considered complete relevance assessments [Cleverdon, 1991].

IR system are usually evaluated on the basis of how many relevant documents have been retrieved, and how many relevant documents have been missed. For example, *precision* measures the ratio of the retrieved relevant documents over all retrieved documents, while *recall* measures the ratio of retrieved relevant documents over all the possible relevant documents for a query. As ranking plays a fundamental role in IR (see Chapter 3), many evaluation measures are rank dependent (such as Average Precision and Discounted Cumulative Gain), i. e. the contribution of a relevant document to the overall user satisfaction is weighted according to the rank position at which the document is retrieved.

Evaluation frameworks and evaluation measures are tailored to how users are likely to use the IR systems [Belew, 2000; Goffman, 1964]. Different usage conditions determine different IR tasks, and evaluation frameworks/measures are generally task-oriented. In this thesis we shall evaluate ranking approaches on two different tasks: *ad-hoc document retrieval* and *diversity document retrieval*. Many other tasks are examined in IR; examples are enterprise search (e.g. [Hawking, 2004]), patent search (e.g. [Fujii et al., 2004]), medical and health-sciences systematic reviews (e.g. [Bouamrane et al., 2011]), HARD task at TREC (e.g. [Allan, 2003]), etc: and each task is characterised by its own evaluation framework and measures.

2.3.2 Ad-hoc Document Retrieval

In the evaluation framework of the ad-hoc document retrieval task it is assumed that users are interested to retrieve all documents that satisfy their information needs, i. e. users require to retrieve as many relevant documents as possible [Voorhees and Harman, 2005]. The user model of this task then prescribes that retrieved documents are examined in a linear fashion throughout the ranking, until a determined k rank position or until the ranking terminates. The goal of the task and its user model are reflected in the measures that are used to evaluate systems.

Evaluation of Ad-Hoc Document Retrieval

Average precision (AP) is the average of the precision values obtained after each relevant document is retrieved [Turpin and Scholer, 2006; Voorhees and Harman, 2005], i. e.

$$AP = \frac{1}{\sum_{i=1}^{n} r_{d_i}} \sum_{i=1}^{n} r_{d_i} \left(\frac{\sum_{j=1}^{i} r_{d_j}}{i} \right) \tag{2.14}$$

where r_{d_i} is 1 if document d_i is relevant, 0 otherwise; and i is the rank position of document d_i (in our notation, rank positions go from 1 to n). For a set of queries, values of AP are averaged to obtain the mean average precision (MAP) over the query-set: MAP assesses the overall performance of an IR system over a set of queries.

MAP is the common measure used consistently throughout the ad-hoc retrieval task. Alternative measures are also adopted. Precision at a specific ranking position k, i. e. $P@k$, is useful to assess the system effectiveness achieved after retrieving k documents.

Discounted Cumulative Gain (DCG) [Järvelin and Kekäläinen, 2002] takes into account both the extent a document is relevant to a query (i. e. graded relevance assessments) and the positions in which relevant documents are retrieved. The intuition of the position-biased discount is that relevant documents that have been retrieved lower in the ranking (e.g. at position k) have a lower contribution to the effectiveness of the system measured by DCG than relevant documents retrieved at earlier ranks (e.g. at position $k - m$, with $m > 0$ and $k - m > 0$). Specifically, DCG at rank k is defined as:

$$DCG(k) = r_{d_1} + \sum_{i=2}^{k} \frac{r_{d_i}}{\log_2 i} \tag{2.15}$$

where in this case r_{d_i} is the graded relevance assessment of the document ranked at position i. DCG values can be normalised, so as to facilitate comparison across queries. Normalisation is achieved by diving the actual DCG values by the values a perfect system would achieve. Specifically, a perfect IR system in the ad-hoc evaluation context would retrieve all and only the relevant documents for the user's query, and would rank them according to their (graded) relevance assessments. If the DCG value at rank k obtained by the perfect system is indicated with IDCG(k), i. e. the ideal DCG value, then the normalised DCG (nDCG) score at k is calculated as

$$nDCG(k) = \frac{DCG(k)}{IDCG(k)} \tag{2.16}$$

The reciprocal rank (RR) metric corresponds to the inverse of the rank position at which the first relevant document appears. For a set of queries Q, the mean reciprocal rank (MRR) is the average of the RR values for each query:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{RR}(q_i)} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank-first-doc-relevant}(q_i)} \tag{2.17}$$

where $|Q|$ corresponds to the number of queries in Q.

2.3.3 Diversity Document Retrieval

The task of diversity document retrieval stems from the need for document rankings that cover all possible different intents (also called subtopics or facets) relevant to the user's information need. In other words, in this task, user satisfaction is not only achieved by retrieving relevant documents: these have also to address the information need in a different way.

The notion of diversity is complex and context dependent. When an ambiguous word such as *apple* is posed to an IR system, diversification stands for addressing all possible interpretations of the word. In the case of our example, this would translate in retrieving at the top of the ranking documents related to the fruit interpretation of apple and the corporation interpretation of apple (i. e. Apple Inc.). On the contrary, when the issued query is unambiguous, such as *Apple Mac Os Lion*, effective diversification policies retrieve at top ranks documents that are topically diverse, while still addressing the user's query In our example, these may be web pages regarding the launch of Apple's operating system and reporting its features; as well as other web pages that indicate where and how to download or purchase the product; web pages that provide reviews or opinions about the system; and finally, other web pages that report problems with the software.

Radlinski et al. [2009] argue that diversity can be manifested in two, not mutually exclusive, forms: intrinsic and extrinsic. Intrinsic diversity refers to the situation where no single result can satisfy the user's information need. In this situation, diversity is an inherit property of the information need, which can only be addressed by multiple diverse documents. On the other hand, extrinsic diversity refers to the situation where result diversification is justified by the fact that the retrieval system is uncertain about the interpretation of the information need. It is then argued that the best approach to maximise the odds of retrieving relevant documents is to present the user with documents that cover the possible intents associated with the query.

Note that diversity may assume connotations other than that of topical diversity. The effectiveness of a system (and thus the user's satisfaction) may be enhanced by diversifying document rankings with respect to opinions, sources, media, etc. When considering the diversity retrieval task within this thesis, we focus on the topical aspect of diversity, although approaches and techniques developed here may be applied also to other forms of diversity.

The TREC Web Diversity Retrieval task has been created to foster developments of retrieval approaches and evaluation techniques within this new evaluation context. The task started in 2009 and is currently ongoing [Clarke et al., 2009a, 2010]; in this thesis we shall use collections, topics, and evaluation methodologies developed in TREC 2009 and 2010. Details of the test collection are provided later in the thesis (Chapter 6), while evaluation measures are outlined in the next section. We also adopt some of the evaluation measures proposed by Zhai et al. [2003] and an alternative test collection used in that work. This collection is based on the Financial Time corpus used both in TREC ad-hoc and interactive tracks. Specifically, the corpus of documents is completed with a set of topics (used in the TREC interactive tracks) for which a set of intents (or subtopics asdenoted by Zhai et al. [2003]) have been identified and against which documents have assessed for relevance. Further details about the collection are given in Chapter 6, while next the evaluation measures are detailed.

Evaluation of Diversity Document Retrieval

The evaluation of retrieval systems for search result diversification has attracted increasing interest from the research community. The early work of Zhai et al. [2003] is concerned with evaluating rankings with respect to the relevance of documents and the coverage of intents. Specifically, intent coverage at a given rank k as achieved by the retrieval of documents is captured by subtopic recall, s-r(k), which is formally defined as

$$\text{s-r}(k) = \frac{| \cup_{i=1}^{k} subtopics(d_i)|}{n_s} \tag{2.18}$$

where $subtopics(d_i)$ returns for a document d_i the set of relevant subtopics or intents that are addressed within the document itself, and n_s is the total number of possible subtopics relevant to the user's query. Intuitively, the fewer documents that have to be examined in order to address all intents or subtopics, the more effective the system.

Zhai et al. [2003] suggested also an extension of the traditional mean reciprocal rank measure, based on the same intuition of intent coverage exploited in subtopic recall. In particular, subtopic mean reciprocal rank (s-mrr is defined as the inverse of the rank at which full subtopic or intent coverage is achieved). The measure may be further adapted to assess partial intent coverage. For example, we indicate as s-mrr@25% the inverse of the smallest rank position at which at least a quarter of all possible relevant intents have been addressed by documents in the ranking.

For a precision oriented measure, subtopic precision at k, s-p(k), is calculated as the ratio between the minimum rank that an optimal system $\mathcal{S}_{opt}$ achieves a subtopic recall value of s-r over the corresponding minimum rank the system under evaluation, e.g. $\mathcal{S}$, achieves the same value s-r of subtopic recall. Formally,

$$\text{s-p(s-r)} = \frac{minRank(\mathcal{S}_{opt}, \text{s-r})}{minRank(\mathcal{S}, \text{s-r})} \tag{2.19}$$

where $minRank(\mathcal{S}, \text{s-r})$ is the function that computes the minimum rank position at which system $\mathcal{S}$ achieves a subtopic recall equal or greater than s-r.

In subtopic recall, each rank position is given the same importance, up to the cutoff k at which subtopic recall is measured. On the contrary, α-nDCG is based on a rank-based user model [Clarke et al., 2008]. This measure builds on DCG/nDCG by extending its original formulation to the case of document diversity retrieval. Similarly to DCG, α-nDCG is characterised by a rank-based discount function that affects the gains rewarded when retrieving a relevant document. Differently from DCG however, only the gains associated with documents that contains intents already addressed by other previously retrieved documents are discounted. That is, the gains for relevant documents addressing novel intents are not discounted. As for DCG, discounting is achieved using a log-harmonic function of the rank positions. Moreover, α-nDCG is characterised by a parameter, α, that has been suggested to control how much diversity is rewarded over relevance [Clarke et al., 2008], with 0.5 providing an adequate balance. This interpretation of α has been however questioned by Leelanupab et al. [2011]. They showed that setting α is crucial, and an incorrect setting may prevent the measure from behaving as expected: the setting $\alpha = 0.5$ may let the measure favour systems that promote redundant relevant intents. They presented and tested an approach to overcome this problem: briefly, this consists in setting the value of α above a certain threshold that is calculated on a query-by-query basis and that depends on the number of query-intents that have been identified in the relevance assessments[1].

Intent Aware Mean Average Precision (MAP-IA) is an extension of MAP to the diversity document retrieval task [Agrawal et al., 2009]. This measure computes AP for each intent separately, and then MAP-IA is obtained by a weighted average over all the intents. Formally, MAP-IA is obtained as

$$\text{MAP-IA} = \sum_{i=1}^{n} P(i|q)\text{MAP}_i \tag{2.20}$$

where $P(i|q)$ is the likelihood of intent i given query q, and MAP_i is the value of average precision obtained when considering documents relevant to intent i. The approach of computing an evaluation measure at intent level and then generating a weighted average over all intents can be applied to measures other than MAP. Following the same approach, Agrawal et al. [2009] extended nDCG to obtain its intention aware version, nDCG-IA. Similarly, Chapelle et al. [2009] framed the Expected Reciprocal Rank for ad-hoc retrieval within the diversity retrieval task, obtaining ERR-IA. It has been argued that the intention aware metrics tend to give no importance to the retrieval of documents relevant to low-weighted intents (i. e. with small $P(i|q)$): this may in general go against the specific goal of the diversity retrieval task [Sakai et al., 2010]. Moreover, the values obtained using intention aware measures are always less than one; this may be problematic when comparing and averaging the measures across topics.

[1] In this thesis however, we use the common settings of α-nDCG (i. e. $\alpha = 0.5$), as it is used in the standard TREC 2009 and 2010 Web Diversity Track evaluation.

Clarke et al. [2009b] proposed to build upon the evaluation approaches underlying α-nDCG, the intention aware measures (e.g. MAP-IA), and rank biased precision [Moffat and Zobel, 2008]. The outcome of their work is the Novelty- and Rank-Biased Precision (NRBP) measure. Similarly to α-nDCG, in NRBP, discounting is performed at two levels: a first level consists of discounting the gain with respect to the rank position, i. e. a (relevant) document retrieved at rank $k + 1$ would generate a lower gain than a (relevant) document retrieved at k, while an additional discount level is introduced to reduce the gain associated with a relevant but redundant[1] document.

Sakai and Song [2011] introduced two novel families of evaluation measures for this retrieval task, called D- and D#-measures. Both measures consider intent probabilities and per-intent graded relevance. D# extends the D-measure as it incorporates s-recall and thus it explicitly considers intent coverage. Sakai and Song argued that their measures overcome the problems that emerge in α-NDCG and in intent aware measures like NDCG-IA. However, it is unclear whether once complete intent-coverage has been achieved once, systems retrieving documents that cover a single intent and systems that provide a complete intent-coverage for a second time are distinguishable according to D#.

For a detailed discussion of the disadvantages and problematics of α-nDCG, NRBP, and the intention aware measures, we refer the reader to the work of Sakai and Song [2011]; Sakai et al. [2010]; for a comparison between these evaluation measures, we refer to the work of Clarke et al. [2011].

On the NP-Completeness of Diversity Document Retrieval Evaluation

Carterette [2011] noticed that evaluation measures for the diversity retrieval task such as subtopic precision, α-nDCG, NRBP, ERR-IA and the other intention aware measures, are based on a common intuition: compare the number of intents addressed by documents retrieved up to a rank k to the maximum number of intents that could have been addressed at the same rank, i. e. the optimal (or ideal) case. Finding such optimal combination is a NP-complete problem:

[1] i. e. a document that addresses an intent that have been already covered by one or more documents retrieved at previous ranks.

specifically, Carterette demonstrated that this problem can be reduced to the set cover problem [Vazirani, 2001]. A similar observation had also been suggested by Stirling [1977] when considering ranking approaches that would satisfy the totality of possible users that issue the same query (more details about this ranking problem shall be given in Section 3.5.2).

The optimal solution to the above problem can be found with a brute force algorithm. This algorithm requires $O(Sn!)$ operations, where S is the total number of intents associated with a query, and n is the number of documents that have to be examined. A common greedy approximation algorithm is usually adopted to make the computation of the evaluation measures feasible. This algorithm does not lead to the optimal solution (see Carterette [2011] for an example); the optimal solution can however be approximated within a constant factor of $1 - \frac{1}{e}$.

Note that an exact solution may be found with an alternative family of algorithms based on dynamic programming, called pseudo-polynomial time algorithms. Such algorithms exhibit exponential time complexity (specifically, $O(Sn2^n)$. Although the family of pseudo-polynomial algorithms may be suitable for finding the exact solution for the ideal document ranking problem, no experimental validation of this solution has been yet performed.

The NP-complete nature of this problem has serious implications on the reliability of the evaluation measures for this retrieval task: the effectiveness of a system in fact may be significantly overestimated. Carterette had showed that these estimation errors are not random errors, and thus do not average out across a large set of query topics. However, he also showed that only a subset of topics is affected by this problem, while for the majority of topics the problem is not present.

2.3.4 Approaches to Diversity Document Retrieval

In Chapters 3 and 4 we shall introduce a number of general ranking approaches that can be used across a wise set of retrieval tasks; we later shall test these approaches on both the ad-hoc document retrieval task and the diversity retrieval task. While these approaches are general, i. e. are applicable to a number of different retrieval tasks and do not use any specific information apart the terms contained in the collection's documents[1] – although these evidence may be incorporated in the approaches – other strategies may be devised that are specific to the document diversity retrieval task (and often these strategies are domain dependant).

Agrawal et al. [2009] devised a ranking strategy which maximises the probability that a document ranked at any rank position satisfies the user given the documents ranked at previous positions. To do so, they turn to examine the conditional probability that a query belongs to a category, given that all documents ranked up to a specific point fail to satisfy the user. To instantiate this strategy, it is necessary to resort to a taxonomy for the classification of intents, and to document and query classifiers.

Santos et al. [2010] proposed a ranking strategy that is based on uncovering the relationship between documents and possible sub-queries associated to the user's query. Such sub-queries are meant to identify the possible intents of a query. Several approaches can be used to generate sub-queries: query logs can be mined to uncover common reformulations of the original query, the query can be processed to obtain a series of meanings that would identify intents (for example using the Wikipedia disambiguation pages), etc. Santos et al. [2010] tested their approach by using the k-means clustering algorithm to partition the documents originally retrieved in response to the user's query: these clusters are then used as representative of sub-queries. A similar approach had been proposed by Leelanupab et al. [2010b]. Specifically, they use the fact that documents retrieved for the original query may be divided into partitions; then they iterate through each partition selecting documents according to policies akin to the strategy that shall be presented in Section 3.8 (namely Maximal Marginal Relevance). In their work, Leelanupab et al. investigated different methods to obtain partitions, all based on the original content provided in the documents; specifically, they employed Latent Dirichlet Allocation [Blei et al., 2003], Probabilistic Latent Semantic Indexing [Hofmann, 1999], k-means algorithm [Croft et al., 2009] (similarly to Santos et al. [2010]), and the partitions created according to the diversity

[1] i. e. these approaches do not rely on external source of evidence such as ontologies, Wikipedia, ect, nor they rely on query-expansion/refinement or are domain specific, e.g. make use of the Web graph for diversifying.

relevance assessments. According to the taxonomy proposed by Radlinski et al. [2009], these works explicitly address the problem of extrinsic diversity.

In a recent work Zuccon et al. [2012] proposed to frame the task of retrieving a set of diverse documents among the top-k retrieved documents into the operations research problem of facility location analysis. They showed that the ranking approaches we shall present in Chapters 3 and 4 can be framed as facility location problem, and in particular as "obnoxious facilities" that should be dispersed as far as possible from each other. Alternatively, they examined the algorithms that are generated if documents in the top k ranks were treated as "desirable facilities", to be placed as closed as possible to their costumers. In this proposal, as well as in the ranking approaches of Chapters 3 and 4, diversity is implicitly addressed; according to Radlinski et al. [2009]'s taxonomy, these approaches tackle intrinsic diversity.

Chapter 3

Ranking Documents

3.1 Introduction

In the first part of this chapter we examine why Information Retrieval systems benefit from a strategy for ranking documents retrieved in response to a user request. In particular, we are interested to derive a ranking approach that is *optimal*. We state that a ranking approach is optimal if it minimises the costs (or effort) the user has to sustain when examining the ranking. The intuition is that the system that minimises the costs a user has to sustain would also provide maximal user satisfaction. In this view, a cost is associated with retrieving an irrelevant document. Then, if for example a user is interested in retrieving a single relevant document, the cost incurred by the user depends upon how many irrelevant documents are ranked before the first relevant document. The optimal ranking approach is the one that minimises the cost incurred by the user: in our example it is the approach that retrieves a relevant document at the first rank position. Note that costs may be associated with not retrieving a relevant document. Similarly, benefits may be also considered, for example associated with retrieving a relevant document [Fuhr, 2008]. The assessment of the optimality of an approach indeed depends upon the specific IR task. For the ad-hoc retrieval task (Section 2.3.2), a ranking is optimal if it retrieves all the relevant documents before any irrelevant document. If the diversity retrieval task (Section 2.3.3) is considered instead, optimality is achieved if relevant documents addressing different query-intent are retrieved at top ranks.

In Information Retrieval ranking is traditionally implemented by the probability ranking principle (PRP). PRP not only sets a ranking rule to be followed, but it also guarantees the optimality of the document ranking under a number of assumptions. However, empirical evidences drawn, for example, from users behaviours and search engines logs [Das Sarma et al., 2008; Radlinski and Dumais, 2006], suggested critics to PRP's assumptions, calling for ranking approaches that "go beyond" PRP. In particular, while such critics may not apply to tasks such that of ad-hoc retrieval, they do apply to tasks such as diversity document retrieval. These issues are discussed in the second part of the chapter, where we examine PRP, discussing its optimality and assumptions. Furthermore, we present a number of alternative ranking approaches that overcome PRP's assumptions and limitations.

3.2 Motivations for ranking documents

Early approaches to document retrieval were based on the Boolean model (see Section 2.2.1). In this model, documents are indexed with respect to representative keywords. Requests (queries) are also composed of keywords organised so as to form an expression conform to Boolean logic. An IR system based on the Boolean model would then retrieve all and only the documents that satisfy the Boolean expression formed by the query with respect to the presence/absence of keywords.

Boolean retrieval systems therefore do not provide a means to measure the extent to which a document satisfy a query. In other words, if Boolean retrieval is seen as a decision process, then the output of the process (i. e. the decision) would simply be a yes or no answer (i. e. a dichotomous decision). This decision is taken with respect to a criterion, identified by the Boolean query, which defines the regions of acceptance and rejection in the decision space [Dowdy et al., 2004]. Documents are accepted, and therefore retrieved, if they fall in the acceptance region (i. e. they completely fulfil the query); while, they are rejected and thus not retrieved, if they fall in the rejection region.

No confidence level of associated with the decision, nor a score that indicates the extent the document satisfies the query. Therefore, the decision process implemented by the Boolean model does not provide a natural ranking of documents [Harman, 1992]. In fact, the retrieved documents are treated as a set with no formal order. This is a key factor when deciding how to display the retrieved documents to a users: which document should be returned first? And, how should subsequent documents be ordered?

In Boolean systems, the display order of documents is therefore arbitrary, i. e. it is based upon some feature external to the matching rule. For example, the Westlaw[1] search system (one of the largest commercial legal search service) returns the retrieved documents in reverse chronological order [Manning et al., 2008]. However, without a criteria that establishes what is the most suitable, or optimal, ordering of documents in response to queries, any choice of document ordering might be deemed as arbitrary: each retrieved document is assumed to be as important as any other retrieved document [Salton et al., 1983]. Gebhardt [1975] argues that a retrieval system admitting only Boolean operators and not imposing an order on retrieved documents based on weights with respect to the presence of query keywords in documents is "far from being optimal".

Several extensions of the Boolean retrieval model have been proposed (e.g., Salton et al. [1983]), as well as alternative models (e.g. the vector space model, the logical inference model, the probabilistic model, etc), that assign importance weights, similarity scores, or probabilities, to retrieved documents. These allow for an ordering of documents, and therefore for a ranking, based on the notions of similarity, importance, usefulness or relevance. In the following we focus on *how document rankings are formed within a probabilistic framework*, rather than examining ranking within extensions of the Boolean model or others. This is because the treatment of ranking within a probabilistic framework can be developed in a more rigorous perspective, and because probabilistic models for Information Retrieval have proven to be highly powerful and effective. Furthermore, note that optimal retrieval in IR has been precisely defined only for probabilistic models, where the optimality of a ranking strategy can be formally demonstrated or confuted [Fuhr, 1992].

[1] http://www.westlaw.com/, last visited June 18, 2012.

3.3 The Ranking Process

As said in Section 3.1, the optimal ranking is the one that guarantees the lowest costs to users, or in other words is a document ranking that yields the highest user satisfaction because it delivers useful information to the user.

The document ranking process consists of ordering documents so as to achieve the highest retrieval effectiveness and ultimately minimise the user's costs delivering the highest user satisfaction. Therefore a ranking algorithm strives to order documents so as to produce an optimal ranking. Ranking is formed according to a ranking criterion, or rule, which expresses an objective function that has to be maximised. We consider *sequential ranking algorithms*, where documents are ranked in sequence, and once a document has been ranked according to the ranking criterion, the choice is not revised. *Non-sequential ranking algorithms* are also possible [Zuccon et al., 2012], but these may lead to higher computational costs (up to exponential-time solutions under particular settings and scenarios), and are thus usually not considered in real search settings, where execution time is an issue.

A sequential document ranking algorithm can be formally described by Algorithm 1, which requires as inputs a set of documents D, a query q, a ranking criterion $\mathfrak{C}(D, q)$ (or $\mathfrak{C}$ for brevity) and returns a ranked list $\mathcal{RL}$ containing the documents in D ordered according to the ranking criterion $\mathfrak{C}$.

Algorithm 1: A generic (sequential) document ranking algorithm.

```
Input: a set of documents $D$
       a query $q$
       a ranking criterion $\mathfrak{C}(D, q)$
Output: a ranked list $\mathcal{RL}$
while $D \setminus \mathcal{RL} \neq \varnothing$ do
    select $d$ from $D \setminus \mathcal{RL}$ such that $d = \arg\max_{d \in D \setminus \mathcal{RL}} \mathfrak{C}(D, q)$;
    insert $d$ at the tail of $\mathcal{RL}$;
end
return $\mathcal{RL}$;
```

3.4 Ranking with Probability of Relevance

Because the factors that influence relevance are complex and evolve over time and changing of context[1], relevance cannot be assessed with certainty: this is the intuition underling probabilistic models of Information Retrieval, where the relevance of a document to a query is assessed probabilistically.

In probabilistic approaches, a document is said to be relevant to a user's query with a probability[2] $P(R|d,q)$. There are many possible interpretation of what a probability of relevance represents. On one hand, the description of the information need users submit to an IR system, i. e. queries, is inevitably incomplete, and often even ambiguous [Spärck-Jones et al., 2007]. The IR system then cannot discern with certainty whether a document is relevant or not to the user: it can only estimate how likely the document is to be relevant to the user issuing the query. On the other hand, uncertainty is associated with the relevance of a document to an information need, where relevance is treated as a continuous, or multi-valued variable: this generates the notion of degree of relevance [Robertson and Belkin, 1978]. Also, for the same pairs of document-query, different users might have different assessments of relevance, because of their contexts, expertise, etc [Maron and Kuhns, 1960; Stirling, 1977]. There are two main interpretations of what *probability of relevance* is:

(a) **A relationship between a single document and a class of users.** For example, the Wikipedia page dedicated to the topic "Computer virus" (`http://en.wikipedia.org/wiki/Computer_virus`), is likely to be relevant to computing science students that have issued the query "origins of virus"; on the contrary is unlikely to be relevant to medical students that issued that same query. This interpretation of probability of relevance underpins the works of Maron and Kuhns [1960] and Stirling [1977].

[1] We refer the interested readers to the works of Cosijn and Ingwersen [2000]; Mizzaro [1997]; Schamber et al. [1990]; Xu and Chen [2006] who analysed the concept of relevance, its dependency upon user's context, and its evolving nature.

[2] The notation used throughout this thesis with respect to the probability of relevance is introduced in Appendix A.

(b) **A relationship between a single user and a class of documents.** With respect to the example given for interpretation (a), a computing science student issuing the query "origins of virus" may consider relevant documents that discuss the topic of computer viruses (i. e. that belong to the same class) such as `http://en.wikipedia.org/wiki/Computer_virus` or `http://www.antivirusworld.com/articles/history.php`, while he may consider not relevant documents that investigate the origins of the natural (i. e. not computer related) infectious agent. This is the interpretation underlining the works of Robertson and Sparck-Jones [1976], Robertson [1977], Robertson and Belkin [1978] and of all recent developments in information retrieval (e.g. Hiemstra [2001]; Ponte and Croft [1998]).

These two interpretations are fundamentally different, but in the rest of the thesis we will adopt the second one, which is the common interpretation currently adopted in IR. However, it is interesting to note that the a ranking approach may be applied to both interpretations, with little adaptations. In fact, a formal link between the two interpretations have been provided by Robertson et al. [1982]. In particular, that work suggested a unifying theory of probability of relevance, including both interpretations (a) and (b), which are further complemented by a lower level and a higher level interpretation. While the lower level interpretation (referred to as Model 0 by Robertson et al. [1982]) groups both documents and users (or user classes), the higher level interpretation (Model 3) considers relevance judgements as the byproduct of the interaction between the events characterised by the pairs formed by individual users with group of documents, and group of users with individual documents. In this respect, in Section 3.5.2 we shall show how the Probability Ranking Principle, which is indeed based on interpretation (b) of the concept of probability of relevance, can be derived and applied also in the case of interpretation (a).

Once the probability of being relevant to a user's query is associated to each document, we can ask ourselves: how should the documents be returned to users? It has been argued that if documents are returned in a ranking, and if the ranking is optimised so as to minimise the number of documents that should be examined by users to satisfy their information needs (i. e. users effort), then the probability user satisfaction is maximised [Gordon and Lenk, 1991, 1992]. This argument is akin to the principle of *sequential optimisation* is search theory [Benkoski et al., 1991; Dobbie, 1963]. The effectiveness of the document ranking is then influenced by the sequential procedure followed for generating the ranking. The goal of a ranking criterion is therefore to produce an optimal ranking so as to minimise user's effort, while maximising the probability of retrieving documents relevant to the information need.

The problem of document ranking can then be stated as follows:

Definition: Given a user u, a set of documents D, and a user's query q, provide an ordering of D such that the user satisfaction is maximum at any given rank position.

In the following, we assume user satisfaction to be a fundamental dichotomous variable that is defined by u. We also assume that user satisfaction or usefulness of a document corresponds to the relevance of the document to the user's query: in other words, we consider a relevant document as being useful, or giving satisfaction to the user. As satisfaction, we also consider relevance as being a dichotomous variable, if not differently stated: a document can be either relevant or not relevant.

Next, we examine a number of ranking criteria that have been proposed for ranking documents in answers to users queries. We first examine the Probability Ranking Principle; then we turn our attention to alternative approaches to rank documents.

3.5 The Probability Ranking Principle

Consider the following settings. A set RE of documents is retrieved in response to query q issued by user u. A probability distribution $P(R|d, q)$ can be defined to identify the likelihood of a document $d \in RE$ to be relevant to q issued by u. A natural and straightforward ranking approach would be to rank documents in decreasing order of their relevance probability. That is, the document that is ranked at rank i is the document with highest probability of relevance that has not been yet ranked.

It is difficult to trace back in the IR literature the origins of this document ranking criterion. The approach is well known as the Probability Ranking Principle (PRP) and is commonly attributed to Robertson [1977]: although the author is not responsible for the proposal of PRP, he has the merit of rigorously state it and analyse its optimality. Robertson himself attributed the ranking criterion to other authors, such as Maron and Kuhns [1960] and Cooper. Other works have devised ranking solutions based on the ranking criterion of PRP during the same period; see for example the works of Dyke [1959], Verhoeff et al. [1961], Mulvihill and Brenner [1968], Cooper [1972]. Nevertheless, the Probability Ranking Principle plays a central role in the development of Information Retrieval models and theories.

The definition of PRP relies upon a number of assumptions. Specifically:

1. Relevance is a *dichotomous* judgment, i. e. a document is either relevant or not and the user cannot express a degree of relevancy for a document;

2. The probabilities of document relevance are estimated *as accurately as possible* on the basis of the data that is available to the system;

3. The relevance of a document to a request is independent of the relevance value of other documents (we refer to this as the *independence assumption* between relevance assessments);

4. PRP is applied to *one query only* at a time, and not to the interaction between user and system comprising issuing a series of multiple reformulated queries;

5. The effectiveness of a system in terms of user performances is measured by *recall and expected precision.*

Assumption 1 means that, from a user's perspective, relevance is dichotomous: a user judges a document either relevant or not. A degree of relevance is not admitted under this assumption, and therefore multivalued relevance is excluded. The assumption does not undermine the presence of probability of relevance from a system point of view though, as the system does not know with certainty[1] if a document is relevant to a user: the best a system can do is to estimate the relevancy of a document. This is analogous to the case of an urn containing x white balls and y black balls: the colour of a ball that is drawn from the urn is either white or black; while the drawing of a grey ball is impossible. However, it is possible to determine how much is the likelihood to draw a white ball (i. e. $\frac{x}{x+y}$), or conversely that of drawing a black ball (i. e. $\frac{y}{x+y}$).

Note that not upholding Assumption 1 does not automatically imply the sub-optimality of the ranking criterion: Bookstein [1993] had examined the optimality of PRP in case of multi-valued relevance. Conversely, Assumptions 2 and 3 are crucial for the optimality of PRP.

Next, we provide a definition of PRP based on Assumptions 1–5; similar definitions have been given that rely on different assumptions[2]: we examine the relations between Assumptions 1–5 and alternative formulations in Section 3.6, where we shall discuss how PRP's optimality is affected by one or more assumptions not holding.

The original statement of the Probability Ranking Principle given by Robertson [1977] states:

> "If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

The principle can be rewritten in the light of the assumptions discussed in the previous paragraphs.

Definition (PRP): If Assumptions 1–5 hold and an Information Retrieval system's response to a query q submitted by user u is a ranking of documents in

[1] Note however that Assumption 1 restricts to point estimations of such probabilities, thus not accommodating for measuring the uncertainty associated with the estimations.

[2] See for example the works by Gordon and Lenk [1991, 1992].

decreasing order of probability of relevance to q, then the overall effectiveness of the system to user u will be the *optimal* achievable on the basis of the data available.

Formally, the definition translates in the following ranking criterion: at rank i, PRP ranks document d_i such that

$$d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} P(R|d,q) \tag{3.1}$$

where $\mathcal{RL}$ is the list of documents that have been ranked, d is a document belonging to the set of retrieved documents RE for query q but not ranked yet (i. e. $RE \setminus \mathcal{RL}$), and $P(R|d,q)$ is the probability of document d to be relevant to the information need expressed by q.

In the next section, we demonstrate the optimality of PRP; an alternative demonstration of PRP's optimality in terms of utility is given by Gordon and Lenk [1991], while Robertson [1977] shows that PRP is optimal with respect to recall and expected precision, as well as expected search length (see [Cooper, 1968] for the last result).

3.5.1 Proof of Optimality

The optimality of PRP when Assumptions 1–5 hold can be demonstrated in terms of decision or utility theory [Fuhr, 1992; Robertson, 1977]. To form a document ranking, the ranking process starts at the first position in the ranking, evaluates all candidate documents and allocates to the current position the document that maximises the criterion set by the ranking approach. Once the first position is secured, the ranking process moves to the next available position and re-iterates the previous steps, but considering a set of candidate documents that does not comprise the document that has been already ranked. Then, at each ranking step, for every document in the set of not ranked documents, the system has to decide whether or not to retrieve it. Costs can be associated with the events of retrieving a non relevant document and not retrieving a relevant document:

$$\begin{aligned} \mathcal{C}(\text{retrieve}|\text{relevant}) = &\quad c & (3.2) \\ \mathcal{C}(\text{retrieve}|\text{non-relevant}) = &\quad \bar{c} & (3.3) \end{aligned}$$

where $c < \bar{c}$.

Let $P(R|d_j, q)$ indicate the probability of relevance given document d_j and query q, and $P(\bar{R}|d_j, q)$ the probability of *non* relevance given the same pair of query and document. The optimality of the ranking rule under Assumptions 1–5 depends upon whether the costs involved with ranking documents are minimised for each ranking position. To prove that PRP provides the optimal decision for ranking documents, we consider the case when d_j has been ranked before d_i. In particular, we examine the costs associated to this ranking choice. The total cost associated to rank a document d is given by the cost for ranking a relevant document, multiplied by the chances that d is relevant, plus the cost for ranking a non-relevant document, multiplied by the chances that d is not relevant. In case of document d_j, the total cost $\mathcal{T}_{\mathcal{C}}(d_j)$ associated to the decision of ranking the document is:

$$\mathcal{T}_{\mathcal{C}}(d_j) = c \cdot P(R|d_j, q) + \bar{c} \cdot P(\bar{R}|d_j, q) \tag{3.4}$$

The choice of ranking d_j before any other document d_i in answer to query q is then optimal *if and only if* for all $d_i \in RE \setminus \mathcal{RL}$:

$$\begin{aligned} \mathcal{T}_{\mathcal{C}}(d_j) &\leq \mathcal{T}_{\mathcal{C}}(d_i) \\ c \cdot P(R|d_j, q) + \bar{c} \cdot P(\bar{R}|d_j, q) &\leq c \cdot P(R|d_i, q) + \bar{c} \cdot P(\bar{R}|d_i, q) \\ c \cdot P(R|d_j, q) + \bar{c} \cdot \big(1 - P(R|d_j, q)\big) &\leq c \cdot P(R|d_i, q) + \bar{c} \cdot \big(1 - P(R|d_i, q)\big) \\ c \cdot P(R|d_j, q) - \overline{(}c) \cdot P(R|d_j, q) + \bar{c} &\leq c \cdot P(R|d_i, q) + \bar{c} - \bar{c}P(R|d_i, q) \\ (c - \bar{c}) \cdot P(R|d_j, q) &\leq (c - \bar{c}) \cdot P(R|d_i, q) \qquad (3.5) \end{aligned}$$

The cost of retrieving a relevant document is less then the cost of retrieving a non-relevant document (i. e. $c < \bar{c}$), because non-relevant documents are not useful to users as they have to endure examining them without finding information which addresses their information needs. Therefore, $c - \bar{c}$ is a negative quantity and inequality 3.5 becomes:

$$(|\bar{c} - c|) \cdot P(R|d_j, q) \geq (|\bar{c} - c|) \cdot P(R|d_i, q) \tag{3.6}$$

The magnitude of the costs' difference (i. e. $|\bar{c} - c|$) cancels out and can be anyway ignored for rank equivalence reasons, obtaining:

$$P(R|d_j, q) \geq P(R|d_i, q) \tag{3.7}$$

Ranking d_j before d_i is the optimal choice if the probability of relevance of d_j with respect to q is higher than that of d_i. From here, we can derive that the optimal ranking rule under Assumptions 1–5 is to rank documents in decreasing probability of relevance to the query: PRP is then the optimal ranking strategy.

□

Note that Assumption 1 can be relaxed and PRP can be extended to multivalued (or graded) relevance [Bookstein, 1993]. In fact, PRP is optimal also in the case of multivalued relevance if increasing retrieval costs are associated to decreasing values of relevance. The optimality of PRP can be demonstrated considering the expected costs of ranking a document: the demonstration is out of the scope of the thesis; we refer the interested reader to the work of Bookstein [1993].

3.5.2 The Probability Ranking Principle for Classes of Users

In the formulation and analysis of PRP we have assumed interpretation (b) of probability of relevance: the probability of relevance is a relationship between a single user and a class of documents. This is the leading interpretation in current Information Retrieval research, and is the one we adhere throughout this thesis. In this subsection, however, we examine the counter-interpretation (i. e. interpretation (a)), where relevance is thought as a relationship between a single document and a class of users. We provide the problem statement related to ranking documents under such an interpretation of (probability of) relevance and we examine the optimal ranking criterion under particular additional assumptions.

This ranking criterion resembles PRP's ranking rule, which has been defined in Section 3.5 assuming interpretation (b) of probability of relevance.

Under interpretation (a), it is assumed that the IR system has data available about the frequency previous users judged a document relevant to specific queries. That is, for each document, the system has previously collected user relevance judgements with respect to a query. Relevance judgements might be characterised with respect to each single user (i. e. document d has been judged relevant to query q by users u_1 and u_2, and not relevant by the remaining users) or to user-types (or classes), which define categories of users sharing common features or having similar needs. For example, with respect to an academic or University retrieval system, user-type U_1 might represent the class of users that are computer science students or researchers, U_2 users that are medical students or researchers, and so forth. Grouping users into user-types or classes allows to reduce the size of the considered user population.

Table 3.1: A document/user-type matrix representing the relevance judgements made on five documents by user-types $U_1, \dots, U_{14}$ with respect to query q. A full dot (i. e. $\bullet$) in a cell (i, j) of the matrix represents the case where document d_i has been found relevant by user-type U_j when issuing the query q. R' represents relevance according to interpretation (a) and $P(R'|d, q)$ encodes the probability that d is relevant to q.

User-types

		U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}	U_{13}	U_{14}	$P(R'\|d, q)$
Documents	d_1	●	●													$\frac{1}{7}$
	d_2			●	●	●	●									$\frac{2}{7}$
	d_3							●	●	●	●	●	●	●	●	$\frac{4}{7}$
	d_4	●		●	●			●	●	●	●					$\frac{1}{2}$
	d_5		●			●	●					●	●	●	●	$\frac{1}{2}$

Now, suppose an unknown user u (i. e. the system does not know which user-type u belongs to) issues a query q, and the system retrieves five documents for the query. Each document is associated with a relevance judgement with respect to each user-type, which has been collected during past operations. In Table 3.1 we provide an example of such a situation, where each document (a row in the table) has been judged relevant to q by specific user-types (columns in the table, indicated by $U_1, \ldots, U_{14}$), before q was issued by user u.

In which order should the retrieval system return documents to user u who issued query q? If the user-type of user u was known, the ideal retrieval strategy would have been to return all and only those documents that have been previously judged relevant to users belonging to the same user-type of u. However, recall that the user-type of u is unknown and the system lacks data for guessing u's user-type. Therefore, user u is considered equally likely to belong to any of the user-types $U_1, \ldots, U_{14}$. Stirling [1977] showed that if Assumptions 1–5 are casted within the context of interpretation (a) of probability of relevance, then the *optimal strategy* is to rank documents in order of decreasing probability of relevance, where the probability of relevance is computed according to the mean of relevance to user-types. For example, in the situation described in Table 3.1, document d_1 has been deemed relevant by two user-types out of fourteen, and therefore is likely to be relevant to user u with probability equal to $\frac{2}{14}$. In his work, Stirling refers to this ranking criterion as the Probability Ranking Rule (PRR); apart the different interpretation of probability of relevance, PRR resembles the same ranking criterion of PRP.

Returning to the example of Table 3.1, PRR would return to user u the following document ranking: d_3, d_4, d_5, d_2, d_1 (note that d_3, d_5, d_4, d_2, d_1 is also a valid ranking according to PRR, as d_4 and d_5 have the same probability of relevance).

3.6 Beyond PRP: When Assumptions Fail

In the previous sections we have discussed the optimality of PRP as document ranking approach. Its optimality is however tailored to Assumptions 1–5. But, what does happen if an assumption does not hold? And, when does an assumption fail to hold? Are these assumptions likely to be upheld within IR tasks? To answer these questions, we develop from the work of Gordon and Lenk [1992] and we analyse when PRP is sub-optimal, i. e. when it does not return the best ordering of documents with respect to the issued query.

In their analysis, Gordon and Lenk have identified three critical assumptions for the optimality of PRP. We report them as presented in Gordon and Lenk [1991, 1992], and then we draw links to Assumptions 1–5 as exposed in Section 3.5.

Calibration: Gordon and Lenk argued PRP assumes that systems are well calibrated, that is, that the estimations of the probability of relevance made by an IR system correspond to the users' assessments of relevance. In fact, if an IR system assigns high relevance probabilities to documents that are on the contrary found irrelevant by the users (and vice-versa, i. e. low relevance probabilities assigned to documents that are subsequently judged as relevant), then the computed predictive probability of relevance do not correspond to the users' assessments. In such situations, the performance of the system is far from optimal.

Independent Relevance: this assumption concerns with the behaviour of users when judging the relevance of documents. It states that a document's relevance assessment will not change after the user evaluates other documents. It also implies that documents are assessed in isolation, i. e. the relevance of a document is not influenced by the relevance assessment of other documents.

Certainty in Estimations: the probability of relevance of a document is known by the system with certainty and is reported as a scalar number.

3.6.1 Assumption of Calibration

The calibration assumption discussed by Gordon and Lenk is equivalent to Assumption 2 of Section 3.5: i. e. relevance probabilities are estimated as accurately as possible leading thus to systems that are well calibrated as the relevance estimates are as much adherent as possible to the relevance assessments provided by users. Gordon and Lenk [1992] showed that performance of ill-calibrated systems are not optimal, and they suggested how to detect ill-calibrated systems, given that users feedbacks about the documents' relevance assessments are available. With this respect, they propose a calibration score, that can be used to detected whether a system is ill-calibrated. Gordon and Lenk's idea is based on the score proposed by Brier [1950] to measure forecasting errors, and is equivalent to a quadratic loss function of the difference between user's relevance assessments and system's estimated probabilities. In practice, the score is formalised as[1]

$$BS = \frac{1}{n} \sum_{i=1}^{n} \left(J(R|d_i, q) - P(R|d_i, q)^2 \right) \tag{3.8}$$

where n is the total number of documents reported in the document ranking, and $J(R|d_i, q)$ is the relevance assessment of document d_i made by a user with respect to query q. Note that $J(R|d_i, q)$ is either 0 or 1, as relevance is assumed to be dichotomous (Assumption 1). A large value of Brier's score BS indicates an ill-calibrated system, because there is high discrepancy between the relevance assessments made by users and the probability of relevance estimated by a system.

In practice, the condition of upholding the calibration assumption is dependent on the system that is employed and the data that is used for computing relevance estimates. No statement about the calibration of a system can be made without user feedback, and calibration does not seem to relate to intrinsic characteristics of particular IR scenarios, or of systems. It is therefore impossible to say *a priori* whether a system is well calibrated or ill-calibrated, and thus whether the calibration assumption is held.

3.6.2 Assumption of Independent Relevance

The independence assumption is equivalent to Assumption 3, which posits that users' relevance assessments of documents are mutually independent. If documents are assessed independently, the relationship between query and document becomes “both necessary and sufficient to establish relevance” [Goffman, 1968]. And, we have shown in Section 3.5 that the optimal ranking is obtained by sorting documents with respect to their “absolute” probabilities to be relevant (to the query): the greater the probability, earlier the document should be presented to the user.

[1] Here we provide a formalisation of the Brier score that differers from the one given by Gordon and Lenk [1992]: this is because we consider a single document ranking, while they compute Brier's score over a number of queries. The latter can be obtained from the former by averaging the Brier's scores obtained over all the queries.

But, what happens if the independence assumption fails to hold? Suppose users do not assess documents independently, but they examine documents in a linear fashion, proceeding from the top of the document ranking, down to its bottom. Because relevance assessments are not independent, documents are not judged in isolation, and therefore the relevance of the current document depends also on the relevance of previous documents. Certain documents might be useless if examined after others. Suppose for example that two documents concerning the same topic of the query are nearly duplicates. If judged in isolation (i. e. assuming independent relevance), the documents would be assessed as being relevant. But, what would be the relevance judgement assigned by a user to the second document, if relevance is not judged in isolation? It is likely that the user assesses the second document as non-relevant, because the same information was already conveyed in the near-duplicate document that was ranked in a previous position, and therefore it does not provide any utility to the user. Conversely, it might happen that a document which would have been judged not-relevant if assessed in isolation, turns to be relevant after examining another document. Therefore, if the independence assumption does not hold, PRP is not optimal, because document and query alone are not enough anymore to assess relevance: conversely, relevance is also dependent on the documents that have been already ranked.

3.6.3 Assumption of Certainty in Relevance Estimations

The third assumption discussed by Gordon and Lenk concerns with the certainty in the estimation of relevance probabilities measured by an IR system. The assumption prescribes that probabilities of relevance are predicted as scalar values, as opposed to associate with each document a probability distribution that encodes the confidence or uncertainty of a system with respect to the relevance estimation. For example, consider the situation where documents d_1, d_2, d_3, d_4 and d_5 were retrieved in response to a query q with the following probabilities of relevance $P(R|d_i, q)$:

$$\begin{aligned}
P(R|d_1, q) &= 0.5 \\
P(R|d_2, q) &= 0.5 \\
P(R|d_3, q) &= 0.5 \\
P(R|d_4, q) &= 0.6 \\
P(R|d_5, q) &= 0.4
\end{aligned} \tag{3.9}$$

When ranking the previous documents, PRP generates the ordering

$$< d_4, d_1, d_2, d_3, d_5 > \tag{3.10}$$

but the orderings $< d_4, d_2, d_1, d_3, d_5 >$ and $< d_4, d_2, d_3, d_1, d_5 >$, etc, i. e. rankings where the positions of d_1, d_2 and d_3 are permuted among them, are equally acceptable, as they are estimated as equally likely to be relevant. The ranking formed by PRP is optimal if the assumption of certainty in the probability estimations holds (together with the others).

However, how would PRP rank documents if instead of point probability estimations (i. e. scalars), an IR system generates probability distributions?

Consider the situation where $P(R|d_1, q)$ is estimated with certainty, i. e.

$$Var(P(R|d_1, q)) = 0 \text{ (null variance)}$$

while the remaining probabilities are estimated according to random variables that are normally distributed with the following means and variances:

$$\begin{aligned}
E[P(R|d_2, q)] = 0.5, \text{ and } Var(P(R|d_2, q)) &= 0.05 \\
E[P(R|d_3, q)] = 0.5, \text{ and } Var(P(R|d_3, q)) &= 0.1 \\
E[P(R|d_4, q)] = 0.6, \text{ and } Var(P(R|d_4, q)) &= 0.08 \\
E[P(R|d_5, q)] = 0.4, \text{ and } Var(P(R|d_5, q)) &= 0.25
\end{aligned} \tag{3.11}$$

The probability distributions of documents $d_2, \ldots, d_5$ are reported in Figures 3.1(a) – 3.1(d). The horizontal axis (x-axis) corresponds to values of probability of relevance $P(R|d_i, q)$, while values on the vertical axis (y-axis) represent how likely the distribution for document d_i is to assume a certain value

$P(R|d_i, q) = x$. Figure 3.2 presents a comparison of the distribution of all five documents (note that document d_1 is represented just by a single point[1] in correspondence to 0.5 with probability 1 because its variance is null, i. e. the probability is estimated with certainty).

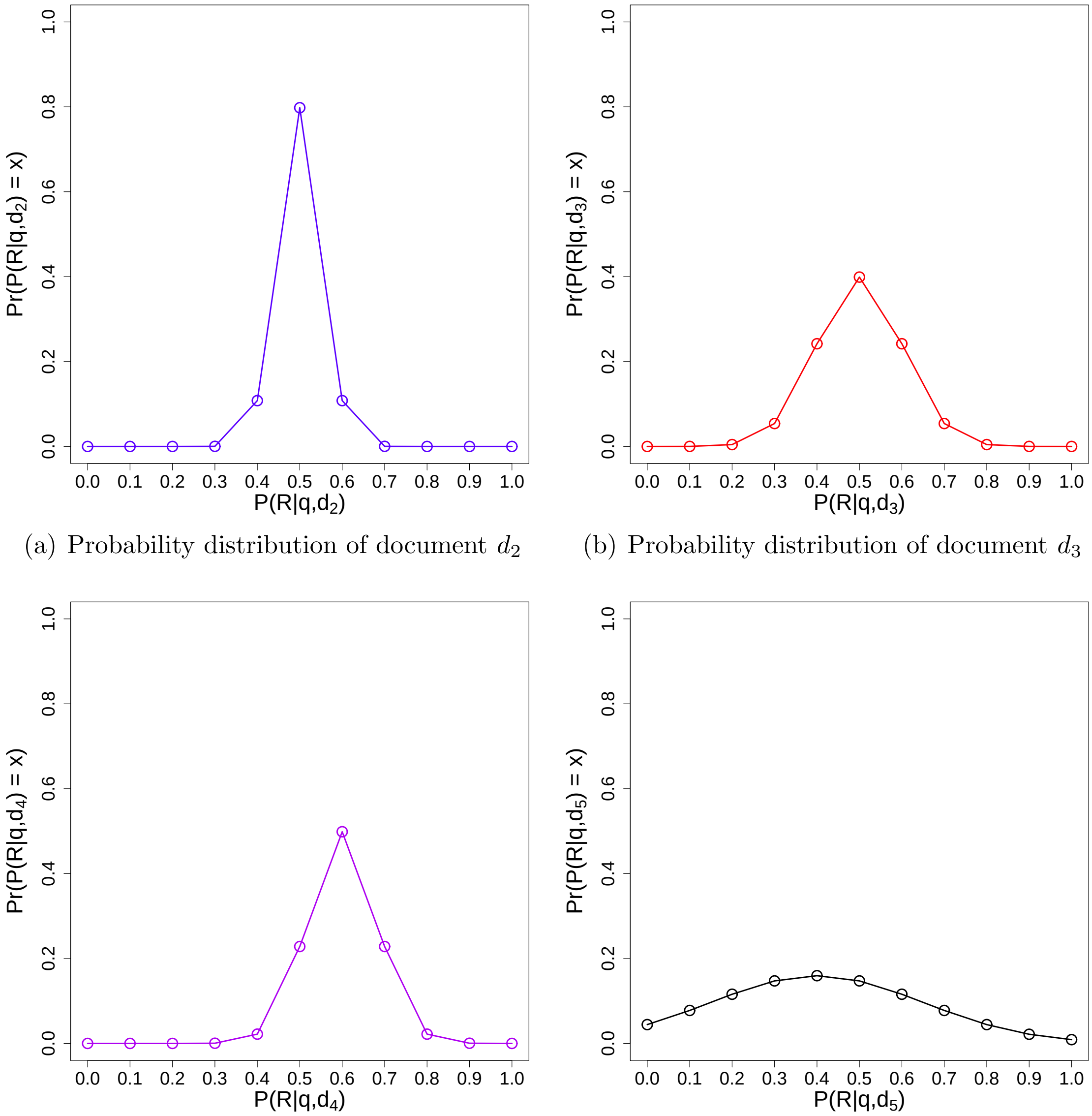
This example represents the case where an IR system estimates with certainty that the probability of relevance of d_1 is 0.5, but it can only estimate the relevance probabilities of the remaining documents with some degree of uncertainty, summarised by the corresponding variances. In particular, documents d_2 and d_3 have the same mean of d_1, i. e. 0.5, but their variance is greater than zero. Specifically, d_2 is estimated to have a probability of relevance equivalent to $P(R|d_2, q) = 0.6$ with probability of about 0.108 (similarly to when the probability of relevance is fixed at 0.4). While, d_3 has a probability of 0.242 to have a relevance probability of $P(R|d_3, q) = 0.6$ (and symmetrically, a relevance probability of 0.4). Documents d_4 and d_5 have respectively a higher and a lower mean when compared to d_1, d_2, d_3. Their variance is also different: d_4 has a skewed probability distribution, while d_5 is characterised by a wider bell-shape. Figure 3.2 shows that d_5 is more likely to have a high probability of relevance $P(R|d_i, q)$ than other documents. In fact, the probability that d_5 obtains a relevance estimation of 0.8 or above (i. e. $Pr\big(P(R|d_5, q) \geq 0.8\big)$) is higher than that corresponding to any other document (e.g., $Pr\big(P(R|d_4, q) = 0.8\big) = 0.022$ while $Pr\big(P(R|d_5, q) = 0.8\big) = 0.044$).

Given this example, we shall consider our initial question: *how would PRP rank documents if instead of point probability estimations (i. e. scalars), an IR system generates probability distributions?*.

Because PRP does not consider relevance distributions and thus does not cater for variance/uncertainty in the relevance estimates, we have to determine which point estimate PRP has to consider to generate a document ranking.

A first possibility is to *select the mean value* of each probability distribution. With respect to the relevance distributions outlined by Equations 3.9 and 3.11, and considering the mean as the source of the point estimate, PRP generates the ranking of Equation 3.10 or equivalents (i. e. $< d_4, d_2, d_1, d_3, d_5 >$, $< d_4, d_2, d_3, d_1, d_5 >$, etc): PRP ignores variances. Is this the optimal choice?

[1] We avoided to plot the points corresponding to values $x \in [0, 1] \setminus \{0.5\}$ for clarity in the figure: these points have zero ordinate.



(a) Probability distribution of document d_2

(b) Probability distribution of document d_3

(c) Probability distribution of document d_4

(d) Probability distribution of document d_5

Figure 3.1: Probability distributions of documents $d_2, \ldots, d_5$, following the example defined by Equation 3.11.

This policy might not generate the best document ranking. This is because uncertainty is associated with each probability estimation, and if relevance values less likely to occur than the mean values do instead occur, then there might be a better ranking than that of PRP. This is the case when the ranking that

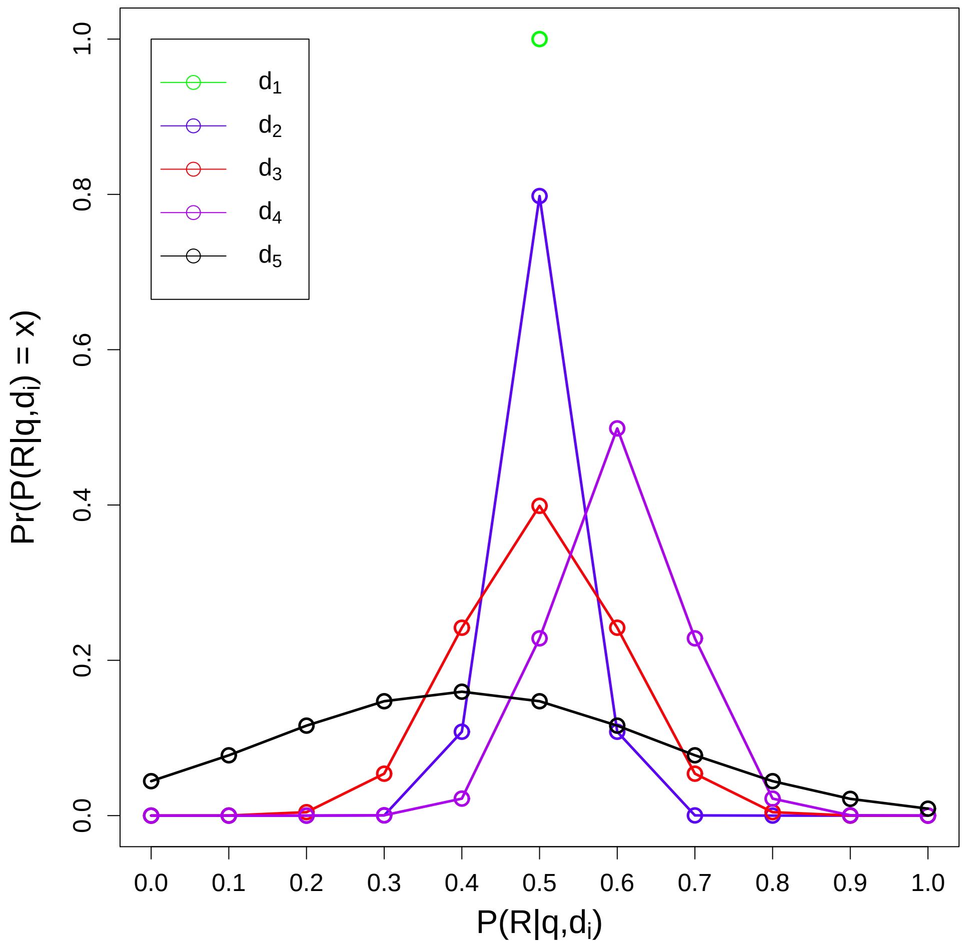


Figure 3.2: Comparison of the probability distributions of documents $d_1, \ldots, d_5$.

maximises the relevance probabilities differs from that created considering the means.

An alternative strategy for the selection of the point estimates to be used by PRP might be to *select the most probable value* of the probability of relevance, and ranking accordingly. With respect to the previous example, this policy would not change the outcome of the selection of the point estimate, as the relevance distributions resemble normal distributions. In this case, the most likely estimation also corresponds to the mean of the distribution. However, consider a further document, d_6, which is characterised as follows:

$$\begin{aligned} Pr(P(R|d_6,q)=x) &= \frac{1}{3}\cdot\left|x-\frac{1}{2}\right| \\ E[P(R|d_6,q)] &= 0.5 \\ Var(P(R|d_6,q)) &= 0.003 \end{aligned} \tag{3.12}$$

The relevance distribution of d_6 is plotted in Figure 3.3(a), while Figure 3.3(b) compares the relevance distribution of d_6 to those of the other documents we consider in the example. The mean value of the probability of relevance of d_6 and its most likely estimate differ. Specifically, while the mean value is 0.5, the most likely value of d_6's probability of relevance is either 0 or 1, with probabilities:

$$\max[Pr(P(R|d_6,q)=0)] = 0.1\bar{6}$$
$$\max[Pr(P(R|d_6,q)=1)] = 0.1\bar{6}$$

If the first strategy (i. e. consider the mean value) is used to derive a point estimate from a relevance distribution, then d_6 is regarded as being equivalent to documents d_1, d_2 and d_3: in fact, they all have mean equals to 0.5. PRP would then produce the ranking

$$< d_4, d_1, d_2, d_3, d_6, d_5 > \tag{3.13}$$

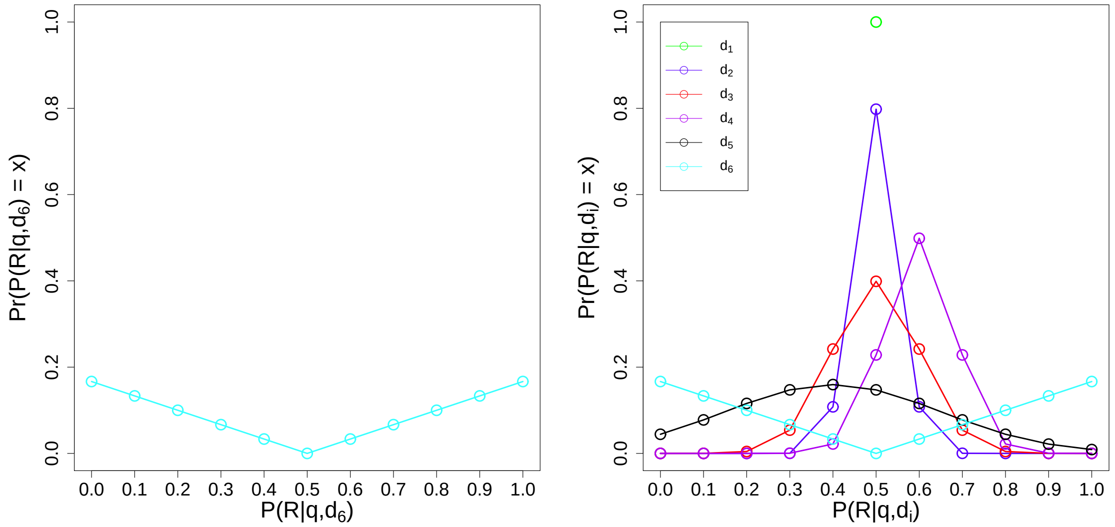
or equivalent permutations of the ranking (i. e. all the rankings containing permutations of the sub-list $< d_1, d_2, d_3, d_6 >$, preceded by d_4 and followed by d_5).

However, if the second strategy is employed (i. e. select the most probable value as point estimate of the relevance distribution) then the point estimate to be selected for d_6 does not correspond to the mean value of d_6's distribution. In fact, the most likely values of the probability of relevance of d_6 are 0 and 1. While, for documents $d_1, \dots, d_5$ the point estimates are left unchanged. Therefore, under these circumstances, PRP would provide either one of the following two rankings:

$$\begin{aligned} &< d_4, d_1, d_2, d_3, d_5, d_6 > \\ &< d_6, d_4, d_1, d_2, d_3, d_5 > \end{aligned} \tag{3.14}$$

or any equivalent permutation of these rankings obtained by permuting the orders of d_1, d_2 and d_3 among themselves. Note that rankings 3.13 and 3.14 are not equivalent.

Does the second strategy provide the optimal ranking then? This cannot be claimed: the optimal strategy in presence of variance in the relevance estimates is indeed ultimately dependant upon the level of risk a system or a user (or both) are willing to undertake. Because PRP does not cater for variance or uncertainty in the relevance estimation, it cannot appropriately model the risk associated to retrieve a document at a certain rank with respect to the likelihood of the point estimates sampled according to some arbitrary rule (such the two strategies discussed in the previous paragraphs) from the probability distribution. In Section 3.9 we shall examine a ranking strategy that accounts for such risks according to a mean-variance analysis of the retrieval results.



(a) Probability distribution of document d_6

(b) Comparison of the probability distributions of documents $d_1, \ldots, d_6$.

Figure 3.3: Probability distributions of documents $d_1, \ldots, d_6$, following the example defined by Equations 3.11 and 3.12.

The ranking scenario that considers relevance distributions is akin to Pandora's Problem, as noted by Varian [1999]. The problem consists in Pandora being faced with the choice of opening n boxes sequentially, and obtaining a reward for opening a box, which follows a random distribution. Also, a cost is associated with opening a box, and the reward obtained by opening a box is weighted by a monotonically diminishing function: i. e. the reward associated with a box decreases with delaying the choice of opening the box to a subsequent moment. This is equivalent to the IR scenario of a user achieving higher benefit when being presented with a relevant document at an earlier rank than at a later rank. In Pandora's Problem, the optimal box-opening strategy *is not* to open first those boxes with the highest expected benefit. In particular, Weitzman [1979] noted that[1]:

> "[o]ther things being equal, it is optimal to sample first from distributions that are more spread out of riskier in hopes to striking it rich early and ending the search."

Within the IR ranking scenario, Weitzman's statement should be revised with respect to the level of risk users accept to tolerate: as we shall see in Section 3.9, this intuition has been developed within IR from a similar perspective.

3.6.4 When do the Assumptions Fail?

So far we have observed that if one or more of the assumptions underlying PRP fails, then the ranking principle does not provide the optimal ranking in response to a user's information need. Next, we consider when the assumptions are likely to fail.

Independent Relevance. With respect to the assumption of independent relevance, it has been long noted that in many scenarios the relevance of a document depends upon what is already known at the time the document is examined by the user [Goffman, 1964]. Therefore, the relevance of a set of documents not only depends on the individual relevance of its elements, but also on the relationships between the documents. In particular, Eisenberg and Barry [1988] showed that in a typical relevance assessment activity the order of document presentation affects the relevance scores assigned to documents. This rules out the assumption

[1] See also Varian [1999].

that documents are judged in isolation and their relevance judgements are independent to each other, at least within the conditions considered in that study. Similar remarks, however, have been presented by Bookstein [1983]; Chen and Karger [2006]; Gordon and Lenk [1992], who argued that document relevance depends upon the information acquired during the course of the retrieval process. An example of such situation is when a user is presented, sequentially, with two identical documents that have been estimated as being the most relevant to the information need. While the user might judge the first document as being relevant, and therefore regarding it as useful, it is unlikely that he would provide the same judgement for the second document. That is, a duplicate document adds little (if any) information: this therefore might not be the optimal document ranking to present to a user. On the contrary, two documents that may be judged as non relevant if retrieved individually, may be relevant when taken together if they each tackle complementary aspects of the information need. The fact that a document has already been judged relevant by the user may provide some indication of the possible relevance of a subsequent document. Finally, even if the relevancy of a first document is unknown, the system may be aware of the presence of a correlation between the usefulness or relevance of the two documents.

Calibration. The calibration assumption states that the probability of relevance made by an IR system correspond to those assessed by users. However, it has been pointed out that this often is not the case. For example, in web search, Spink et al. [2001] have shown that web queries tend to be short and often ambiguous, and consequently the IR system is unlikely to be provided with enough evidence so as to generate well calibrated probability estimations. Furthermore, the fact that queries are often ambiguous implies that documents estimated as being highly relevant are also likely to refer to the most common interpretation of ambiguous queries. For example, the query "virus" might refer to: (1) the small infectious agent studied in medicine and life sciences; or (2) to the computer virus, i. e. computer programs, often malicious, that copy and replicate themselves within and across computers; or (3) to the movie titled "Virus" and directed by John Bruno[1]; or (4) the improbable Church of Virus[2]; etc.

Queries' ambiguity gives rise to the notion of *extrinsic diversity*: documents should be ranked so as to diversify the senses the information need is addressed, so as to account for the uncertainty associated with the information need.

It has also been observed that queries might be unambiguous, and yet the calibration assumption fail to hold. This is the case of queries that might be addressed with respect to, for example, different aspects, or sub-topics. Other information needs instead might imply that the user would get benefit from a ranking that encompasses different opinions on a topic: for example in product search users might find useful to be presented with reviews expressing different opinions about the product they are searching for. That is, the information need that is represented by the query submitted to an IR system is faceted. This gives rise to the notion of *intrinsic diversity*[Radlinski et al., 2009] and users would benefit from a ranking that covers different aspects of the information need. A system that assesses relevance solely through statistical features dependent on occurrences of query terms is unlikely to generate relevance estimates that are well calibrated with respect to relevance assessments made by users.

Supporting evidence and further discussion on the notion of extrinsic and intrinsic diversity are given by Radlinski et al. [2009]. While, Spärck-Jones et al. [2007] examined the implications of ambiguous queries in IR, and posited the need for ranking principles that move beyond PRP. This is further supported by the work of Sanderson [2008], who showed that traditional IR systems based on PRP are not able to cope with ambiguous queries (where the ambiguity is because of the word sense or of the reference aspect, or both). A similar conclusion was reported also by Chen and Karger [2006], who showed that PRP's approach fails to generate rankings that cover many aspects or subtopics of an ambiguous query, therefore being likely to fail in estimating well calibrate relevance probability if users are interested in aspects that are not reported by the system.

[1] See `http://www.virusthemovie.com/` and `http://www.imdb.com/title/tt0120458/`, last visited June 18, 2012.

[2] See `http://www.churchofvirus.org/`, last visited June 18, 2012.

Certainty in Relevance Estimations. This assumption prescribes that relevance probabilities are estimated by the IR system with certainty. This is often upheld when considering traditional IR retrieval models, where by construction (or definition) estimates of relevance depend on statistical occurrences of terms and no uncertainty is associated with such derivation. Consider for example the TF-IDF [Spärck-Jones, 1993], BM25 [Robertson and Walker, 1994], language model (LM) [Ponte and Croft, 1998], and Divergence From Randomness (DFR) [Amati and van Rijsbergen, 2002] weighting schemas: specific functions of term occurrences are used as estimators of relevance, and uncertainty in the estimations is not accommodated for within the weighting schemas. Specifically, TF-IDF tailors relevancy of documents to the frequencies of occurrence of query terms in documents (TF component) and their specificity as evaluated across the collection (IDF component). Similarly, BM25 relies on the eliteness of terms within documents, on documents lengths (e.g., the ratio between a specific document length and the average lengths of documents in the corpus), and on a saturation function that sets a limit on the contribution terms frequencies bring to the estimations of relevance. In LM, probabilistic language models are computed for queries and documents using term occurrence and collection statistics, and uncertainty is not considered when constructing the language models and deriving estimations of relevance. In DFR, the attention is shifted from term frequencies to the distribution of terms across documents, and relevance estimates are derived from implicit evidences that a specific term is not randomly distributed. Again, also in DFR-based weighting schemas, no uncertainty is associated with the relevance estimates.

In doing so, traditional weighting schema provide a point estimation of relevance (or probability of relevance). Within probabilistic models, this corresponds to the best guess generated by considering the maximum likelihood estimation or the posterior (relevance) probability estimation.

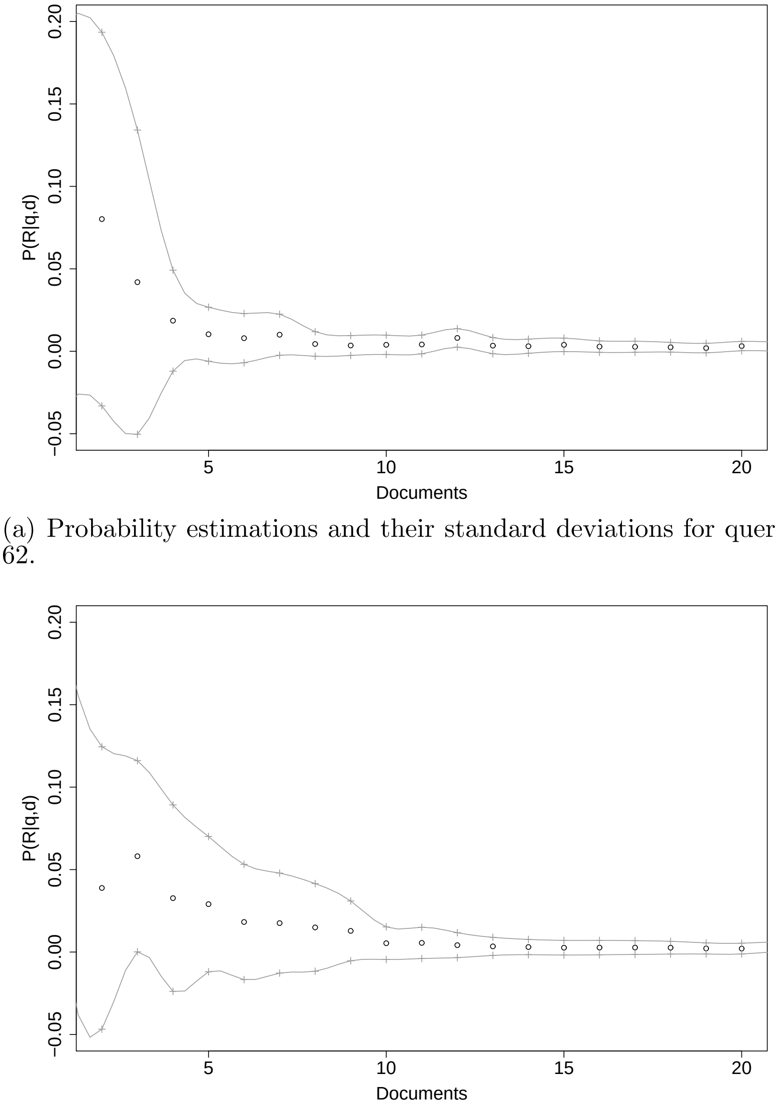
It is unclear, however, why uncertainty in the relevance estimation should not be considered by a retrieval function. In fact, different models and weighting schemas give rise to different estimation of relevance, and documents that might obtain a high probability of relevance with a particular schema might as well obtain a low relevance estimate when using an alternative schema. This has been noticed for example in data fusion, where rankings derived from different weighting schemas, models, or evidences are merged into a unique ranking [Croft, 2002; Frank Hsu and Taksa, 2005]. With this respect, we have shown in [Zuccon et al., 2011b] that aggregating the relevance estimates generated by IR systems, as those that participated to TREC 2009 Web Retrieval Track, and considering the second moment of the relevance distributions (i. e. the variances) give rise to a different and more effective ranking than that obtained by not considering higher moments. In that work, variances are employed to represent estimates' uncertainty. A similar argument had been put forward by Varian [1999], who argued that the "ordering [of documents] should depend not only on estimated first moment of the distribution [i. e. the mean values], but on higher moments as well".

We expand the analysis presented in [Zuccon et al., 2011b], and consider the variability of the scores assigned to retrieved documents as reported by systems that participated to TREC 2010 Web Retrieval Track. We argue that the variability in documents' scores is a reflection of the variability of the relevance estimates as provided by different systems. For example, in Figure 3.4(a) we report the means and the standard deviations of 20 documents that have been retrieved by various IR systems in answer to query 62 of TREC 2010 Web Retrieval Track. Similarly, Figure 3.4(b) reports the values obtained for query 86[1]. Note that different retrieval systems provide different estimations of documents' probability of relevance. The considered systems may differ with respect to the retrieval models and weighting schemas employed: often however, they differ only with respect to parameter settings, or evidences used to draw estimations from, in particular when considered systems employed by the same research group. This analysis shows that relevance probabilities differ widely across systems, which in turn might differ only with respect to parameter settings, and exhibit high variances.

Also Zhu et al. [2009] have observed that current IR models ignore the uncertainty associated with relevance estimates. To overcome this problem, they proposed a technique, inspired by Markowitz [1991] Portfolio Theory of financial markets and to the work of Wang and Zhu [2009] and Wang [2009] on a

[1] Details about the data used for this experiments are reported in Appendix B

Portfolio Theory for IR, that attempts to model uncertainty in the relevance estimation within a language model framework using an asymmetric loss function that represents the level of risk that one is willing to accept.



(a) Probability estimations and their standard deviations for query 62.

(b) Probability estimations and their standard deviations for query 86.

Figure 3.4: Probability estimations and their standard deviations for twenty documents retrieved by TREC systems for queries 62 (Figure 3.4(a)) and 86 (Figure 3.4(b)) of TREC 2010 Web Retrieval Track.

Empirical observations of PRP not being optimal. It has been observed that the key assumption underlying PRP does not always hold. Alternative ranking approaches must then be sought, in the following applications or scenarios:

- web search [Agrawal et al., 2009; Chen and Karger, 2006] and personalised web search [Radlinski and Dumais, 2006], e.g. Radlinski and Dumais [2006] develop three methods for improving personalised web search through the promotion of diverse documents in the top search results;
- image search, e.g. Muller et al. [2010] consider the need for diversity in cross-language image retrieval, and report about the development of a shared task, ImageCLEF, for the evaluation of IR systems that promote search result diversity in image retrieval;
- news and blog aggregation, e.g. Munson et al. [2009] argue that opinion and topic diversity can provide benefit when aggregating news; they investigate the need for diversity in this domain and propose algorithms for aggregating news considering both topicality and diversity;
- document summarisation, e.g. Carbonell and Goldstein [1998] propose a strategy that when applied to the task of document summarisation is able to select appropriate passages for text summarisation by reducing redundancy;
- biomedical passage retrieval, e.g. Andreopoulos et al. [2009] propose an extrinsic diversification method based on clustering for the retrieval of passages from the biomedical literature.

3.7 Alternative Ranking Approaches

Next we consider alternative ranking approaches to PRP. The alternatives we consider in this thesis share a common structure. Specifically, the following four elements can be identified:

(I) Relevance Estimation (rel. est.): the estimation of the probability of relevance as given by the underlying relevance model;

(II) Diversity Estimation (div. est.): the estimation of the extent a pair of documents differs (this might be biased by the query);

(III) Composition Function (comp. func.): the function that is used to mix (i. e. compose) the relevance estimation and the diversity estimation (e.g., addition, multiplication, etc);

(IV) Objective function (obj. func.): the final ranking criterion that is optimised during the sequential ranking process.

In particular, the common structure of ranking approaches alternative to PRP we consider in this thesis is as outlined by the following equation:

$$\mathfrak{C}(D,q) = \underbrace{\arg\max_{d \in RE\setminus\mathcal{RL}} \left[\overbrace{f(P(R|d,q))}^{\text{rel. est. (I)}} \underbrace{\circ}_{\text{comp. func. (III)}} \overbrace{g(d,d',q)}^{\text{div. est. (II)}} \right]}_{\text{obj. func. (IV)}} \tag{3.15}$$

As shown by Carterette [2011] and in Section 2.3.3, document ranking may become an NP-hard problem under specific circumstances. This is the case for the evaluation context of the diversity retrieval task (Section 2.3.3), which considers interdependent document relevance. We overcome the high computational costs involved in finding a perfect solution to an NP-hard problem, such as that of ranking under interdependent document relevance, by considering *sequential ranking* approaches, i. e. the ranking is formed by moving ahead through rank positions, without revising documents that have been ranked at previous ranks. Sequential ranking may not provide a global optimal solution, but it provides a trade-off between effectiveness and efficiency.

Next, we present the approaches that we shall study within this thesis. Further analytical relations among ranking approaches alternative to PRP, as well as the ranking principle that we shall formalise in Chapter 4 (i. e. the quantum PRP), will be explored in Chapter 6.

3.8 Maximal Marginal Relevance

Carbonell and Goldstein [1998] recognised that in specific search scenarios relevancy is not the only criteria that should be used to rank documents. While not connecting this intuition to PRP's assumptions, Carbonell and Goldstein argued that redundant documents might harm users' satisfaction, even if the redundant documents are relevant to the information need. In particular, they stated that ranked lists of document should be formed according to a criterion promoting relevance novelty. This objective is translated into a heuristic ranking approach, called (MMR), that first approximates relevance novelty by independently estimating relevance and novelty (or diversity) and then blends the estimations by means of a linear combination.

Formally, MMR ranks documents such that a document is selected to be retrieved at rank i if

$$d_i = \operatorname*{arg\,max}_{d \in RE \setminus \mathcal{RL}} \left(\lambda Sim(d,q) - (1-\lambda) \max_{d' \in \mathcal{RL}} Sim(d,d') \right) \tag{3.16}$$

where $Sim(d,q)$ is the similarity between document and query (q), $Sim(d,d')$ is a similarity function between documents, which is used as indicator of novelty (i. e. the fewer a pair of documents is similar, the more novel information one contains with respect to the other), and λ is the hyper-parameter which linearly combines query-similarity and document-similarity. Following this criterion, at each rank position the document that contains the highest marginal relevance, i. e. relevant information that is not similar to that already presented, is selected. The hyper-parameter can be inferred by the user's model. For example, values $\lambda < 0.5$ would characterise users with preference for document rankings which highly convey novel information, i. e. ideally addressing several facets of the information need; while, greater values of λ, i. e. $0.5 < \lambda < 1$, would be suited for users that prefer to focus on reinforcing relevant documents. In other words, as λ tends to one, the ranking criterion tends to those of PRP. Values of $\lambda = \{0, 1\}$ represent the limit situations where respectively only novelty is considered, rejecting any evidence provided by the relevance estimation, or only relevance is considered.

Several modifications or instantiations of MMR have been proposed that move away from the original approach of Carbonell and Goldstein [1998]; similarly, many applications and scenarios have been shown to benefit from MMR's ranking criterion.

Zhai and Lafferty [2006] framed MMR within a risk minimisation framework, where MMR's ranking intuition is encoded with a conditional loss function[1], aiming to balance the relevance and the novelty of a document with respect to a query. Leelanupab et al. [2010a,b] proposed a modification of MMR that considers average marginal relevance instead of maximal marginal relevance (i. e. the maximum similarity of documents, $\max_{d' \in RA} Sim(d, d')$ in eq. 3.16, is substituted by the average similarity of documents, $\text{avg}_{d' \in \mathcal{RL}} Sim(d, d')$) and where document-pairs similarities are derived from categorisation techniques (i. e. K-Means clustering, Probability Latent Semantic Indexing [Hofmann, 1999], Latent Dirichlet Allocation [Blei et al., 2003]). They showed that such an approach improves empirical performances in the context of Web diversity ranking and image diversity ranking. A similar study had been independently developed by He et al. [2011], leading to a cluster-based framework for ranking diversification, similar to that proposed by Leelanupab et al. [2010b], achieving, as the previous study reported, better retrieval performances when compared with rankings obtained by PRP. Moshfeghi et al. [2011] used the original formulation of MMR and a variant based on average marginal relevance to combine relevance estimates and document similarities as extracted from emotion features, and reported that these approaches provide increments in retrieval performances if compared with PRP in a text retrieval diversification context. Finally, Santos et al. [2012] have investigated the impact of novelty, as captured by MMR, on search result diversification. From the empirical evidence gathered on a standard web search scenario, they argued that novelty has little if no any effect when employed as diversification strategy. Santos et al. [2012] suggested that improvements in web search can be achieved when novelty-based approaches like MMR are deployed in combination with strategies that explicitly capture the coverage of the query's intents.

[1] i. e. the loss function associated with a document d_i conditionalised with respect to the language model formed by the query and the documents $d' \in \mathcal{RL}$ ranked until the current position.

3.9 Portfolio Theory for IR

Wang [2009]; Wang and Zhu [2009] examined the assumption of certainty in the estimation of relevance brought forward by PRP, and argued against this assumption. In particular, similarly to what we reported in Section 3.6.3, they stated that when estimating relevance, uncertainty in the estimation can arise because of the limited sample size or for estimation errors [Wang, 2009]. They also report that the independence assumption is not realistic in many situations. Their conclusion was that IR systems have to address uncertainty in the relevance estimation during the ranking process, as well as not assume independency between the relevance assessments.

To model both uncertainty and assessments' correlations within the ranking process, Wang and Zhu [2009] suggested an analogy with financial models used in economics, drawing a link with the Modern Portfolio Selection Theory of Markowitz [1991]. This financial theory prescribes that stocks or shares should not be selected only on the basis of the expected return. In fact, it is argued that by investing in more than one stock, an investor would achieve a reduction of the riskiness associated to the stock portfolio, benefitting from its diversification. In fact, the risk investors take when buying a stock is that the return will be lower than expected; this can be measured by the standard deviation or variance[1] from the expected value of the stock. Therefore, to minimise the risk of a portfolio, investors should not only invest capitals in more than one different stock, but they should also make sure that the different kind of stocks selected in their portfolio are not directly related. This is because the risk of a diversified portfolio will be less than the risk of holding only stocks of an individual kind.

The concept of portfolio diversification with respect to risk and relations between stocks have been transposed to IR giving rise to the Portfolio Theory (PT) ranking strategy Wang [2009]; Wang and Zhu [2009]. The ranking strategy aims to minimise the risk associated with ranking documents under uncertainty in their relevance estimates. This is achieved by balancing the expected relevance value and its variance. The resulting ranking criterion combines

[1] Recall that the standard deviation is the square root of the variance, which in turn is the average of the squared differences from the mean.

(i) the estimated document relevance,

(ii) an additive term which synthesises the risk inclination of the user,

(iii) the uncertainty (measured by the variance) associated with the probability estimation, and

(iv) the sum of the correlations between the candidate document and documents ranked in previous positions.

Formally, for each rank position i, documents are selected according to the following equation:

$$\begin{aligned} d_i &= \arg\max_{d \in RE \setminus \mathcal{RL}} \Bigg(E\Big[P(R|d,q)\Big] - bw_d \operatorname{var}\Big[P(R|d,q)\Big] + \\ & -2b \sum_{d' \in \mathcal{RL}} w_{d'} \operatorname{cov}\Big[P(R|d,q), P(R|d',q)\Big] \Bigg) \end{aligned} \tag{3.17}$$

where $E\Big[P(R|d,q)\Big]$, $\operatorname{var}\Big[P(R|d,q)\Big]$, and $\operatorname{cov}\Big[P(R|d,q), P(R|d',q)\Big]$ are respectively the mean value of the relevance estimation of document d to query q, the variance of these estimations, and the covariance of the relevance estimates for documents d and d'. Furthermore, b is a parametric coefficient that encodes the risk propensity of a user and w_d is a weight, inversely proportional to the rank position, which expresses the importance of the rank position itself. In particular, $b < 0$ encodes situations where users are incline to risk, while $b > 0$ represent risk-averse users; finally, when $b = 0$, then only the mean of the relevance estimates are considered, resembling a PRP-like ranking criterion.

Note that Equation 3.17 can be re-written using the fact that by definition the following relation holds with respect to the covariance of two random variables $\mathcal{X}$ and $\mathcal{Y}$:

$$\operatorname{cov}\Big[\mathcal{X}, \mathcal{Y}\Big] = \sigma(\mathcal{X})\sigma(\mathcal{Y})\rho(\mathcal{X}, \mathcal{Y}) \tag{3.18}$$

where $\sigma(\mathcal{X})$ is the standard deviation associated with $\mathcal{X}$ (i. e. $\sigma(\mathcal{X}) = \sqrt{\operatorname{var}\Big[\mathcal{X}\Big]}$) and $\rho(\mathcal{X}, \mathcal{Y})$ is the Pearson's correlation between the estimates of $\mathcal{X}$ and $\mathcal{Y}$. By plugging the previous relation into Equation 3.17, we obtain the following criterion:

$$\begin{aligned} d_i &= \operatorname*{arg\,max}_{d \in RE \backslash \mathcal{RL}} \Bigg(E\Big[P(R|d,q)\Big] - bw_d \sigma\big(P(R|d,q)\big)^2 + \\ &- 2b \sum_{d' \in \mathcal{RL}} w_{d'} \sigma\big(P(R|d,q)\big) \sigma\big(P(R|d',q)\big) \rho\big(P(R|d,q), P(R|d',q)\big)\Bigg) \end{aligned} \tag{3.19}$$

which in turn can be trivially re-written as a function of the mean value of the relevance estimates $P(R|d,q)$ and $P(R|d',q)$ using the relations between mean, standard deviation, and correlation:

$$\begin{aligned} d_i &= \operatorname*{arg\,max}_{d \in RE \backslash \mathcal{RL}} \Bigg(\overbrace{E\Big[P(R|d,q)\Big]}^{\text{first term}} - \overbrace{bw_d E\Big[\big(P(R|d,q) - E[P(R|d,q)]\big)^2\Big]}^{\text{second term}} + \\ &\underbrace{- \; 2b \sum_{d' \in \mathcal{RL}} w_{d'} E\Big[\big(P(R|d,q) - E[P(R|d,q)]\big)\big(P(R|d',q) - E[P(R|d',q)]\big)\Big]}_{\text{third term}}\Bigg) \end{aligned} \tag{3.20}$$

Following the criterion formalised in Equation 3.20, a document is ranked with respect to its relevance estimates and the relationships between these and the estimates of the documents ranked at previous position. However, in Section 2.2 we have observed that traditional IR models do not provide a number of estimates of the probability of relevance of a document to a query: instead, they provide a unique estimation of such probability. In such situation, the first term of Equation 3.20 corresponds to the unique estimate of the relevance probability provided by the system; while, the second and third terms are equivalent to zero[1], and Equation 3.20 resorts to PRP's ranking criterion (Equation 3.1). In order to remedy to the lack of estimations of $P(R|d,q)$, Wang and Zhu [2009] proposed to derive both variance and covariance (or standard deviation and correlation) from

[1] Because under that conditions the following equation holds:

$$E\Big[\big(P(R|d,q) - E[P(R|d,q)]\big)^2\Big] = E\Big[\big(P(R|d,q) - (P(R|d,q)\big)^2\Big] = E[0] = 0 \tag{3.21}$$

and similarly for $E\Big[\big(P(R|d,q) - E[P(R|d,q)]\big)\big(P(R|d',q) - E[P(R|d',q)]\big)\Big]$.

alternative evidences. In particular, the following approximations have been put forward:

- $\sigma\big(P(R|d,q)\big)$ is substituted by a quantity σ_d, which is either treated as a parameter to be learnt or optimised, or is approximated by a function of the variation of the length of document d [Zhu et al., 2009, see Section 3.2.2]. However, it is not clear how this approximation could be instantiated in practice, as Zhu et al. [2009] provided an approximation of the variance associated with the score contribution of a query term to the relevance estimation of a document provided by a portfolio-like language model technique.

- $\rho\big(P(R|d,q), P(R|d',q)\big)$ is substituted by the Pearson's correlation computed between the document term vectors of d and d', i. e. $\rho(d,d')$. The coefficient of such vectors could be arbitrarily derived employing a number of weighting schemas, e.g., TF, TF-IDF, BM25, etc.

Therefore, we rewrite Equation 3.20 with respect to the introduced approximations, obtaining the final criterion that is used when ranking documents according to PT:

$$d_i = \operatorname*{arg\,max}_{d \in RE \setminus \mathcal{RL}} \left(P(R|d,q) - b w_d \sigma_d^2 - 2b \sum_{d' \in \mathcal{RL}} w_{d'} \sigma_d \sigma_{d'} \rho(d,d') \right) \tag{3.22}$$

In summary, PT relaxes the assumptions of certainty in relevance estimates and independency in relevance assessments by considering mean, variance, and covariance of the estimations. The ranking criterion relies though on a number of parameters (e.g. the risk propensity of a user, b) and approximations (e.g. σ_d and $\rho(d,d')$).

Wang and Zhu [2009] showed that the criterion is successful in ranking documents both in the ad-hoc retrieval scenario and in the subtopic retrieval scenario, as defined by Zhai et al. [2003], providing improvements over the performance obtained by PRP. Similarly, Wang [2009] reported improvements in the collaborative filtering context. While, Aly et al. [2010] employed PT in the context of video story retrieval, using the TRECVID 2005 collection [Smeaton et al., 2006], and exhibiting performance gains with respect to alternative methods. Finally, we proposed a revision of PT's ranking approach, employing relevance estimates produced by different IR systems (namely, those participating to the TREC 2009 Web Track) to produce means, variances, and covariances [Zuccon et al., 2011b]. In that work, it was found that our reformulation of PT was effective for fusing document rankings provided by a number of different retrieval systems.

3.10 Interactive PRP

Fuhr [2008] explored ranking in Interactive IR scenarios, i. e. where users directly interact with the retrieval systems issuing and reformulating queries, examining results, and moving between different information needs. In particular, Fuhr noted that the independence assumption, central to the optimality of PRP, is certainly not valid in interactive settings, because, the author argued, relevance depends upon the previously seen documents and information. Regardless of the underlying model of interaction[1], Fuhr proposed to model ranking in Interactive IR as a decision making process that considers two actors: a user and an IR system. In particular, during the search process, users move between situations, while systems presents choices to users in each situation they reach. Choices are evaluated linearly, and when users select or disregard a choice, systems are required to produce new choices based upon the previous decision. The model considers also efforts and benefits: users' actions such as formulating a query or assessing a document entail an effort, or cost, while correct choices or decisions, such as examining a relevant document, imply benefits.

In the framework introduced by Fuhr [2008] the interactive PRP (iPRP) is derived as the optimum ordering of the choices presented in each situation. Note that iPRP is not a sequential ranking algorithm as those of Sections 3.8 and 3.9, because a ranking choice can be revised after an interaction has occurred. However, we might regard the formation of the ranking at each interaction step as sequential. Within this context, Zuccon et al. [2011a] proposed an effective instantiation of iPRP for a given situation and for each rank position i, iPRP can

[1] Four major models for describing how users interact with retrieval systems have been individuated in the IR literature: (i) the stratified model by Saracevic [1997]; (ii) the episodic model by Belkin [1996]; (iii) the interactive feedback and search process model by Spink [1997]; and (iv) the polyrepresentation model by Ingwersen [1993].

be instantiated such that a document d_i is selected following the criterion:

$$d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} \Big(e + P(R|d,q)(Q(Re|d)b_{d,i} + (1 - Q(Re|d))g) \Big) \tag{3.23}$$

where $Q(Re|d)$ is the probability that the user does not revise his choice of selecting document d, i. e. judges irrelevant document d after examining it, e is the effort of examining document d, g is the additional effort required for correction if the user judges a viewed document as irrelevant, and $b_{d,i}$ is the benefit of ranking document d at rank i if the document is relevant.

We develop further iPRP, considering what the ranking would be in the first retrieval pass, i. e. before any user interaction with the document ranking has occurred.

Under iPRP, it is assumed that users evaluate choices in linear order Fuhr [2008], thus previous documents would influence the ranking of the next document. Since, we are performing the ranking in the first instance on the system side, the costs associated with the user can be ignored. Therefore, both e, the effort of examining a document, and g, the effort for correction, are set to zero during the first pass. Furthermore, the probability $Q(Re|d)$ of a user not revising his choice can be treated as constant for all documents, and thus can be dropped for rank equivalence reasons. In our instantiation, the benefit of ranking document d at rank i is approximated by a similarity function between document-pairs, and in particular between d and the documents ranked until position $i-1$. The reason for this choice is twofold: on one hand, the benefit depends on the documents ranked previously, and this can be achieved through a similarity function. On the other hand, it allows us to compare the instantiation of iPRP with instantiations of other models for ranking interdependent documents, and in particular with MMR (which uses a generic similarity function) and PT (which in turn depends upon correlations between documents). Under these assumptions iPRP's ranking function reduces to:

$$\begin{aligned} d_i &= \arg\max_{d \in RE \setminus \mathcal{RL}} \big(P(R|d,q) b_{d,i} \big) = \\ & \arg\max_{d \in RE \setminus \mathcal{RL}} \Big(P(R|d,q) \frac{\sum_{d' \in \mathcal{RL}} Sim(d,d')}{|\mathcal{RL}|} \Big) \end{aligned} \tag{3.24}$$

where $Sim(d, d')$ represents a similarity function between document-pairs, $|\mathcal{RL}|$ is the size of the list containing the documents that have been ranked so far (i. e. until position $i - 1$; therefore $|\mathcal{RL}| = i - 1$).

3.11 Summary

In this chapter we discussed the problem of document ranking in Information Retrieval. We have first presented the motivations for producing a ranking of documents given a user's query, as opposed to return an unordered set of documents. We have then provided an algorithmic definition of a sequential ranking process. We have argued that although ideally IR systems aim to reach perfect ranking, i. e. return all and only the relevant documents as response to a query, this is not possible because the factors influencing relevance are complex and do evolve over time.

We instead focused on the notion of optimal ranking, where we are interested to know which ranking criterion yields the highest user satisfaction with respect to the ranked documents. We framed the ranking problem in terms of probability of relevance. In this context, we examined the probability ranking principle (PRP) that has played a central role in the development of Information Retrieval theories, models and systems. Specifically, we proved its optimality; however, because this depends upon a number of assumptions, we also examined them, showing what happens if the assumptions do not hold and discussing when this is likely to happen. In particular, we recognised that the independence assumption is critical for PRP's optimality. We also argued that this assumption often does not hold; this is the case in many IR applications, such as Web search, image search, news and blog aggregation, biomedical passage retrieval, etc. We pointed out how this assumption is linked with the notion of documents diversity in Information Retrieval, which represents one of the evaluation contexts where we shall compare different ranking approaches.

Finally, we have presented three alternative approaches to PRP that aim to overcome all or part of PRP's assumptions: Maximal Marginal Relevance, Modern Portfolio Theory, and Interactive PRP. Other approaches have also been proposed in the IR literature. However, we do not consider them in the context of this thesis, because we seek for alternative approaches that are as general as PRP, not relying on specific external evidences or additional techniques. Whereas, approaches other than MMR, MPT, iPRP (and the quantum PRP of Chapter 4) are (1) specific to particular search applications; or (2) make use of evidences drawn from sources others that the documents in the collection (i. e. evidences external to the document collection), as for example query logs, ontologies, Wikipedia pages, etc; or (3) rely on additional techniques, such as clustering and topic modelling. For example, Calegari and Pasi [2010] proposed a ranking diversification approach that uses a normalised granular view of an ontology: documents are associated to topical granules of the ontology, and the diversification technique strive to retrieve at top rank documents that may be categorised in topically different granules. While, Chandar and Carterette [2010] proposed and evaluated a diversification method that exploits the graph formed by the hyperlinks on the Web. Leelanupab et al. [2010b]'s approach to document ranking diversification rely instead on categorisation techniques such as Latent Dirichlet Allocation, probabilistic Latent Semantic Indexing, and K-Means clustering, for inferring which intents are covered by documents. Thereafter rankings are formed through instantiations of the Maximal Marginal Relevance approach using intra- and inter-category similarities.

Chapter 4

Ranking Documents with Quantum Probabilities: The Quantum Probability Ranking Principle

4.1 Introduction

In quantum theory, systems are represented by state vectors over a Hilbert space. Hilbert spaces[1] are a generalisation of Euclidean vector spaces to any finite or infinite number of dimensions. Formally, a Hilbert space $\mathcal{H}$ is a *complex*[2] vector space on which an inner product can be defined. A formal description of Hilbert spaces is given in Appendix C.

State vectors in Hilbert spaces can be projected into subspaces representing outcomes (i. e. physical quantities). In doing so, a probability measure is obtained, that assigns to a particular system the probability of being observed in a particular configuration. Logic relationships, and in particular conditionals in logic, can be represented as geometrical objects in Hilbert spaces. Quantum probability theory develops from these underpinnings, leading to numerous points of departure from the traditional Kolmogorovian probability theory.

Differences between the two probability theories arise when measuring *incompatible observables*. The incompatibility between two observables translates into the impossibility to simultaneously perform measurements on them: one

[1] We refer to the work of Cohen [1989] for a complete introduction to the basics of Hilbert spaces.

[2] That is, the components of the vectors belonging to the space $\mathcal{H}$ are complex values numbers $\mathbb{C}$.

measurement has to follow the other and measuring one observable can affect a subsequent measurement on other observables. This gives rise to phenomena of distortion or interference. From a mathematical perspective, when observables are incompatible, the probability of an outcome is not necessarily equal to the sum across the joint probabilities of all outcome combinations, as opposed to what Kolmogorovian probability prescribes.

The Probability Ranking Principle is based upon the axioms of Kolmogorovian probability theory, and does not consider incompatible observables. Moreover, its assumptions restrict the principle to consider document independently: relevance assessments are assumed to be gathered independently. This assumption in practice partially discards the sequential nature of document ranking, in the sense that the retrieval of a document at a specific rank position does not influence the ranking process at other positions. As discussed in Section 3.5.1, this behaviour may be optimal in the task of ad-hoc retrieval. However, doubts have been cast on the optimality of PRP in contexts other than ad-hoc retrieval, e.g. diversity retrieval [Stirling, 1977; Wang and Zhu, 2009; Zhai et al., 2003].

Because the mathematical framework of quantum theory (and in particular quantum probability theory) intrinsically accounts for incompatible observables, it may be suited for modelling document ranking in situations as those considered within the diversity retrieval task. An advantage of a document ranking criteria derived from quantum probability theory is that it would have formal mathematical underpinnings so that its suitability to the ranking task may be mathematically demonstrated, as opposed to heuristic approaches such as MMR (Section 3.8).

If a ranking principle is to be developed based upon quantum theory's mathematical underpinnings, an understanding of the differences between Kolmogorovian and quantum probability theory is necessary. To facilitate this, next we present a physical experiment that exemplifies the incompatibility between measurement and the rise of interference: the double slit experiment. Our discussion of the double slit experiment develops in a purely probabilistic manner. For an alternative discussion of the experiment in terms of Hilbert spaces, we refer the interested reader to Appendix C.

Once the experiment has been described and the presence of interference introduced from a physical perspective, we turn to consider how the document ranking process may be modelled using quantum theory. To do so we develop upon the double slit experiment, devising an analogy between the physical experiment itself and the document ranking process. Through the analogy we then develop the ranking rule of the quantum probability ranking principle, which is presented in Section 4.6.

4.2 The Double Slit Experiment

The double slit experiment is originally due to Thomas Young and had been used to demonstrate the wave theory of light. Its modern version, which considers atomic particles such as electrons[1], has been used within quantum theory to study interferences of individual particles and with particle detectors at the slits Feynman [1951]. In this section, we use the experiment to illustrate the nature of quantum probabilities and the presence of quantum interference term between probabilities associated to events. In Section 4.3 we use the double slit experiment to draw an analogy with the document ranking process.

The settings of the double slit experiment are pictured in Figure 4.1 and consists of the following. The right hand-side screen in Figure 4.1 serves as measuring device by recording the arrival position of particles on the screen. An emitter of particles is positioned in front of the measuring screen (left hand-side of Figure 4.1): the emitter releases particles, such as electrons, which are directed towards the measuring screen. An additional screen is placed between the emitter of particles and the measuring screen. The interposed screen is characterised by the presence of two slits, i. e. two holes in the screen. These are indicated by A and B in the schema of Figure 4.1.

The execution of the experiment consists in closing one of the slits, say B, while keeping the other slit, say A, open and let the source of particles emit a number of these. At the same time, the arrival distribution of particles is

[1] The use of electrons in the experiment was first realised by Clauss Jönsson in 1961 [Jönsson, 1974, this is a translation from German of the original paper that appeared in Zeitschrift für Physik, 161, 454]. Other particles may be used in the experiment, e.g. neutron, proton, photon, ion, etc.

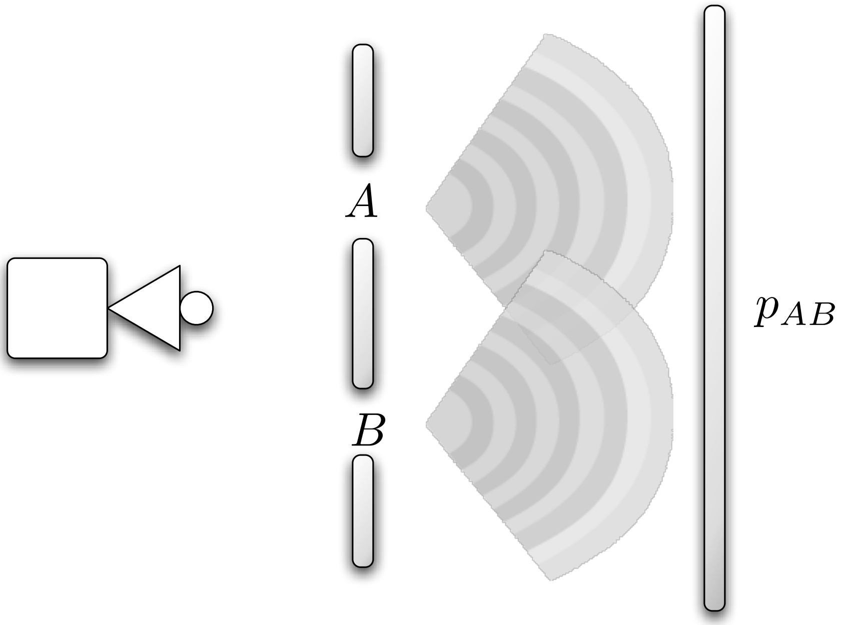


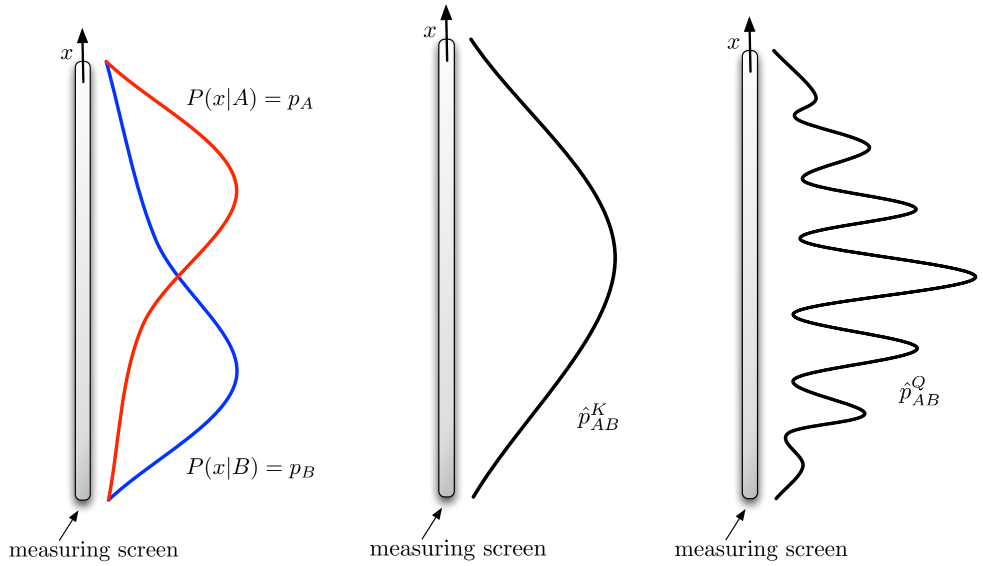
Figure 4.1: A schematic representation of the settings of the double slit experiment.

recorded on the measuring screen. Similar measurements are performed in the converse situation, i. e. when slit B is open and A is close. A final repetition of the experiment is then performed varying the configuration of the slits by letting both slits open.

What is of interested in the double slit experiment is to compare the arrival distributions of particle on the measuring screen in the three configurations.

The first observation that can be made is that the events "arrival of a particle at a location x on the measuring screen when slit A is open and B is closed" and "arrival of a particle at a location x on the measurement screen when list B is open and A is closed" are mutually exclusive and disjoint.

Probabilities can be associated to the previous events: $P(x|A)$ shall represent the probability of a particle being detected at x when A is open but B is closed, while $P(x|B)$ shall denote the probability of the other event. For simplicity of notation, in the following we shall indicate $P(x|A)$ with $p_A(x)$, and similarly for $P(x|B)$. Furthermore, we shall omit x when denoting the probabilities of a particle arriving at location x under specific conditions on the slits; we shall then



(a) Distribution of p_A and p_B in the double slit experiment.

(b) Distribution of $\hat{p}^{\mathcal{K}}_{AB}$ in the double slit experiment as estimated by Kolmogorovian probability.

(c) Distribution of p_{AB} as measured in the double slit experiment.

Figure 4.2: Figure (a) presents an example of the probability distributions measured in the double slit experiment in the situations "slit A open, slit B closed" (p_A) and "slit B open, slit A closed" (p_B). Figures (b) and (c) represent the estimation of the probability distribution associated with the event hitting the measuring screen under the condition "both slits open" as provided by Kolmogorovian probability ($\hat{p}^{\mathcal{K}}_{AB}$) and the probability distribution that is actually measured (p_{AB}), respectively.

use p_A for indicating $p_A(x)$, and similarly for p_B. An example of two possible distributions of p_A and p_B is given in Figure 4.2(a).

Central to the experiment is the probability distribution of the event of measuring an "arrival of a particle at a location x on the measurement screen when both lists A and B are open", i. e. $P(x|A, B)$. We shall indicate this probability as $p_{AB}(x)$, or simply p_{AB}.

Observe that particles are corpuscular elements: when both slits are open, a particle can pass through one slit only, although different particles can pass through different slits. It can be further observed that probability p_{AB} can be calculated considering the probabilities p_A and p_B. Next, we derive the estimations of joint probability p_{AB} given by the two probability theories, and we compare such estimations with the probability distribution that is actually measured in the settings of the double slit experiment.

Kolmogorovian probability theory. We consider first the case of Kolmogorovian probability theory. Assuming that the axioms of Kolmogorovian probability theory hold in the double slit experiment, the probability of a particle being measured in a particular location on the measuring screen when both slits A and B are open (i.e p_{AB}) is equal to the sum of the probabilities associated to the disjoint events of measuring a particle at that same location when only A is open and when only B is open. We indicate with $\hat{p}^{\mathcal{K}}_{AB}$ the probability distribution p_{AB} estimated using the axioms of Kolmogorovian probability theory. This can be formally derived as

$$\begin{aligned} \hat{p}^{\mathcal{K}}_{AB} &= P(\text{measuring a particle at x}|\text{slits A and B are open}) \\ &= P(\text{measuring a particle at x}|\text{slit A is open and B is closed}) \\ &+ P(\text{measuring a particle at x}|\text{slit B is open and A is closed}) \\ &= P(x|A) + P(x|B) \\ &= p_A + p_B \end{aligned} \tag{4.1}$$

If Kolmogorovian axioms hold in the context of the double slit experiment, it is possible to infer the value of p_{AB} (and its distribution over the measuring screen) from those of p_A and p_B using Equation 4.1. When the distribution of p_A and p_B depicted in Figure 4.2(a) is used, the predicted value $\hat{p}^{\mathcal{K}}_{AB}$ is equivalent to that pictured in Figure 4.2(b), and indeed corresponds to the sum of the two probability distributions associated with p_A and p_B.

Figure 4.2(c) pictures the actual probability distribution p_{AB} as measured in the settings of the double slit experiment. It is trivial to note that p_{AB} and $\hat{p}^{\mathcal{K}}_{AB}$ greatly differ. This suggests that the empirical measurements obtained for p_{AB} in the double slit experiment settings are not consistent with the predictions of Kolmogorovian probability theory, i. e.

$$p_{AB} \neq p_A + p_B = \hat{p}^{\mathcal{K}}_{AB}$$

in the settings of the double slit experiment described above.

Quantum probability theory. We now turn to consider quantum probability theory. The method that is used in quantum probability theory to calculate probabilities is radically different from that of Kolmogorovian probability theory. This is because a new concept is introduced in quantum probability theory: the complex probability amplitude. This concept is more primitive than that of probability, and it stems from the probability interpretation of the wave function proposed by Born. We shall indicate the amplitude associated to the event of measuring the "arrival of a particle at a location x on the measuring screen when slit A is open and B is closed" as $\phi(x|A)$. As before, we shall simplify the notation, indicating $\phi(x|A)$ with $\phi_A(x)$, or simply ϕ_A. Similarly, $\phi(x|B)$ and $\phi(x|A, B)$ characterise the settings when B is open while A is closed and those when both A and B are open. We shall indicate them with the notation ϕ_B and ϕ_{AB}, respectively.

Amplitudes are complex numbers, and therefore are characterised by a magnitude and a phase. A mathematical link can be established between probability amplitudes and probabilities. Given the complex probability amplitude of an event X, the probability of the event is given by the square of the modulo of the amplitude; formally:

$$p(X) = |\phi(X)|^2 \tag{4.2}$$

We now reconsider the double slit experiment using the notion of probability amplitude. In particular the probability amplitudes ϕ_A and ϕ_B can be measured from the empirical data. Similarly to p_{AB} in the Kolmogorovian case, we can compute ϕ_{AB} in the quantum case using ϕ_A and ϕ_B. In particular, quantum probability rules states that the probability amplitude ϕ_{AB} is calculated as the square modulo of the sum of the probability amplitudes of the disjoint events (see for example Feynman [1951]):

$$\begin{aligned}
\phi_{AB} &= \phi(\text{measuring a particle at x}|\text{slits A and B are open}) \\
&= \phi(\text{measuring a particle at x}|\text{slit A is open and B is closed}) \\
&+ \phi(\text{measuring a particle at x}|\text{slit B is open and A is closed}) \\
&= \phi_A + \phi_B
\end{aligned} \tag{4.3}$$

where $\phi(x|y)$ is read as "probability amplitude of event x given event y" (in analogy to $P(x|y)$ representing the probability of event x occurring given that event y has occurred).

Following the rules of quantum probability theory, it is then possible to compute the value of the probability p_{AB} from the observations of ϕ_A and ϕ_B and applying the rule of Equation 4.2. The estimation obtained in this way is denoted by $\hat{p}^{\mathrm{Q}}_{AB}$, indicating that the estimation has been obtained using quantum probability theory as opposed to Kolmogorovian. Thus, the probability of the joint event corresponds to the square modulo of the joint probability amplitude, which in turns is the sum of the probability amplitudes ϕ_A and ϕ_B, as Equation 4.3 showed.

It is possible to relate $\hat{p}^{\mathrm{Q}}_{AB}$ to the probabilities p_A and p_B. This requires some algebraic calculations involving *complex numbers*: recall in fact that probability amplitudes are complex numbers. In particular, a complex number z can be written in the form $re^{i\theta} = r\cos\theta + i\sin\theta$ (called polar form), where r is the *magnitude* of z, θ is its *phase*, and i is the *imaginary unit*[1]. This is called the polar form of a complex number. Alternatively, z can be written as $z = a + ib$, where $a, b \in \mathbb{R}$ are called respectively the *real part* of z and the *imaginary part* of z. The *modulus* of a complex number is denoted by $|z|$ and is defined as the square root of the sum of the squares of its real and imaginary parts, i. e. $|z| = \sqrt{a^2 + b^2}$. Similarly to what is observed for complex-valued vectors and inner products, every complex number has a *complex conjugate*, represented by $\bar{z}$ and defined as $\bar{z} = a - ib$. Finally, the following equalities are needed to re-write $\hat{p}^{Q}_{AB}$with respect to p_A and p_B:[2]

- $$z\bar{z} = |z|^2 \tag{4.4}$$
- $$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2} \tag{4.5}$$
- $$\bar{z} = re^{-i\theta} \tag{4.6}$$
- $$|z|^2 = r^2 \tag{4.7}$$

[1] Recall that $i = \sqrt{-1}$ and $i^2 = -1$.

[2] The proofs of Equations 4.4-4.7 are reported in Appendix D.

Given the previous definitions and relations, the following equalities hold:

$$\begin{aligned}
\hat{p}^{\Omega}_{AB} &= |\phi_{AB}|^2 \\
&= \phi_{AB} \cdot \overline{\phi_{AB}} \\
&= (\phi_A + \phi_B) \cdot (\overline{\phi_A + \phi_B}) \\
&= (\phi_A + \phi_B) \cdot (\overline{\phi_A} + \overline{\phi_B}) \\
&= \phi_A\overline{\phi_A} + \phi_B\overline{\phi_B} + \phi_A\overline{\phi_B} + \phi_B\overline{\phi_A} \\
&= |\phi_A|^2 + |\phi_B|^2 + \phi_A\overline{\phi_B} + \phi_B\overline{\phi_A} \\
&= |\phi_A|^2 + |\phi_B|^2 + r_A e^{i\theta_A} r_B e^{-i\theta_B} + r_B e^{i\theta_B} r_A e^{-i\theta_A} \\
&= |\phi_A|^2 + |\phi_B|^2 + r_A r_B \cdot \left[e^{i(\theta_A - \theta_B)} + e^{-i(\theta_A - \theta_B)}\right] \\
&= |\phi_A|^2 + |\phi_B|^2 + 2 \cdot r_A r_B \cdot \cos(\theta_A - \theta_B) \\
&= p_A + p_B + 2 \cdot \sqrt{p_A}\sqrt{p_B} \cdot \cos(\theta_A - \theta_B) \qquad (4.8) \\
&= P(x|A) + P(x|B) + 2 \cdot \sqrt{P(x|A)}\sqrt{P(x|B)} \cdot \cos(\theta_A - \theta_B)
\end{aligned}$$

Note that $r_A = \sqrt{p_A}$ (similarly for p_B) because the following equivalences are true[1]:

$$r_A = \sqrt{r_A^2} = \sqrt{|\phi_A|^2} = \sqrt{p_A}$$

Therefore, Equation 4.8 represents the estimation of the probability p_{AB} provided following the rules of quantum probability theory. In particular, such estimation depends upon the values of p_A and p_B. However, an additional term has to be considered when calculating such estimation: $2 \cdot \sqrt{p_A}\sqrt{p_B} \cdot \cos(\theta_A - \theta_B)$. This is called the *quantum interference term*, or simply interference term. This term does not only depend on the values of the probabilities p_A and p_B. In fact, it also depends on the cosine of the angle formed by the difference of the phases associated to the underlying complex probability amplitudes ϕ_A and ϕ_B. In practice, the cosine modulates the interference term according to the phase difference between the amplitude involved. The phase difference is not bounded a priori (because θ_A – and similarly θ_B – can assume any value in $[0, 2\pi] + 2k\pi$, with $k \in \mathbb{Z}$). Nothing can then be said about the cosine of the phase difference, apart that it can only be a real value between -1 and 1.

[1] Recall Equation 4.7.

Equations 4.1 (i. e. $\hat{p}^{\mathcal{K}}_{AB}$) and 4.8 (i. e. $\hat{p}^{\mathcal{Q}}_{AB}$) only differ for the presence (or absence) of the quantum interference term. That is, when the axioms of Kolmogorovian probability theory are abandoned in favour of quantum probability theory, the predicted value of the probability of observing a particle at a location x on the measuring screen, under the condition "both slits open", differs for the presence of the interference term.

What is important to observe is that $\hat{p}^{\mathcal{Q}}_{AB}$ provides an exact prediction of the probability distribution that is found in the empirical data collected from the double slit experiment [Feynman, 1963]. In fact, the prediction of $\hat{p}^{\mathcal{Q}}_{AB}$ is consistent with observed distribution p_{AB}, represented in Figure 4.2(c). The double slit experiment proves then the existence of physical situations for which the concepts and the mathematical formalism of classical physics (and in particular of Kolmogorovian probability theory) are not adequate.

For simplicity of notation, in the following we indicate with θ_{AB} the phase difference between the probability amplitudes ϕ_A and ϕ_B, i. e. $\theta_{AB} = \theta_A - \theta_B$. Furthermore, we indicate with I_{AB} the quantum interference term that arises between ϕ_A and ϕ_B, i. e. $I_{AB} = 2 \cdot \sqrt{p_A}\sqrt{p_B} \cdot \cos(\theta_{AB})$. With this notation, Equation 4.8 can be restated as:

$$\begin{aligned} \hat{p}^{\mathcal{Q}}_{AB} &= p_A + p_B + \overbrace{2 \cdot \sqrt{p_A}\sqrt{p_B} \cdot \cos(\theta_{AB})}^{\text{quantum interference term}} \\ &= p_A + p_B + I_{AB} \end{aligned} \tag{4.9}$$

In this section we have presented a discussion of the double slit experiment from a purely probabilistic point of view. Nevertheless, Equation 4.8 can as well be obtained considering the states of the system of the double slit experiment as defined over a Hilbert space, where state vectors corresponding to the considered events are defined, and for which probabilities are derived. We report this derivation in Appendix C.

4.2.1 Further Considerations: $p_{AB} = \hat{p}^{\mathcal{Q}}_{AB} = \hat{p}^{\mathcal{K}}_{AB}$

We have shown that Kolmogorovian probability theory is *in general* inadequate to model what happens in situations such that described by the double slit experiment. There are however two specific circumstances for which the predictions

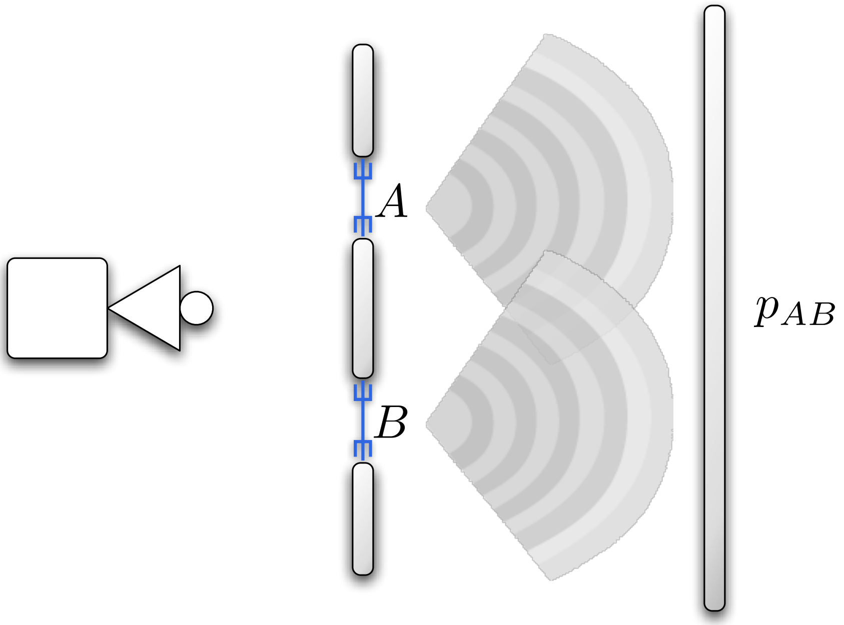


Figure 4.3: A setting of double slit experiment where detectors are *also* placed in correspondence with the two slits.

obtained using Kolmogorovian axioms, i. e. $\hat{p}^{\mathcal{K}}_{AB}$, are consistent with the data observed empirically.

The first situation is a particular case encoded within the quantum probability prediction, $\hat{p}^{\mathcal{Q}}_{AB}$. In fact, the equalities

$$p_{AB} = \hat{p}^{\mathcal{Q}}_{AB} = \hat{p}^{\mathcal{K}}_{AB}$$

are verified when the quantum interference term of Equation 4.8 is zero. This happens only when the cosine of the phase difference between the amplitudes ϕ_A and ϕ_B is zero[1]. That is, when the two phases are perpendicular to each other:

$$I_{AB} = 0 \Longleftrightarrow \cos(\theta_{AB}) = 0 \Longleftrightarrow \theta_{AB} = \frac{\pi}{2} + k\pi \tag{4.10}$$

with $p_A, p_B \neq 0$ and $k \in \mathbb{Z}$ (the set of integer numbers). It is trivial to observe in fact that if the interference is null, the predictions provided by the two theories are equivalent, resembling the data obtained empirically.

[1] We exclude a priori the trivial case when one, or both, probabilities p_A, p_B are zero.

A perhaps more interesting situation in which the Kolmogorovian prediction $\hat{p}^{\mathcal{K}}_{AB}$ resemble the distribution observed empirically is when the settings of the double slit experiment are slightly modified. The modification involves the position of the detectors. In the double slit experiment described in Section 4.2, the detectors of particles were attached to the rightmost screen, that effectively acted as a measuring device. Consider instead the settings of Figure 4.3, where to the detectors placed on the measurement screen are added also detectors in correspondence with the slits A and B. The same regimes of execution used previously for the double slit experiment can be repeated using the new setting, i. e. with one slit open per time (two situations: either A open or B open) or with both slits open. However, in the new settings and when both slits are open, we can know exactly from which slit the particles passed and there is no need to use the statistical data collected in the two configurations with only one slit open. As for the previous settings of the double slit experiment, we compare the Kolmogorovian prediction, $\hat{p}^{\mathcal{K}}_{AB}$, with the data that is empirically observed, p_{AB}. What is found in the new settings is that the Kolmogorovian prediction resembles the empirically observed data.

The previous result can be summarised as follows. If the system is never observed in any intermediate state (i. e. no detector is placed on the slits – original configuration of the experiment), p_{AB} is not given by the Kolmogorovian prediction[1]. In fact, the empirical measurements show evidence of interference effects, which can be adequately modelled by quantum probability theory. Conversely, if the intermediate states of the system are observed (i. e. p_A and p_B are directly measured by detectors placed on the slits – revisited configuration of the experiment) then interference effects are not found[2], and correct predictions are obtained using Kolmogorovian probability theory.

4.3 An Analogy with Document Ranking

In this section we propose an analogy between the double slit experiment and the document ranking process in IR. We do not claim or assume that quantum phenomena occurs in IR as they do in Physics; the goal of the analogy is to provide

[1] Apart the case of amplitudes with perpendicular phases.

[2] i. e. the quantum interference term is "washed out".

a simple physical view of document ranking that may help when modelling this IR process using quantum probabilities.

The analogy is the following. The particle corresponds to the user, who is characterised by an information need. The user examines documents for which expresses relevance judgements. Each slit corresponds to a document. Here we consider settings where only two slits are present: in the analogy this corresponds to only two documents being considered. In practical IR settings, more than two documents have to be considered for ranking. To accommodate this, we shall extend the analogy to consider more documents, alternatively more slits, later on in this chapter. Note however that the findings and considerations that apply for the double slit experiment in the usual configuration also apply to settings that involve more slits and, with respect to our analogy, more documents. In fact, we shall also consider the trivial case of screens with only one slit.

In the analogy, the event of a particle passing from the left of the screen to the right (i. e. through a slit) is comparable with the user examining a set of documents, e.g. read the associated snippets or the documents themselves. A measurement in the experiment corresponds to assessing the satisfaction of the user given the presented documents, or more concretely the decision of the user to stop or continue the search. Note that this interpretation of measurement in the double slit experiment applies to both settings of the experiment, i. e. when the detectors are placed on the measurement screen and when the detectors are placed in correspondence to the slits. Here however we consider the first setting only, although we shall use also the second setting when discussing how to derive PRP within the analogy.

Detecting a particle with probability p_{AB} (i. e. $P(x|A,B)$) on the measuring screen is analogous to being satisfied, or choosing to stop the search (represented by event S), with probability $P(S|d_A,d_B)$, after being presented with documents d_A and d_B. Similarly, $P(S|d_A)$ (alternatively, $P(S|d_B)$) corresponds to the likelihood of being satisfied by observing document d_A (d_B). As before, we simplify our notation, indicating with $p_{d_A d_B}$ the probability $P(S|d_A,d_B)$ (and similarly for p_{d_A} and p_{d_B}).

It can be argued that the probability of a document d_A inducing the user to stop searching because the information need has been satisfied by the document is proportional to the probability of relevance of the document to the information need, i. e. $P(S|d_A, d_B) = p_{d_A} \propto P(R|q, d_A)$, given that the user posed the query q as expression of the information need.

Note that in the physical experiment p_{AB} stands for the probability of detecting a particle at a particular *location* on the measuring screen (when both slits are open): location x. Similar observations are valid for p_A and p_B. To keep the analysis simple within the analogy, in the following we ignore the detection location x, and we assume that a specific location is considered, which produced the observed values of probability. We discuss the meaning of location x within the analogy in Section 4.8. For the same reason we ignore how the position of the measuring screen with respect to the particle emitter and the screen with the slits influences the values of the probabilities that are measured. Similarly, we also ignore the influence of the distance between the slits on the screen have on the values of the probabilities. These caveats are further discussed in Section 4.8. However, we take into consideration the fact that different slits[1] produce different probability distributions.

The settings of the analogy are pictured in Figure 4.4, where the user is presented with two documents d_A and d_B, which replace the slits that characterise the physical experiment. Probabilities p_{d_A} and p_{d_B} can be empirically measured, similarly to what happens in the first two executions of the physical experiments. In particular, in the analogy the action of opening a slit translates into showing the corresponding document to the user in the context of the analogy. Vice versa, masking or closing a slit translates to not showing the corresponding document. Therefore, within the analogy, probabilities p_{d_A} and p_{d_B} may be interpreted as the probabilities of the user being satisfied when examining document d_A (or d_B) in "isolation", i. e. independently from the other document.

In the context of the analogy, we observe the probabilities that correspond to two events: that of the user being *satisfied* by the presented documents, and that of the user being *not satisfied* with the presented documents. Similar considerations can be drawn if the events regarded the decision to stop or not the search, given the presented documents. The probability that the first event occurs is given by $p_{d_Ad_B}$. The probability of the opposite event (which we indicate with

[1] Slits may differ for example for their width.

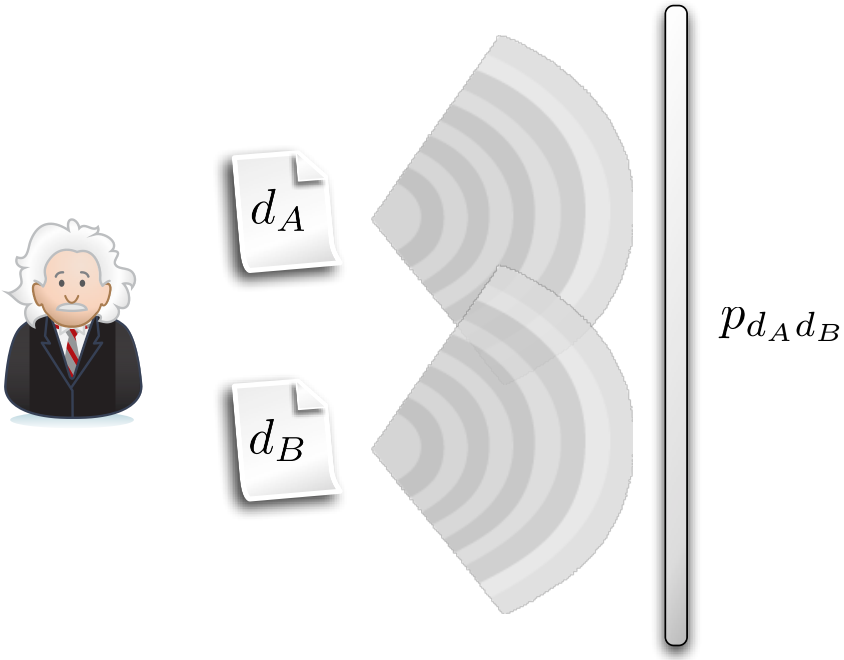


Figure 4.4: A schematic representation of the analogy between the double slit experiment and the document ranking process in IR.

$p_{\sim(d_A d_B)}$) is instead given by the counter probability $1 - p_{d_A d_B}$. In the physical experiment involving the double slit, the aim was to study p_{AB} and to show that quantum probability theory provides correct estimations of such probabilities, while the Kolmogorovian axioms are inadequate for modelling the outcomes of the experiments. Here, we take this result for granted. We are instead interested to explore the different ranking criteria that stem from the adoption of the analogy and of quantum probability theory as replacement of the Kolmogorovian one. This investigation is developed in the next section, and shall lead to two different ranking criteria, one of which resembles PRP, that are optimal under different circumstances or settings (thus ultimately, assumptions).

4.4 Ranking Documents within the Analogy

We consider the case where a document has been already ranked, and the algorithm implementing the ranking process needs to decide which document should

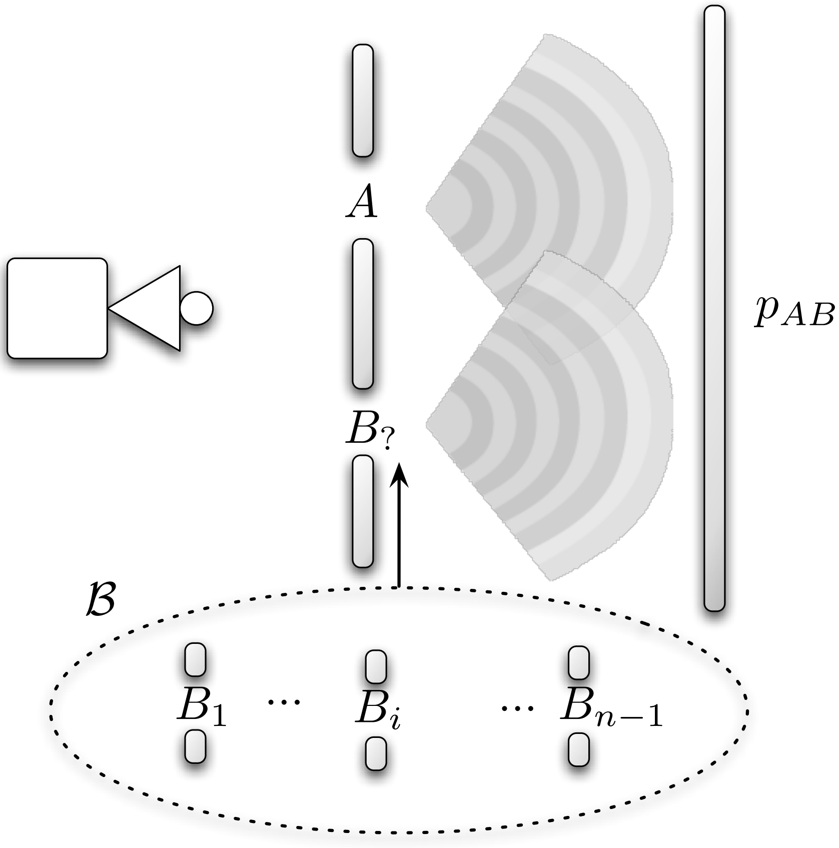


Figure 4.5: A setting of double slit experiment where slit A is kept fixed during different repetitions of the experiment, while the second slit is varied among the slits of set $\mathcal{B}$.

be ranked after the first among a set of candidate documents

$$\mathcal{B} = \{d_{B_1}, \ldots, d_{B_i}, \ldots, d_{B_{n-1}}\}$$

We defer the discussion of how to select the first document and how to select documents after the second one has been ranked later in this section.

Replicating the settings of our analogy with the double slit experiment, $n-1$ different experiments can be constructed, each characterised by a different document $d_B \in \mathcal{B}$. From a physical point of view, this corresponds in creating $n-1$ experiments, each characterised by a different slit $B \in \mathcal{B} = \{B_1, \ldots, B_i, \ldots, B_{n-1}\}$. This situation is pictured in Figure 4.5 (the IR counterpart is pictured in Figure 4.6). Different pairs of documents (slits) may generate different probabilities

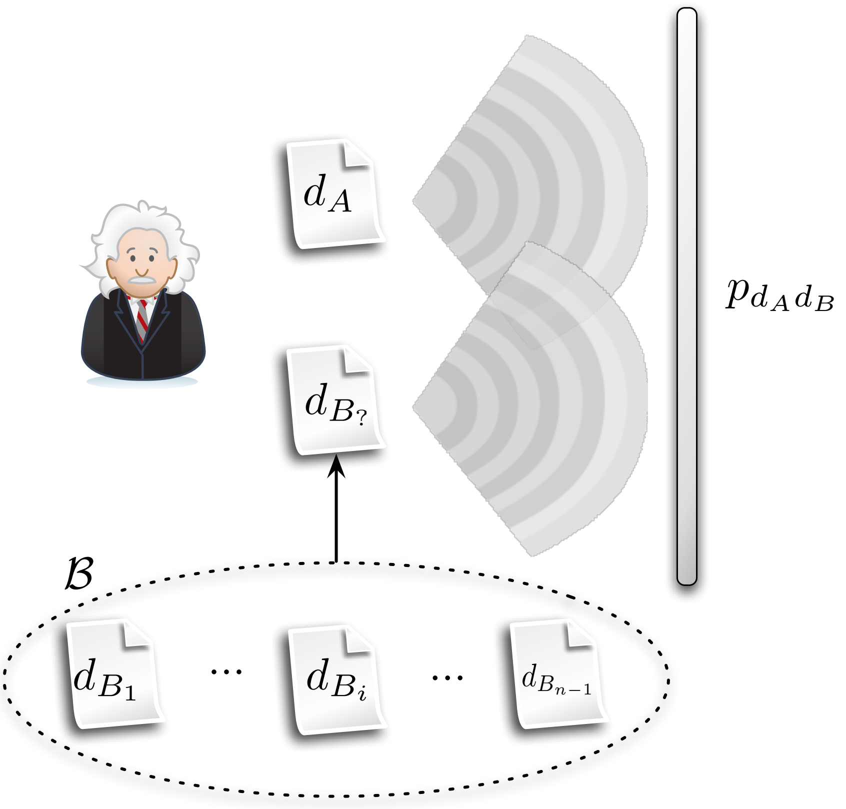


Figure 4.6: The IR analogous of the situation pictured in Figure 4.5.

$p_{d_A d_B}$: there is no reason for which to expect that $p_{d_A d_{B_i}} = p_{d_A d_{B_j}}, \forall d_{B_i}, d_{B_j} \in \mathcal{B}$. A question now arises: which document (slit) $d_B \in \mathcal{B}$ should be selected such that, once coupled with document (slit) d_A, the probability $p_{d_A d_B}$ is maximised? Answering this question corresponds to define a criteria for selecting documents such that, once coupled with the already ranked document, the likelihood of delivering maximum satisfaction to the user is maximised. In terms of the physical experiment, this would correspond to selecting slits among the set of available ones such that the likelihood of hitting the detector panel (at a particular location) is maximised.

Different probability theories provide different answers to the previous question. Therefore, the pair of documents that are obtained when using Kolmogorovian probability theory to model our analogy may be different from those obtained when using quantum probability theory. In the following we examine these two alternatives, and we derive the corresponding ranking criteria. Before that, recall that the probability of satisfaction induced by one or more documents can be linked to the probability of relevance and in general to the utility of the choice for the user, similarly to what discussed in Chapter 3. We shall use this in Section 4.7 where we prove that the ranking criteria derived in Section 4.4.2 is optimal, under specific circumstances.

4.4.1 Kolmogorovian Probability

We first consider the ranking criteria that is produced when using Kolmogorovian probability theory in the context of our analogy. To this aim, recall that the Kolmogorovian axioms provide the following estimation of p_{AB}:

$$\hat{p}^{\mathcal{K}}_{AB} = p_A + p_B$$

The following equalities can then be derived:

$$\begin{aligned}\operatorname*{argmax}_{B \in \mathcal{B}}(p_{AB}) &= \operatorname*{argmax}_{B \in \mathcal{B}}(\hat{p}^{\mathcal{K}}_{AB}) \\ &= \operatorname*{argmax}_{B \in \mathcal{B}}(p_A + p_B) \qquad (4.11) \\ &= \operatorname*{argmax}_{B \in \mathcal{B}}(p_B) \qquad (4.12)\end{aligned}$$

The passage from Equation 4.11 to Equation 4.12 is motivated by the fact that p_A is constant for every $B \in \mathcal{B}$, and therefore 4.11 and 4.12 are rank equivalent.

When Kolmogorovian probability theory is used to model the double slit experiment and a set of slits are available for being paired with slit A, the slit to select so as to maximise p_{AB} is the slit characterised by the highest p_{d_B} (probability of detection at the measuring screen when slit B is open and A is closed). In our analogy, this is equivalent to select the document with highest probability (of relevance, satisfaction, etc.) p_B among the set of documents that have not been ranked yet, i. e. $d_B \in \mathcal{B}$.

4.4.2 Quantum Probability

We now consider the ranking criteria that is produced when using quantum probability theory in the context of our analogy. To this aim, recall that when using quantum probability theory, the following estimation of p_{AB} is produced:

$$\hat{p}^{\mathrm{Q}}_{AB} = p_A + p_B + I_{AB}$$

Therefore, the following equalities can be derived:

$$\begin{aligned}
\operatorname*{argmax}_{B\in\mathcal{B}}\big(p_{AB}\big) &= \operatorname*{argmax}_{B\in\mathcal{B}}\big(\hat{p}^{\mathrm{Q}}_{AB}\big) \\
&= \operatorname*{argmax}_{B\in\mathcal{B}}\big(p_A + p_B + I_{AB}\big) && (4.13) \\
&= \operatorname*{argmax}_{B\in\mathcal{B}}\big(p_B + I_{AB}\big) && (4.14)
\end{aligned}$$

As for the derivations obtained in the previous section, Equation 4.13 and 4.14 are rank equivalent because p_A is constant for every $B \in \mathcal{B}$.

When quantum probability is used, the slit that maximises the probability of detection at the measuring screen, when both slits are open (i. e. p_{AB}) and slit A is fixed, is not the slit that is characterised by the highest value of p_B. Instead, quantum probability suggests that the slit that maximises p_{AB} has to be selected such that the sum of the probability of detection at the measuring screen when only that slit is open and the interference produced by the selected slit and slit A is maximised. Within our analogy, this translates as follows. When using quantum probability theory and when a first document d_A has been already selected in the first position of the ranking, the likelihood associated to a user being satisfied by the ranking of documents formed by d_A and a subsequent document d_B is maximised when the sum of the probability of relevance (or satisfaction, etc) associated to d_B and the interference between d_A and d_B is maximised.

Quantum interference appears to play a fundamental role both in the modelling of the physical experiment, and in the analogy between the experiment and the document ranking process in IR. Interference is a physical phenomena that is empirically observed in the experiment: how does this relate to IR? What is the interpretation of quantum interference within the settings of our analogy? We defer the discussion of these issues to Chapter 5, that explicitly deals with the interpretation of interference in IR and its calculation within the ranking principle that we shall propose later in this chapter.

4.4.3 Ranking the First Document and Ranking Subsequent Documents

In Sections 4.4.1 and 4.4.2 we have shown how documents would be ranked if the ranking process is cast within an analogy with the double slit experiments and Kolmogorovian probability theory and quantum probability theory are employed. In particular, we focused on the case where a first document has been retrieved and ranked at the first position of the ranking, and a second document has to be selected so as to provide the next document to rank. In order to generalise the observations made so far into a sound ranking principle, two situations have yet to be examined:

- the selection of the first document in the ranking;
- the selection of documents to be ranked at rank three, four, ..., n, i. e. the selection of subsequent documents.

4.4.3.1 First document

To analyse how the document to be ranked at the first position in the ranking should be selected according to our analogy, we modify the settings of the experiment.

We consider the situation where only one slit is present in the middle screen, i. e. the initial configuration with two slits is modified so as to accommodate only one slit. For example, let the middle screen have only slit A, as pictured in Figure 4.7. In order to reach the measuring screen, particles can only pass through slit A. At the measuring screen the probability p_A is observed. In these settings there is no sense in closing the only slit that is present, as no particles would be found to hit the measuring screen. The analogy with the document ranking process is directly derived from the new configuration by replacing the particle emitter with the user and the only slit A with a document d_A. This represents the situation where no document has yet been selected to be retrieved at the first position of the ranking.

Let $\mathcal{A} = \{A_1, \dots, A_i, \dots, A_n\}$ be the set of slits that are available. The following question can be asked: which slit $A \in \mathcal{A}$ does provide the highest probability of arrival at the measuring screen (given a particular location on the screen)? Recall that such probability is given by p_A. Also observe that in such settings, both Kolmogorovian probability theory and quantum probability theory provide the same estimations: i. e. $\hat{p}_A^{\mathcal{K}} = \hat{p}_A^{\mathcal{Q}} = p_A$. It is straightforward then to observe that both when employing Kolmogorovian probability theory and quantum probability theory, the slit that provides the highest likelihood of particles arriving at the measuring screen is that with highest probability p_A.

In the context of the analogy, this result implies that the document to be ranked at first position in the document ranking is the one characterised by the highest probability of relevance (or satisfaction, etc), regardless of the probability theory that is employed (i. e. Kolmogorovian or quantum).

4.4.3.2 Subsequent Documents: Kolmogorovian Case

So far it has been showen in the framework of the analogy how different probability theories select documents to be ranked at position 1 and 2 of a document ranking. Here, we focus on how subsequent documents should be ranked. To this aim, we extend the settings of the analogy between the double slit experiment and the document ranking process to the case of three slits, assuming that two slits have already been selected, and it is required to find which is the third slit, among a set of candidates, that should be selected so as to maximise the probability of particles reaching the measuring screen (at a specific location). Note that once the criteria for selecting three slits (documents) is found, it can be easily generalised to the case of ranking the n-th slit (document), given that $n-1$ slits (documents) have already been selected (retrieved).

The configuration of the experiment with three slits is pictured in Figure 4.8, and is as follows. Let A and B be the slits that have been already selected using the previous configurations of the experiment. Let C be the slit that has to be selected among a set of available slits, i. e. $C \in \mathcal{C} = \{C_1, \dots, C_i, \dots, C_{n-2}\}$. We are interested to find a criterion to select the slit $C \in \mathcal{C}$ that maximises p_{ABC},

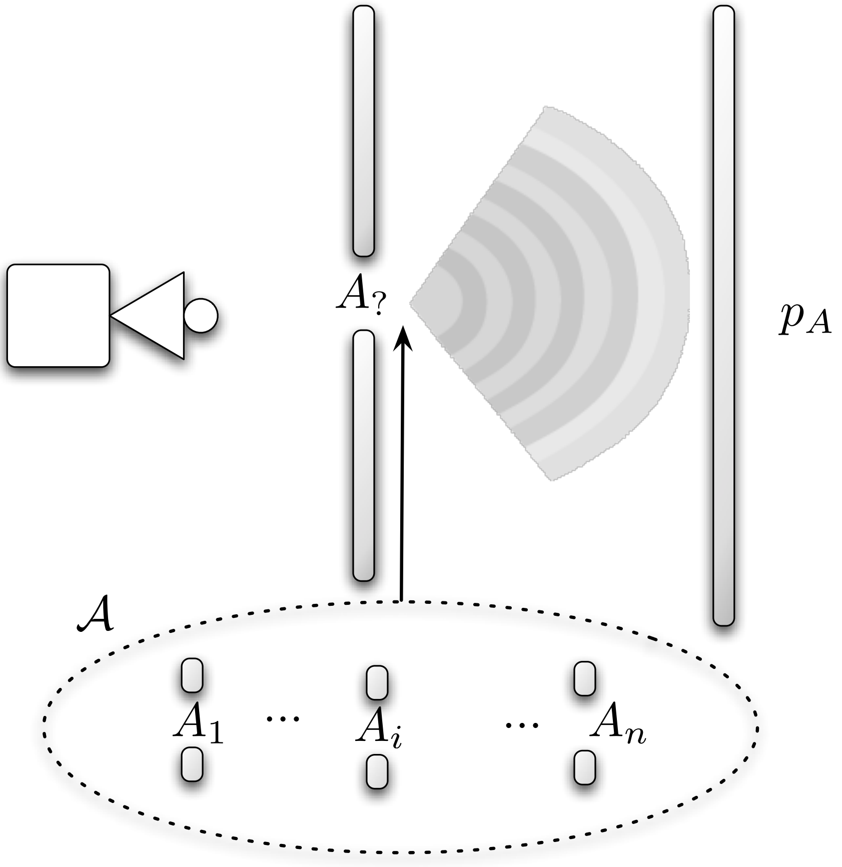


Figure 4.7: A setting of double slit experiment where only one slit is considered. During different repetitions of the experiment a different slit is selected among the slits contained in set $\mathcal{A}$, and the probability p_A is recorded.

i. e. the probability of a particle being detected at the measuring screen when all three slits A, B and C are open. To this aim we run three executions of the experiment, each characterised by a different slit kept open, while the others are closed. This procedure allows to record the probabilities p_A, p_B and p_C.

Here we follow the axioms of Kolmogorovian probability theory and we mathematically derive which slit C should be selected to maximise probability p_{ABC}. Recall that under the common Kolmogorovian axioms, the probability of a joint event is given by the sum of the probabilities associated to the disjoint events that lead to that event, i. e.

$$\hat{p}^{\mathcal{K}}_{ABC} = p_A + p_B + p_C$$

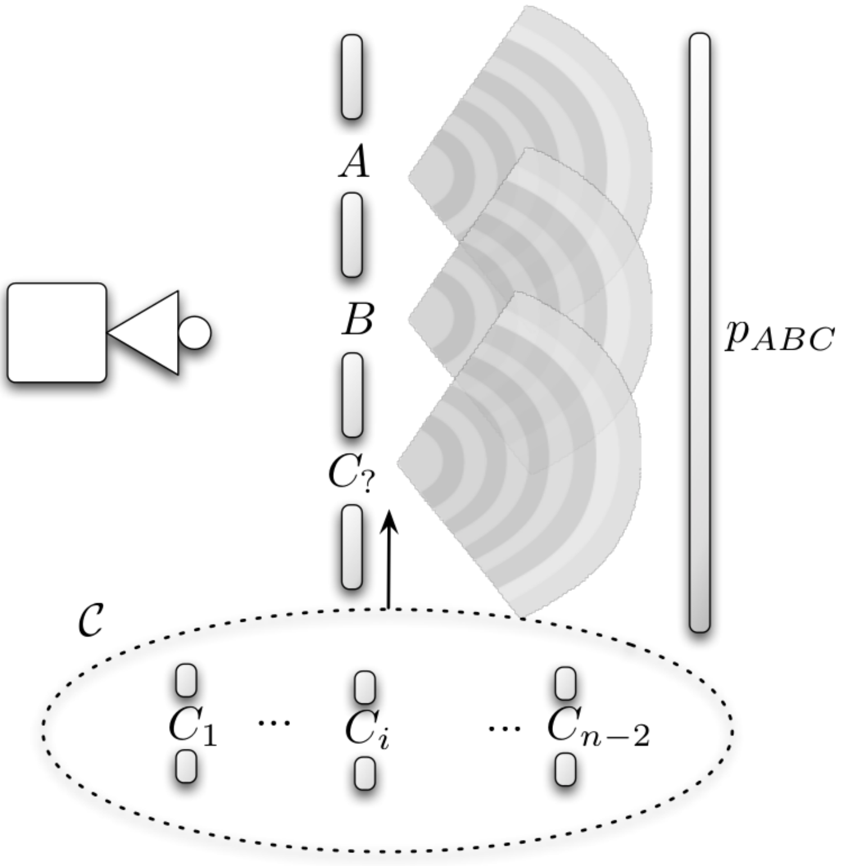


Figure 4.8: A setting of double slit experiment where slits A and B are kept fixed during different repetitions of the experiment, while the third slit is varied among the slits of set $\mathcal{C}$.

Therefore the condition for selecting the slit C that maximises p_{ABC} can be derived as follows:

$$\begin{aligned}
\underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_{ABC}\big) &= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(\hat{p}^{\mathcal{K}}_{ABC}\big) \\
&= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_A + p_B + p_C\big) && (4.15) \\
&= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_C\big) && (4.16)
\end{aligned}$$

where p_A and p_B can be ignored for rank equivalence reasons, i. e. their values are constant regardless of the slit $C \in \mathcal{C}$ that is used. When using the Kolmogorovian axioms, the probability of particles hitting the measuring screen when all three slits are open (two of which have been fixed in precedence) is maximum if the slit that produces the highest probability of hitting the measuring screen for particles when all slits are closed (apart itself), i. e. p_C, is used. In the context of the analogy, this translates in selecting the document with highest probability of relevance among the available ones.

This criteria can be generalised even further, regardless of the number of slits that are added to the experiment. In fact, following the Kolmogorovian axioms and given that m documents have been already ranked, the document that should be ranked at position $m+1$ is the one with highest probability of relevance among the $n-m$ documents that have not been ranked already.

4.4.3.3 Subsequent Documents: Quantum Case

Here we consider the case of the three slit experiment introduced in Section 4.4.3.2 when quantum probability theory is used instead of Kolmogorov's one. Note that the settings of the experiment are the same as those illustrated in the Kolmogorovian case.

When employing quantum probability theory, the following estimation of p_{ABC} is found:

$$
\begin{aligned}
\hat{p}^{\mathrm{Q}}_{ABC} &= |\phi_A + \phi_B + \phi_C|^2 \\
&= (\phi_A + \phi_B + \phi_C) \cdot \overline{(\phi_A + \phi_B + \phi_C)} \\
&= (\phi_A + \phi_B + \phi_C) \cdot (\overline{\phi_A} + \overline{\phi_B} + \overline{\phi_C})) \\
&= \phi_A\overline{\phi_A} + \phi_B\overline{\phi_B} + \phi_C\overline{\phi_C} + \phi_A\overline{\phi_B} + \phi_A\overline{\phi_C} + \phi_B\overline{\phi_A} + \phi_B\overline{\phi_C} + \phi_C\overline{\phi_A} + \phi_C\overline{\phi_B} \\
&= |\phi_A|^2 + |\phi_B|^2 + |\phi_C|^2 + r_A e^{i\theta_A} r_B e^{-i\theta_B} + r_A e^{i\theta_A} r_C e^{-i\theta_C} + r_B e^{i\theta_B} r_A e^{-i\theta_A} \\
&+ r_B e^{i\theta_B} r_C e^{-i\theta_C} + r_C e^{i\theta_C} r_A e^{-i\theta_A} + r_C e^{i\theta_C} r_B e^{-i\theta_B} \\
&= |\phi_A|^2 + |\phi_B|^2 + |\phi_C|^2 + 2 \cdot r_A r_B \cdot \cos(\theta_A - \theta_B) \\
&+ 2 \cdot r_A r_C \cdot \cos(\theta_A - \theta_C) + 2 \cdot r_B r_C \cdot \cos(\theta_B - \theta_C) \\
&= p_A + p_B + p_C + 2 \cdot \sqrt{p_A}\sqrt{p_B} \cdot \cos(\theta_A - \theta_B) \\
&+ 2 \cdot \sqrt{p_A}\sqrt{p_C} \cdot \cos(\theta_A - \theta_C) + 2 \cdot \sqrt{p_B}\sqrt{p_C} \cdot \cos(\theta_B - \theta_C) \\
&= p_A + p_B + p_C + I_{AB} + I_{AC} + I_{BC} \qquad (4.17)
\end{aligned}
$$

employing the same results used for the derivation of Equation 4.8.

This result can be generalised to n slits. Note in fact that no higher order interferences are created; that is, interference appears always among pairs of slits, while not involving tuples of slits. This is in accordance to empirical data experimentally collected in settings where more than two slits are used [Sinha et al., 2010].

Using the results of Equation 4.17, it is possible to obtain the criteria to select a slit $C \in \mathcal{C}$ such that p_{ABC} is maximum:

$$\begin{aligned}
\underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_{ABC}\big) &= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(\hat{p}^{\mathcal{Q}}_{ABC}\big) \\
&= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_A + p_B + p_C + I_{AB} + I_{AC} + I_{BC}\big) \qquad (4.18) \\
&= \underset{C\in\mathcal{C}}{\operatorname{argmax}}\big(p_C + I_{AC} + I_{BC}\big) \qquad (4.19)
\end{aligned}$$

Therefore, under the conditions studied in this section, the maximal p_{ABC} is achieved when selecting the slit $C \in \mathcal{C}$ that maximises the sum of the probability of the particles being detected at the measuring screen (at a specific location) when all slits are closed except C, and the interferences generated between C and the other slits included in the experiment.

Within the context of our analogy, Equation 4.19 suggests that when deciding upon which document should be ranked after documents d_A and d_B, the document that maximises the sum of its probability of relevance and of its interferences with documents d_A and d_B should be selected.

Similarly to the Kolmogorovian case, we can further generalise the result obtained for quantum probabilities. Following Equation 4.19 and assuming that m documents have been already ranked, the document that should be ranked at position $m+1$ is the document, among the $n-m$ documents that have not be ranked already, that maximises the sum of its probability of relevance and of its interferences with the m already ranked documents.

4.5 Resembling the Probability Ranking Principle

In Section 4.2.1 it has been observed that the probability estimate of p_{AB} produced by Kolmogorovian probability theory, i. e. $\hat{p}^{\mathcal{K}}_{AB}$, is consistent with the empirical measurements collected in the double slit experiment (and incidentally with $\hat{p}^{\text{Q}}_{AB}$) in two specific situations. These situations are the following:

1. the phases of the complex probability amplitudes ϕ_A and ϕ_B are perpendicular (and therefore $\cos(\theta_A - \theta_B = 0)$, or
2. detectors are also positioned in correspondence of the slits, such that the intermediate states of the system become observable.

How may these conditions be translated within the analogy between the physical experiment and the document ranking process in IR? In the physical context, two complex probability amplitudes having perpendicular phases entail no relationships or dependencies between the events they are associated with. When detectors are instead placed at the slits, the attempt of measuring the particles' positions (i. e. determine from which slit particles pass) disrupts the particles' trajectories. These intermediate measurements lead to empirical data that can be modelled by Kolmogorovian probability theory, thus eliminating any interference pattern. Within the context of our analogy, these findings may be translated in terms of document ranking as follows:

1. the event of assessing document d_A's relevancy is assumed to be "orthogonal" to the event of assessing d_B's relevancy, where the orthogonality characteristic indicates that no features are shared between the relevance assessments;
2. the judgement of relevance (or satisfaction) that a user provides for a document are assumed to be made in isolation, i. e. only the measurement of relevance on the selected document influences the relevance assessments.

Under any or both of these conditions, or assumptions, the optimal strategy is to select at each stage of the experiment (i. e. when considering one slit, two slits, three slits, ..., n slits) the slit that provides maximal probability of arrival on the measuring screen. Within the context of the analogy, this acquires the meaning of selecting at each rank position the document characterised by the highest probability of relevance. This ranking criteria resembles that of PRP.

4.6 The Quantum Probability Ranking Principle for IR

We have shown that in general the empirical data collected in the double slit experiment is best described using quantum probability theory. This happens in particular when detectors are only plugged to the measuring screen, and no intermediate measurements of the system's state are performed. In fact, although quantum probability can model the situation of detectors on the slits, the produced model does not differ from that obtained using Kolmogorovian probability theory.

When quantum probability theory is used to model the outcomes of the double slit experiment, and our analogy with document ranking is considered, a novel ranking criterion is produced. In particular, if sequential ranking and quantum probabilities are considered, quantum interference transpires, affecting how documents are ranked.

We can formalise these observations, and define the quantum Probability Ranking Principle (qPRP) and the assumptions that are made to guarantee its optimality. The optimality of the qPRP is discussed in Section 4.7.

Assumptions:

1. ranking is performed sequentially;
2. empirical data is best described using quantum probabilities;
3. the physical condition that detectors are only plugged to the measuring screen so that no intermediate measurements of the system's state are performed translates into the condition that the relevance of documents is not assessed in isolations. Vice versa, it is assumed that documents that have been ranked before may influence future relevance assessments.

Given these assumptions, the qPRP can be stated as follows.

Definition *The quantum probability ranking principle (qPRP):* Let $\mathcal{A}$ (where $\mathcal{A} = \varnothing$ is also considered) be the set containing the documents that have been already retrieved until rank $i - 1$ and let $\mathcal{B}$ be the set of candidate documents for being retrieved at rank i. If Assumptions 1–3 hold, in order to maximise its effectiveness, an Information Retrieval system has to rank document $d_B \in \mathcal{B}$ at rank i if and only if $p_{d_B} + I_{\mathcal{A}d_B} \geq p_{d_C} + I_{\mathcal{A}d_C}$, for any document $C \in \mathcal{C}$, where $I_{\mathcal{A}d_B}$ is the sum of the quantum interference terms produced considering all the pairs composed by document d_B and each document in $\mathcal{A}$ (similarly for $I_{\mathcal{A}d_C}$).

When following qPRP, at each rank positions documents would then be selected according to the following criteria:

$$\operatorname*{argmax}_{d_B \in \mathcal{B}} \big(p_{d_B} + \sum_{d_A \in \mathcal{A}} I_{d_A d_B}\big) \tag{4.20}$$

4.7 Optimality of qPRP within the analogy

Here we prove the optimality of the ranking criteria underlying qPRP within the settings and assumptions of the analogy when the data is best described using quantum probability theory. To this aim, we define the costs c and $\bar{c}$

$$\mathcal{C}(\text{retrieve}|\text{relevant}) = \quad c \tag{4.21}$$

$$\mathcal{C}(\text{retrieve}|\text{non-relevant}) = \quad \bar{c} \tag{4.22}$$

Note that within the analogy, cost c may be interpreted as the cost of selecting a slit given that particles hit the measuring screen (at a specific location), and similarly for $\bar{c}$. Costs are assumed to be positive and $\bar{c} > c$, i. e. the cost of retrieving a non-relevant document is higher than that of retrieving a relevant one.

The analogy suggests that the best choice for the document to rank after d_A among a set of candidate documents $\mathcal{B} = \{d_{B_1}, \dots, d_{B_{n-1}}\}$ is not the one for which p_{d_B} is maximal. Instead the ranking of two documents that maximises satisfaction is the one that takes also interference into account.

Within these settings, the total cost of ranking a document d_{B_j} after d_A is represented by $\mathcal{T}_{\mathcal{C}}(d_{B_j})$ and can be written as:

$$\mathcal{T}_{\mathcal{C}}(d_{B_j}) = c \cdot p_{d_A d_{B_j}} + \bar{c} \cdot (1 - p_{d_A d_{B_j}})$$

We now focus on the choice of ranking document d_{B_j} after d_A and before any other document d_{B_i}. This is equivalent to selecting slit B_j instead of any other slit B_i in the double slit configuration, given that slit A had been fixed on the screen. This choice is optimal in terms of minimising the total costs associated to the retrieval of documents if and only if, for any $d_{B_i} \in \mathcal{B}$ the following inequality is satisfied:

$$\begin{aligned}
\mathcal{T}_{\mathcal{C}}(B_j) &\leq \mathcal{T}_{\mathcal{C}}(B_i) \\
c \cdot p_{AB_j} + \bar{c} \cdot (1 - p_{AB_j}) &\leq c \cdot p_{AB_i} + \bar{c} \cdot (1 - p_{AB_i}) \\
c \cdot (p_{AB_j} - p_{AB_i}) + \bar{c} \cdot (1 - p_{AB_j} - 1 + p_{AB_i}) &\leq 0 \\
(c - \bar{c}) \cdot (p_{AB_j} - p_{AB_i}) &\leq 0 \qquad (4.23) \\
(p_{AB_j} - p_{AB_i}) &\geq 0 \qquad (4.24) \\
p_A + p_{B_j} + I_{AB_j} &\geq p_A + p_{B_i} + I_{AB_i} \\
p_{B_j} + I_{AB_j} &\geq p_{B_i} + I_{AB_i} \qquad (4.25)
\end{aligned}$$

where the passage from Inequality 4.23 to 4.24 is justified by the fact that $c - \bar{c} < 0$.

□

The optimality of PRP as ranking criteria within the analogy and under the conditions set in Section 4.5 can be demonstrated in a similar way. However the final inequality that shall be found differs from Inequality 4.25 for the absence of the interference terms. Note also that the demonstration obtained within the settings of the analogy resembles that of Section 3.5.1.

4.8 Caveats, Limitations and Discussion

The analogy between the double slit experiment and the document ranking process examined in this chapter led to the definition of a ranking principle alternative to PRP, the quantum Probability Ranking Principle. The two principles are characterised by different ranking functions, because when ranking qPRP also considers the presence of quantum interference, which is instead ignored in PRP.

In this section we discuss the caveats, limitations and issues of our analogy and of qPRP. We instead shall postpone to the next chapter the discussion of a key concept that appears within qPRP: quantum interference.

4.8.1 Role of location x

Within the analogy we ignored the role of the location x on the measuring screen in which the detectors are placed, i. e. where measurements are performed. While in the physical context $P(x|A)$ (or simply $p_A(x)$) is interpreted as the probability that a particle is detected, under specific conditions, at location x, it is unclear how the location of the detection may be translated in the IR context. We suggest the following view. We conjecture that the set of all possible locations on the measuring screen identifies *all the possible features that can determine the relevance assessment* of documents.

Many researchers argued that relevance is a multi-dimensional concept, which encompasses (but it is not limited to) topicality, novelty, understandability, reliability, authoritativeness, scope, appropriateness [Borlund, 2003; Cosijn and Ingwersen, 2000; Mizzaro, 1997; Xu and Chen, 2006]. Then, detecting a particle at a location x could be viewed as an analogous of assessing the relevance of a document with respect to the dimension (or feature) of relevance x. Often though relevance assessments are not only influenced by one specific dimension of relevance. For example, da Costa Pereira et al. [2009] considered four dimensions of relevance, assigning to each dimension a priority and ranking documents with respect to the extent they satisfy the combination of the relevance dimensions weighted by their priorities. To account for the influence of more than one dimension of relevance on the formulation of relevance assessments within the analogy, it is necessary to expand the measurements from one detector placed at x to multiple detectors placed at different locations. To accommodate this, measurements have to be performed on a volume V_x containing the r locations $x_1, \ldots, x_r$ of interest. This solution does not affect the findings obtained throughout this chapter. In fact, under the new conditions, the estimations of p_{AB} provided by the two probability theories become:

$$\hat{p}^{\mathcal{K}}_{AB} = \int_{V_x} p_A(x) + p_B(x) \cdot dx \tag{4.26}$$

$$\hat{p}^{\Omega}_{AB} = \int_{V_x} \left|\phi_A(x) + \phi_B(x)\right|^2 \cdot dx \tag{4.27}$$

while the interference term of Equation 4.9 becomes:

$$I_{AB} = 2 \cdot Re\Big(\int_{V_x} \overline{\phi}_A(x)\phi_B(x) \cdot dx\Big) \tag{4.28}$$

where $Re(.)$ indicates the real part of a complex number. The previous equations are the counterparts of Equations 4.1, 4.3 and 4.9 when considering continuos probability distributions over the surface of the measuring screen.

4.8.2 Factors that influence the magnitude of probabilities

In the physical experiment, many experimental details influence the magnitude of the probabilities that are recorded at the measuring screen. For example, Marcella [2002] reported the angular distribution of particles that are scattered from the slits under specific conditions. He observed that the distance between slits determines the number of interference fringes that are recorded at the measuring screen. Marcella also noticed that characteristics of the slits such as their widths influence the interference patterns that are measured in the physical experiment. Within the analogy between the experiment and document ranking, we ignored the influence that the characteristics of the slits and the distance between screens have on the magnitude of the probabilities that are recorded on the measuring screen (similarly for the magnitude of the interference). We did not map such characteristics and distances into aspects in the ranking process. However in the analogy we recognised that each list is characterised by different features, and that different slits provide different probabilities and interferences. Within the analogy, these had been translated as the intrinsic features of the documents associated to the slits: different documents provide different probability distributions and interferences, i. e. their relevance to the user is different.

4.8.3 Motivations for using quantum theory in IR

In Section 4.2 it has been shown that Kolmogorovian probability theory is not adequate to model and predict the empirical observation that are recorded in the double slit experiment. Similarly, Accardi [1984] showed that the whole quantum theory cannot be developed within the framework of Kolmogorovian probability theory, and quantum probability theory (and possible generalisations [Walach and von Stillfried, 2011]) is necessary. Accardi also proposed a method to distinguish the situations where Kolmogorovian probability theory is inapplicable.

This evidence may justify the use of quantum theory to model physical experiments, and support the thesis that Kolmogorovian probability is inadequate in such contexts. However, an open question is: is quantum theory (and in particular quantum probability theory) the right formalism to describe the empirical data observed in IR and in particular in IR processes such as document ranking?

Limited research have focused on this aspect. For example, Wang et al. [2010] studied whether interferences are present in relevance measurements with respect to a topic and caused by another topic. To uncover this, they analysed whether the probability of relevance of a document to a topic as judged by a user is influenced by a companion topic that is contemporarily presented to the user. From the empirical observations collected, they found that the relevance of a topic to a document is indeed affected by relevance of the companion topic to the same document; the extent of such influence depends on the companion topic that is presented to the user. Finally, the authors also showed that the predictions provided by quantum probability fit the empirical data observed in their experiments.

Di Buccio et al. [2011] instead studied the interactions between users and documents when assessing the relevance of documents to selected TREC topics. They focused on determining the level of uncertainty in the relevance assessments, claiming that the quantum probability framework, and in particular the notion of quantum interference can correctly update the predicted probability of relevance depending on the evolution of the user's relevance state and the user's uncertainty about the assessment. They showed that empirical data collected by their measurements exhibit the presence of quantum interference.

Note that attempts to model outcomes such that of the double slit experiment (and in general of physical experiments that exhibits outcomes not explainable by the Kolmogorovian axioms) by using classical hidden variables have been made in the past [Einstein et al., 1935; Hemmick, 1997]. Similar ideas have been applied also to IR: recall for example the Latent Dirichlet Allocation (LDA) technique, proposed by Blei et al. [2003], which models sets of observations (e.g. documents) using mixtures of unobserved variables (e.g. topics) encoding similarity between data-points (e.g. terms in documents). Both in Physics and IR, these hidden variables are supposed to the Kolmogorovian in nature, thus providing an alternative interpretation of observations than the one of quantum theory. In Physics, the Bell inequality [Bell, 1964] had been proposed to test this idea, and in particular to assess whether these hidden variables do exist by proving that empirical observations lead to the satisfaction of the inequality. Empirical observations have led to results that are inconsistent with the inequality, proving that hidden variables are not enough to describe phenomena captured by the modelling capabilities of quantum theory.

Outside Physics, a particular instance of the Bell inequality, called the Clauser-Horne-Shimony-Holt (CHSH) inequality [Clauser et al., 1969; Laloë, 2001], has been used by Kitto et al. [2010] to analyse the non-separability of bi-ambiguous compounds. Their work relates to IR, as provides insights on how to detect and cope with ambiguous compounds, which might be present, for examples, in users' queries. By applying the CHSH inequality to data collected in a concept combination experiment, they showed that concept combinations exhibit non-separable effects, akin to those of entangled quantum systems. This result suggests that classical cognitive models for concept combinations may not provide a complete account of the processes involved in human reasoning. Moreover, the CHSH inequality may be used to detected the non-separability of concept-combinations [Bruza et al., 2009a; Kitto et al., 2011].

4.8.4 Analysis of situations where interference is absent

In Section 4.2.1 we have discussed when Kolmogorovian and quantum probability theory provide the same probability estimations in the context of the double slit experiment. We observed that this situation happens in two cases: (i) when the probability amplitudes of two considered disjoint events have orthogonal phases, or (ii) when detectors are placed also in correspondence of the slits. Herbut [1992] developed these observations further, and derived a series of formal conditions on the presence of interference. In particular, Herbut stated that a *sufficient* condition for the *absence* of interference is the compatibility between the disjoint events considered in the double slit experiment. He also observed that another *sufficient* condition is the compatibility between the quantum state of the particle and the considered disjoint events. Finally, the author observed that the *necessary* conditions for the *presence* of quantum interference are that (1) the considered disjoint events are incompatible, and (2) the quantum state of the particle and the considered disjoint events are incompatible. It is clear then that *incompatibility* is a central condition for the presence of quantum interference. How incompatibility may be translated in IR terms, or within the document ranking process, is yet an open question. For example, van Rijsbergen [2004] investigated how incompatibility between observables, geometrically represented by projectors on a Hilbert space, translates in terms of logical implications between propositions. He showed incompatible observables imply that the logical disjunction does not respect classical (i. e. Boolean) logic. The author suggested that this is important to IR because if the notions of relevance and topicality are associated (respectively) to two observables, then "observing relevance followed by topicality is not the same [, in the framework of Hilbert spaces,] as observing topicality followed by relevance" [van Rijsbergen, 2004, page 67][1]. Nevertheless, within qPRP we assume that events are incompatible and thus that empirical data is best describable by means of quantum probability theory.

4.9 Summary

In this chapter we investigated how to rank documents according to quantum probability theory. We first introduced the double slit experiment in Section 4.2. This experiment allowed us to exhibit the differences in outcome predictions produced by Kolmogorovian and quantum probability. The main difference in the predictions arises in the "both slits open" case. The Kolmogorovian prediction

[1] We have added the sentence in bracket to contextualise the quote.

is summarised by Equation 4.1; while, the quantum prediction is summarised by Equation 4.9. These equations differ for the presence of the interference term I_{AB} in the quantum probability case. Experiments in quantum Physics have demonstrated that quantum probability theory predicts the outcomes of the double slit experiment with higher accuracy than Kolmogorovian probability.

Subsequently, we abstracted the physical experiment and created an analogy between the experiment and document ranking. The analogy consists in considering a user in place of a particle, documents in place of slits, and probability of relevance in place of probability of detection at the measurement screen (Section 4.3). With this analogy in place, the ranking decision (i.e. which document shall be ranked next?) is analogous to choosing the configuration of slits that maximises the probability of detection at the measurement screen (Section 4.4). The ranking criteria generated from the analogy are characterised by the probability theory that is used to model the double slit experiment. If Kolmogorovian probability is used, then ranking documents according to the analogy resembles PRP (Section 4.5). On the other hand, if quantum probability is used, then a new ranking principle is generated: the quantum probability ranking principle (Section 4.6).

The last part of this chapter discussed the optimality of qPRP for document ranking (Section 4.7) and the caveats that characterise the analogy between the experiment and document ranking (Section 4.8). Specifically, in the last section, we examined (i) the role within the analogy of location x on the measurement screen at which a particle is detected; (ii) the factors that may influence the magnitude of the predicted probabilities, including width of slits and distance between these; (iii) the situations where quantum interference may be absent; (iv) some of the motivations for using quantum theory in IR.

Chapter 5

Interference in qPRP

5.1 Introduction

In the previous chapter we have examined the double slit experiment and we observed the presence of quantum interference when measuring distributions of particles on a screen once scattered from two slits. We showed that the Kolmogorovian rule of additivity of probabilities of disjoint events is not valid in such configuration. In practice, interference is the deviation from this additivity. Similarly, Herbut [1992] suggests that the physical interpretation of quantum interference is "an observable deviation of quantum filtering from quantum occurrence of events" (where in our context quantum filtering is performed by the slits).

Thanks to the analogy developed in Chapter 4, we were able to interpret the document ranking process in the light of the double slit experiment. This resulted in a novel ranking principle: qPRP. Key to the principle is the presence of quantum interference in its ranking criteria. In fact, qPRP and PRP differ because of the presence of the interference term in qPRP. However, how could quantum interference be interpreted in IR, and in particular within qPRP?

In this chapter we attempt to answer this fundamental question. To this aim, we examine from a mathematical perspective why quantum interference arises in first place. In fact, we have observed that mathematically interference appears because of the additivity rule of probability amplitudes and because amplitudes are represented by complex numbers.

The presence of complex numbers is undoubtedly an intriguing characteristic of quantum probability theory, and of quantum theory in general. However, it is unclear what such numbers could or would actually represent or mean in IR. We then examine the role of complex numbers within quantum theory, and their potentials for IR, in Section 5.2. In Section 5.3, we then turn to examine possible interpretations of quantum interference in IR and in qPRP. We also investigate how interference influences document ranking and what governs interference. In our discussion we shall observe that, at the current state of the art, interference cannot be directly derived from the usual statistics used in IR, such as term frequency relationships. This is because, how we shall elaborate in Section 5.2, it is yet unclear how complex numbers may be used in IR. Thus estimations of interference have to be devised to empirically instantiate qPRP. Possible estimations are proposed in Section 5.4 and are empirically tested in Section 5.5.

5.2 Complex Numbers in IR

While traditional models of IR, such as the vector space models, are based on the field of real numbers, quantum models use complex vector spaces (i. e. Hilbert spaces). Complex numbers are one of the key concepts of the mathematical framework of quantum theory. As we have seen in Chapter 4, they allow to describe and model phenomena such as interference.

How to harness the use of complex numbers in quantum-inspired IR models has been largely ignored. This is also the case for most quantum-inspired models proposed in disciplines outside Physics, i. e. the so called “Quantum Interaction” research area Bruza et al. [2009b].

There are two main exceptions. van Rijsbergen [2004] only sketched out the use of complex numbers, proposing to store the term frequency and the inverse document frequency respectively in the magnitude r and the phase θ of a complex number $re^{i\theta}$. However, no further theoretical insight supporting this proposal has been given, and no empirical evaluation has been performed. In the context of semantic space models, De Vine and Bruza [2010] proposed a novel approach for the construction of spaces based on circular holographic representations, where the construction of complex valued vectors plays a fundamental role in preserving the order information in n-grams. However, they do not provide an interpretation of how complex numbers are used.

There have been few attempts to use complex numbers in IR that are not related to the framework of quantum theory. Notably, Park et al. [2005] proposed a model that employs discrete wavelengths transform for document retrieval. There, magnitudes and phases of complex numbers are obtained from the spectra of signals associated with query terms within a document. These are then used to score documents. Here however, we restrict our attention to the use of complex numbers within quantum inspired models for IR.

Next, we first define what complex numbers are useful for in the context of the mathematical framework of quantum theory and in particular of quantum probability theory. We then demonstrate theoretically and empirically that van Rijsbergen's proposal of evoking complex valued representations of informations objects does not hold, and discuss how complex numbers could be made explicit in qPRP.

5.2.1 Use of Complex Numbers in Quantum Theory

As stated, complex numbers are pervasive throughout the mathematical framework of quantum theory, due to the wave nature of matter. As such, they provide more freedom in terms of (quantum) probability distributions, and it is this degree of freedom that we describe in this section. We make bold simplifications for the sake of clarity.

First, we review the concept of quantum probability, and we represent probabilities in terms of vectors in a complex-valued Hilbert space. In its simplest form, a quantum probability is characterised by a quantum probability distribution and an event, which are respectively defined by the *unit* vectors $\mathbf{d}$ and $\mathbf{e}$. The probability $p^{\mathcal{Q}}(\mathbf{e}|\mathbf{d})$ of event $\mathbf{e}$ given distribution $\mathbf{d}$ is then $|\mathbf{d} \cdot \mathbf{e}|^2$, which corresponds, from a geometrical perspective, to the squared cosine between the two vectors. This relationship shows that vector based IR can be interpreted within quantum probability theory [van Rijsbergen, 2004].

Let us analyse further the concept of quantum probability in a concrete example, by considering two vectors on a two-dimensional space. Specifically, we represent the event and the distribution as respectively:

$$\mathbf{e} = \sqrt{\frac{1}{2}}(1,1)^\top \tag{5.1}$$

$$\mathbf{d} = \sqrt{\frac{1}{|1+e^{i\theta}|}}(1,e^{i\theta})^\top \tag{5.2}$$

where $\mathbf{d}$ depends on a parameter, i. e. the angle or phase $\theta \in [0, 2\pi[$, $|\cdot|$ denotes the usual norm of a complex number, and $\sqrt{1/2}$ and $\sqrt{1/|1+e^{i\theta}|}$ are the normalising factors that yield unit vectors.

Unless $\theta \in \{0, \pi\}$, $\mathbf{d}$ is expressed by complex numbers with no null imaginary parts. By varying θ between 0 and π, the probability $p^\Omega(\mathbf{e}|\mathbf{d})$ varies between 1 and 0. Further, an important fact is that multiplying $\mathbf{e}$ and $\mathbf{d}$ by $e^{i\psi}$ would not change the (quantum) probability value, for all $\psi \in \mathbb{R}$. In fact, it is the *phase difference* between the components in the vector that is important, and not their values per se. In our example, the phase difference between the two components of the vector in $\mathbf{d}$ is θ.

5.2.2 How May Complex Numbers Be Used in IR?

A simple IR example can clarify the situation. If we assume that $\mathbf{e}_a = (1,0)^\top$ and $\mathbf{e}_b = (0,1)^\top$ are documents containing term a and b, respectively, then $\mathbf{e} = \sqrt{1/2}(1,1)^\top$ means that the document contains both terms in equal quantities. By varying θ in $\mathbf{d}$, we can express that a document is relevant if it contains either a or b, but not both (case $\theta = \pi$), or is relevant if it contains a, b or both (case $\theta = 0$). Intermediate values of θ enable smooth transitions from one possibility to the other.

The idea of using the phase difference between words could also be used in the Quantum Information Retrieval framework proposed by Piwowarski et al. [2010] and based on quantum probability theory. In that framework, the term vector space is used to represent both documents and information needs. Terms can *interfere* between each other in the measurement of relevance.

Interestingly, one could interpret the negative numbers (i. e. $\theta = \pi$) obtained when performing Latent Semantic Analysis [Landauer, 2006] through the prism of the quantum formalism: in this case, a basis vector would contain two categories of terms that are mutually exclusive, i. e. that generally do not co-occur.

5.2.3 Analysis of the Potentials of Complex Numbers in IR

5.2.3.1 Encoding idf in the Phase

van Rijsbergen [2004, page 25] suggested to use complex numbers as a sort of storage mechanism for information (which then has to be transformed at matching time) where instead of associating to each component of the vector space a $\mathrm{tf} \times \mathrm{idf}$ value, it associates $\mathrm{tf} \times e^{i \cdot \mathrm{idf}}$. As this is the only example of complex number usage in van Rijsbergen's book, let us go beyond its usage as a simple storage scheme, which is not particularly useful in itself, and interpret it directly as a new complex weighting scheme for documents and queries.

Note that to operationalise this proposal, we normalised the idf so it ranges between 0 and 2π, since these are the extremal values that a phase can take. From a theoretical point of view, according to Section 5.2.1, van Rijsbergen's proposal would mean that if a query contains a term a with a high idf and b with an average idf, then a document would have a high probability of being relevant if it contains either a or b, but not both! This counterintuitive behaviour does not really depend on the mapping between idf and the $[0, 2\pi]$ range.

We empirically tested this proposal. In particular we experimented with the standard vector space model ($\mathbb{R}$-VSM) and the "complex" VSM ($\mathbb{C}$-VSM) on the following TREC collections:

- the Associated Press dataset (AP8889) with TREC 1, 2, 3 topics;
- the Wall Street Journal dataset (WSJ8792) with TREC 1, 2, 3 topics;
- the Los Angeles Times dataset (LA8990) with TREC 6, 7, 8 topics;
- the Web Track 2GB and 10 GB datasets (WT2g and WT10g) with the TREC 8 (WT2g), 9, 10 (WT10g) topics.

Details and statistics of these collections can be found in Table 6.2. Both documents and queries were indexed with the Lemur/Indri toolkit[1], after applying Porter stemming and stop-word removal. The real-valued vector space model ($\mathbb{R}$-VSM) and the complex-valued vector space model ($\mathbb{C}$-VSM) were implemented

[1] http://www.lemurproject.org/

	AP8889	**WSJ8792**	**LA8990**	**WT2g**	**WT10g**
$\mathbb{R}$-VSM	.1870	.1789	.1378	.1276	.1038
$\mathbb{C}$-VSM	.1313†	.0967†	.1146†	.0781†	.0232†

Table 5.1: Values of MAP for two matching models based respectively on a real-valued and a complex-valued vector space model ($\mathbb{R}$-VSM and $\mathbb{C}$-VSM). Statistical significance is calculated using a two-tailed paired t-test with $p \ll 0.01$ and is indicated by †.

as C++ extension of the Lemur/Indri APIs. In the real-valued vector space model, vectors component are computed following the tf × idf weighting scheme. Document rankings were evaluated using MAP.

The results of the empirical investigation are reported in Table 5.2.3.1. Statistical significance is measured using a two-tailed paired t-test with $p \ll 0.01$ and is indicated by † in the table. The results show clearly that the encoding of idf in the phase does not perform well, even when compared to the low baseline of the tf × idf weighting scheme.

5.2.3.2 Complex Numbers in qPRP

qPRP implicitly relies on interferences, and hence on complex numbers. It is interesting then to make explicit the representation of documents and to uncover the meaning of complex numbers in this case.

Intuitively, in the context of the diversity retrieval task a phase difference between a pair of documents corresponds to the fact that the documents are relevant to the same topic, and consequently their relevance probabilities should not add up. A possible re-interpretation of the example of Section 5.2.1 is as follows. Assume that d_a (respectively d_b) corresponds to the fact that document d_a (respectively d_b) is relevant. We can see that with a phase difference of $\pi/2$, a ranking containing the documents d_a and d_b would have the same probability of being relevant to the user than a ranking containing only d_a or d_b.

How to explicitly encode the relevance of documents and to define the probability distribution is still not clear at this stage. However, the previous example shows that it might be possible to build up the document representation by ensuring that documents do exhibit the same interference as the one that are empirically shown to work well (e.g., through an empirical estimation of the interference term).

5.3 Interpretation of Interference in qPRP

Quantum interference is central in the formalisation of qPRP. We have discussed why interference is present in quantum theory, and to what it corresponds in its mathematical framework. However, once interference is expressed in terms of IR, these questions may arise:

1. What does quantum interference mean in qPRP and in IR? What is our interpretation of quantum interference?
2. How does the quantum interference term influence document ranking?
3. What governs the quantum interference term?
4. How does quantum interference behave in qPRP by varying θ?
5. How is θ computed in IR and in qPRP?

Next, we attempt to provide answers to these questions.

5.3.1 What does quantum interference mean in qPRP and in IR?

We suggest that, in the context of document ranking, interference occurs between documents in a ranking (or their representations) at relevance level. For example, Chen and Karger [2006] and Zhai et al. [2003] showed that users are more likely to be satisfied by documents addressing different aspects or intents of their information need than by documents with the same content. It might then be sensible to model documents expressing diverse information as having a higher degree of interference than documents that are similar. Similarly, 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 previous ranks might trigger a revision of the user's beliefs about a topic.

Interference might then model dependencies in documents' relevance judgements. And with this respect, in qPRP the relevance of documents ranked until position $n-1$ interferes with the relevance of the document ranked at position n. This characteristic of qPRP suggests that this ranking principle may be suited to address ranking problems such that defined in the diversity retrieval task.

Note that similar interpretations of quantum interference within IR have been implicitly suggested by Wang et al. [2010], Di Buccio et al. [2011] and Melucci [2010]. In the first work, in fact, Wang et al. suggested that topics may interfere, thus affecting the relevance assessments of documents. This is similar to our interpretation, because we suggest that the relevance assessments of documents exhibit interference because of the intents or topics they address.

Di Buccio et al. [2011] instead suggested that quantum interference transpires from the superposition of relevance and non-relevance assessments that occur in the assessment process and that thus reflects the user's uncertainty about an assessment. A similar interpretation of quantum interference in IR is provided by Melucci [2010]. Common to our interpretation is the intuition that interference occurs at relevance level, affecting the assessments of documents. However, while we explicitly link interference as being generated by considering and assessing multiple documents, Di Buccio et al. [2011] suggested that interference is produced by the interactions the user performs during the assessment process. In that case then, assessing documents relevance is only one of the actions that are performed by the user.

5.3.2 How does the quantum interference term influence document ranking?

To answer this question we consider a small scale example and we contrast how PRP and qPRP rank documents. This allows us to gain an understanding of how the quantum interference term influence document ranking, because PRP and qPRP differ for the presence of the interference term.

Consider the situation where document d_A is ranked at the first rank position, and two documents d_B and d_C are left to be ranked. Assume that $P(R|q, d_B)$ is greater than $P(R|q, d_C)$ Which document should be ranked right after d_A?

Because $P(R|q, d_B) > P(R|q, d_C)$, PRP ranks d_B after d_A and before d_C, generating the ranking $\langle d_A, d_B, d_C\rangle$. According to Equation 4.20 the same ranking is produced by qPRP if and only if the difference between the probabilities of relevance of each document is greater than the difference between the associated interference terms:

$$\text{qPRP produces } \langle d_A, d_B, d_C\rangle \Leftrightarrow P(R|q, d_B) - P(R|q, d_C) > I_{d_A d_C} - I_{d_A d_B}$$

However, the document ranking produced by qPRP diverges from that of PRP if the previous condition is not met. In fact, if $P(R|q, d_B) - P(R|q, d_C) < I_{d_A d_C} - I_{d_A d_B}$ the interference $I_{d_A d_C}$ is strong enough to fill the gap given by $P(R|q, d_B) + I_{d_A d_B} - P(R|q, d_C)$. In such situation, ranking using qPRP results in the oder $\langle d_A, d_C, d_B\rangle$. Within the context of the diversity task, this situation may be interpreted as follows. Document d_C carries diverse and novel information related to the query with respect to document d_A; while, document d_B's content is less novel or possibly not novel at all when contrasted to the content of d_A.

Finally, consider the situation where $P(R|q, d_B) = P(R|q, d_C)$ and d_A is ranked at first position. In such case, PRP ranks first either one of these documents. In the same situation, qPRP instead would rank first the document that exhibits higher interference with d_A. For example, d_B is ranked above d_C if and only if $I_{d_A d_B}$ is greater than $I_{d_A d_C}$. These observations are summarised in Table 5.2.

5.3.3 What governs the quantum interference term?

Recall that mathematically the interference term between d_A and d_B is given by:

$$\begin{aligned} I_{d_A d_B} &= \phi_{d_A}\overline{\phi_{d_B}} + \phi_{d_B}\overline{\phi_{d_A}} \\ &= 2 \cdot \sqrt{P(R|q, d_A)}\sqrt{P(R|q, d_B)} \cdot \cos\theta_{d_A d_B} \end{aligned}$$

where $\theta_{d_A d_B}$ is the difference between the phases of the complex probability amplitudes that express the relevance of document d_A and d_B.

	$P(R\|q, d_B) > P(R\|q, d_C)$	$P(R\|q, d_B) = P(R\|q, d_C)$
PRP	d_B before d_C	either
qPRP	d_B before d_C iff $P(R\|q, d_B) - P(R\|q, d_C) > I_{d_A d_C} - I_{d_A d_B}$	d_B before d_C iff $I_{d_A d_B} > I_{d_A d_C}$

Table 5.2: When is d_B ranked above d_C? A comparison between ranking with PRP and with qPRP.

When $\cos\theta_{d_A d_B} > 0$, then[1] $I_{d_A d_B} \geq 0$ and is called constructive interference. At the contrary, destructive interference is obtained when $\cos\theta_{d_A d_B} < 0$. The behaviour of the quantum interference term is then governed by the difference in phase $\theta_{d_A d_B}$ between the complex probability amplitudes ϕ_{d_A} and ϕ_{d_B}.

5.3.4 How does quantum interference behave in qPRP by varying θ?

In qPRP, the phase difference actively affects the document ranking. For example, when $P(R|q, d_B) = P(R|q, d_C)$, document d_B is ranked above document d_C when $\cos\theta_{d_A d_B} > \cos\theta_{d_A d_C}$. In general, when $P(R|q, d_B) \geq P(R|q, d_C)$ the interference term subverts the ordering suggested by the probability of relevance only if (i. e. PRP's ranking)

$$\frac{P(R|q, d_B) - P(R|q, d_C)}{2\sqrt{P(R|q, d_A)}} < \sqrt{P(R|q, d_C)}\cos\theta_{d_A d_C} - \sqrt{P(R|q, d_B)}\cos\theta_{d_A d_B}$$

5.3.5 How is θ computed in IR and in qPRP?

The value of the phase difference $\theta_{d_A d_B}$ depends ultimately upon $\phi(R|q, d_A)$ and $\phi(R|q, d_B)$ and how they have been computed. While $P(R|q, d_A)$ and $P(R|q, d_B)$ are commonly estimated from statistical features of the document collection (such as term frequency, document frequency, etc.), the estimation of the complex probability amplitudes and therefore of the phases θ_{d_A}, θ_{d_B}, etc., is still an open question, as discussed in Section 5.2.

[1] The case $I_{d_A d_B} = 0$ occurs only when $\cos\theta_{d_A d_B} = 0$, or when one or both probabilities $P(R|q, d_A)$, $P(R|q, d_B)$ is zero.

To remedy to this problem and to empirically instantiate qPRP, we suggest that $\theta_{d_A d_B}$ may be estimated for every pair of documents d_A, d_B so as to create a relationship between documents in pairs. Effective estimations are likely to be tailored to the specific tasks qPRP is applied to.

If estimations of $\theta_{d_A d_B}$ have to be sought such that they model a sort of dependencies between documents at relevance level (e.g. in the diversity retrieval task), then similarity functions represent natural candidates. Intuitively, this is because similarity relations between documents can be encoded in phase differences. In particular, suitable estimations of $\theta_{d_A d_B}$ may be in the general form:

$$\theta_{d_A d_B} \approx \arccos(\text{sim}\,(d_A, d_B)) + \pi \tag{5.3}$$

where sim is a function of similarity between documents. Alternative strategies might relate θ to information gain or cross entropy between documents.

5.4 Estimating Interference in qPRP

As discussed in the previous section, without a method to compute complex probability amplitudes, estimations of phase differences are needed to instantiate qPRP in practical settings. We suggested that to estimate phase differences we may resort to compute similarities between documents, which then are encoded in phase differences by means of the inverse of the cosine function (i. e. the arccosine).

Here though, we take a slightly different approach. We focus in fact on the estimation of the quantum interference term, instead of the phase differences. Consider the following estimation:

$$\begin{aligned} I_{d_A d_B} &= \phi_{d_A}\overline{\phi_{d_B}} + \phi_{d_B}\overline{\phi_{d_A}} \\ &= 2 \cdot \sqrt{P(R|q, d_A)}\sqrt{P(R|q, d_B)} \cdot \cos\theta_{d_A d_B} \\ &\approx 2 \cdot \sqrt{P(R|q, d_A)}\sqrt{P(R|q, d_B)} \cdot \beta f_{sim}(d_A d_B) \end{aligned} \tag{5.4}$$

where $f_{sim}(d_A d_B)$ is a function assessing the similarity between documents d_A and d_B, and β is a real-valued free parameter (i. e. $\beta \in \mathbb{R}$). At this stage, we do not set any restriction to the values returned by the similarity function, although ideally[1] we would expect $\beta f_{sim}(d_A d_B)$ to be bounded between -1 and 1, so as to mimic the minimum and maximum values of $\cos\theta_{d_A d_B}$.

Here, estimations are made through similarity functions between documents (similarly to the estimation of phase differences); however there is now no need to compute arccosines and cosines, thus simplifying the computations. In fact, in Equation 5.4 we directly approximated $\cos\theta_{d_A d_B}$ with $\beta f_{sim}(d_A d_B)$. The presence of the parameter β has a dual goal. On one hand, it allows to adjust the magnitude of the estimation obtained through the similarity function. This may also be used for example as a form of normalisation. On the other hand, it allows to control the sign of the interference, i. e. positive or negative, so as to tailor the interference's estimations to the task qPRP is instantiated in. For example, while in the diversity retrieval task the interference between a pair of similar documents may be estimated as negative, in the ad-hoc retrieval task the same pair of documents may be estimated as creating positive interference.

In the following we consider a number of similarity functions, which act on vector representations of documents. While other similarity functions and document representations may be feasible, we think that similarity functions based on vector representations of documents may be good candidate to study estimations of interference because:

1. vector representations conceptually resemble the geometrical nature of the quantum formalism;

2. similarity functions that act on vectorial document representations have been widely studied in IR, see for example the work of Lee [1999].

5.4.1 Constructing Document Representations

We consider representations of documents based on vectors. In particular, we associated to each document a vector, which is defined on the vector space made up by the terms in the collection of considered documents. In such settings, each term in the collection is considered as a dimension of the vector space. The term-vector of a document is then constructed considering statistical features of the

[1] If this condition is not satisfied, values can be forced to range in $[-1, 1]$ through normalisation.

occurrences of terms in a document and in the collection. Different strategies can be employed to compute the components of the term-vector for a document. For example, a binary schema may be employed so that each component of a document's term-vector is 1 if the correspondent term appears in the document, while it is zero if the term is not present in the document. Instead of using only presence and absence of terms in documents to compute the vector's components, occurrence frequencies of terms in documents may be as well used. To this aim, common weighting schemas developed in IR may be used, such as term frequency (TF), inverse term frequency (IDF), BM25, etc. Determine which approach is more suited to estimate interference for a specific IR task is just matter of empirical investigation.

Once vector representations of documents have been computed, we can turn to consider how they may be used when computing similarities between documents. We consider two approaches that are based on different hypothesis:

1. **pairwise:** the user judges the interest of the current document by comparing it to each of the previous ranked documents. In this case, the current candidate and the documents already ranked are compared using f_{sim} in a pairwise fashion;

2. **surrogates:** the user judges the interest of the current document by comparing it to the knowledge acquired from documents ranked in the previous positions. The current candidate is then compared against a surrogate of the documents already ranked, which is obtained interpolating their vector representations.

When considering the surrogate hypothesis, we consider linear interpolation of the documents' term vectors to form a surrogate. Alternative approaches may however perform a weighted interpolations of these vectors, in order to simulate user's memory effects (e.g. documents retrieved at early ranks are weighted less than documents at ranks close to the current one) or estimated importance of documents (e.g. documents ranked at early ranks contribute more in generating the surrogate than lower ranked ones).

5.4.2 Candidate Similarity Functions

Several similarity functions may be employed to estimate interference in the context of document ranking. Many of these similarity functions have been already investigated in IR, for example for the purpose of improving probability estimation for unseen term co-occurrences [Lee, 1999]. Next, we give a brief overview of the similarity functions we investigate here. Note that in general the similarities obtained through any of these functions have to be normalised (if they are not already) so that f_{sim} ranges between the boundaries set by $\cos\theta$ (i. e. in the range $[-1,1]$).

- Pearson's correlation coefficient:

$$r_{d_A,d_B} = \frac{\sum_{k=1}^{n}(d_A(k)-\bar{d_A})(d_B(k)-\bar{d_B})}{\sqrt{\sum_{k=1}^{n}(d_A(k)-\bar{d_A})^2}\sqrt{\sum_{k=1}^{n}(d_B(k)-\bar{d_B})^2}} \tag{5.5}$$

  where $\bar{d_A}$ is the mean of the components of the term-vector representation of document d_A, and $d_A(k)$ is the k-th component of the n-dimensional vector representation of document d_A (similarly for d_B). This notation is used also for the other functions employed here to compute similarities between documents.

- Kullback-Leibler divergence (KLD):

$$KLD(d_A,d_B) = \sum_{k=1}^{n} d_A(k)\cdot\log\left(\frac{d_A(k)}{d_B(k)}\right) \tag{5.6}$$

- Jensen-Shannon divergence (JSD)

$$\begin{aligned} JSD(d_A,d_B) &= \left(\sum_{k=1}^{n} d_A(k)\log\frac{d_A(k)}{\frac{1}{2}(d_A(k)+d_B(k))}\right. \\ &\quad \left.+\sum_{k=1}^{n} d_A(k)\log\frac{d_B(k)}{\frac{1}{2}(d_A(k)+d_B(k))}\right)^{\frac{1}{2}} \end{aligned} \tag{5.7}$$

- Skew divergence (SkD) [Lee, 1999]

$$SkD_\alpha(d_A,d_B) = \sum_{k=1}^{n} d_B(k)\left(\log d_B(k) - \log\left(\alpha d_A(k)+(1-\alpha)d_B(k)\right)\right) \tag{5.8}$$

with $\alpha \in [0, 1]$ representing the degree of confidence in the distribution of terms that is empirically observed from documents. Note that when $\alpha = 1$, the skew divergence becomes equal to the Kullback-Leibler divergence.

- L_1 norm:

$$L_1(d_A, d_B) = \sum_{k=1}^{n} |d_A(k) - d_B(k)| \tag{5.9}$$

- L_2 norm (or Euclidean):

$$L_2(d_A, d_B) = \sqrt{\sum_{k=1}^{n}(d_A(k))^2 - 2\sum_{k=1}^{n} d_A(k)d_B(k) + \sum_{k=1}^{n}(d_B(k))^2} \tag{5.10}$$

- Cosine similarity

$$\cos(d_A, d_B) = \frac{\sum_{k=1}^{n} d_A(k)d_B(k)}{\sqrt{\sum_{k=1}^{n} d_A(k)^2 \sum_{k=1}^{n} d_B(k)^2}} \tag{5.11}$$

- Jaccard similarity coefficient

$$Jac(d_A, d_B) = \frac{|d_A \bigcup d_B|}{|d_A \bigcap d_B|} \tag{5.12}$$

where $|d_A \bigcup d_B|$ is the total number of terms that are in d_A and d_B (without considering how many occurrences of each term are present in the document), while $|d_A \bigcap d_B|$ is the number of terms that d_A and d_B have in common. Note the abuse of notation in Equation 5.12, where d_A indicates the set of terms contained in the correspondent document (similarly for d_B), and $|.|$ indicates the size of a set (i. e. its cardinality).

5.5 Empirical Assessment of Interference Estimations for qPRP

To study in the context of qPRP the effectiveness of the estimations of the quantum interference term proposed in the previous sections, we conduct an empirical investigation on a specific IR task. We in fact consider the task of diversity retrieval, as described in Section 2.3.3. In this evaluation context, we consider instantiations of the interference term in the form:

$$I_{d_A d_B} \approx -2 \cdot \sqrt{P(R|q,d_A)}\sqrt{P(R|q,d_B)} \cdot f_{sim}(d_A d_B) \tag{5.13}$$

that is, instantiations obtained by setting $\beta = -1$. Intuitively, this setting is motivated by the fact that in the diversity retrieval task documents that address the same query-intent should not be retrieved closed-by, as top-ranked documents should each address different query-intents. When employing the instantiation of Equation 5.13, documents that are similar to those already ranked based on f_{sim} would be scored with a negative interference, while diverse documents will be characterised as achieving positive (or "less negative") interferences with the documents already ranked.

In our empirical investigation we employed the TREC 6, 7, 8 interactive collection with topics and subtopics judgements described in Zhai et al. [2003], and the ClueWeb collection (part B only), along with the topics defined for the TREC 2009 Web Diversity track. The collections were indexed using the Lemur/Indri toolkit, where Porter stemming was applied and standard stop words removed.

As baseline we scored documents using Okapi BM25 and ranked them according to PRP, where probabilities were estimated using the BM25 scores[1].

To investigate qPRP in context, we also employed an alternative comparable[2] approach for document diversification: Portfolio Theory for IR (PT, see Section 3.9). This strategy is based on the ranking criteria of Equation 3.22, and had been instantiated considering the variance σ^2 of a document's probability of relevance and the risk propensity of a user as parameters; while, Pearson's correlation between documents' term-vector representations were employed as a measure of correlation between documents. Parameters were tuned conducting a grid search of the parameter space b by σ^2 to select the most effective run of PT in terms of α-ndcg@10 on each employed collection. The weights w to be assigned to each rank position were computed as the inverse of the rank position plus one, i. e. $w_r = \frac{1}{\log(r+1)}$, with r being the rank positions $[1, n]$. More details on common instantiations of PT within this thesis shall be given in Section 6.4.2.1.

[1] For more details about the settings of our experiments, refer to Section 6.4.2.1 where similar settings are throughly discussed.

[2] i.e. that uses the same evidence used by qPRP to perform the ranking; this is in contrast with the techniques discussed in Section2.3.4 that use external evidence, such as query-logs, ontologies, search engine's query suggestions, etc., to explicitly diversity search results.

Note that in both PT and qPRP, the estimations of documents' probabilities of relevance were derived from the BM25 scores of documents, as for PRP. Similarly, Okapi BM25 had been used to compute the component of the documents' term-vectors in both PT and qPRP. For qPRP, we tested both pairwise and surrogates document representations.

All ranking approaches were implemented in C++ using Lemur/Indri's APIs. The resulting document rankings were evaluated using α-ndcg@10, NRBP, IA-P@10 and s-recall@10.

We report the results of our empirical investigation in Tables 5.3 and 5.4. Improvements provided by qPRP over PRP are marked with $*$ if they have been found to be statistical significant when using a paired t-test with $p < 0.01$. Similarly, statistical significant improvements provided by qPRP over PT are marked with †. The best retrieval performance are obtained when using Kullback-Leibler divergence to estimate the cosine of the phase difference of the quantum interference term. However, this result is obtained when using different document representations comparison in different collections, i. e. surrogates in TREC 6, 7, 8 and pairwise in ClueWeb. These differences might be due to:

- the limited number of topics available for the TREC 6,7, 8 collection[1]; and
- the different kind of documents contained in TREC 6, 7, 8, i. e. newswire articles, and in ClueWeb, i. e. Web pages.

Furthermore, while no statistical significance can be calculated on the improvements provided by qPRP in TREC 6, 7, 8 due to the scarce number of topics[2], improvements over PRP and PT obtained on ClueWeb are statistically significant. Note that the instantiation of qPRP based on the estimations obtained by Pearson's correlation consistently provides excellent retrieval performance regardless of the comparison method. While this is not always better than the tuned PT, it is not significantly worse, being in fact significantly better on several diversity measures. However, qPRP has a distinct advantage over PT, as no parameter tuning is required once the function that approximates interference has been set.

[1] Recall that only 20 topics are available for this collection.

[2] See [van Rijsbergen, 1979, pages 178–180] for details.

				Pairwise							
	Measure	**PRP**	**PT**	**Pear.**	**L1**	**L2**	**Cos.**	**Jac.**	**KLD**	**SkD**	**JSD**
TREC 678	**α-ndcg@10**	.416	.424	.418	.354	**.431**	.427	.419	.413	.426	.364
	NRBP	.123	.126	.127	.104	**.128**	.127	.124	.117*	.104	.106
	IA-P@10	.058	.062	.063	.042	.063	**.064**	.061	.062	.041	.043
	s-r@10	.379	.384	.387	.281	.385	.388	.381	**.389**	.267	.286
ClueWeb	**α-ndcg@10**	.093	.105	.094	.076	.099	.099	.097	**.115*†**	.075	.075
	NRBP	.032	.029	.035†*	.029	.029	.034†	.035†	**.043†**	.027	.029
	IA-P@10	.033	.041*	.035	.031	.046	.040	.038	**.047*†**	.032	.031
	s-r@10	.151	.178*	.180*	.121	.173*	.168*	.165*	**.190*†**	.131	.121

Table 5.3: Overview of the results obtained over two TREC test collections when using the method of *pairwise* comparison between document representations. Each similarity or correlation function indicates an instantiation of qPRP where the corresponding function is employed to estimate quantum interference. Statistical significance over PRP is indicated by *, while † indicates statistical significant improvements over PT. The best results for each evaluation measure are reported in bold.

				Surrogates							
	Measure	**PRP**	**PT**	**Pear.**	**L1**	**L2**	**Cos.**	**Jac.**	**KLD**	**SkD**	**JSD**
TREC 678	**α-ndcg@10**	.416	.424	.415	.301	.426	.424	.412	**.433**	.286	.328
	NRBP	.123	.126	.127	.091	.128	.126	.120	**.135**	.099	.096
	IA-P@10	.058	.062	.060	.018	.059	.060	.053	**.067**	.022	.029
	s-r@10	.379	.384	.379	.147	.375	.380	.360	**.402**	.183	.223
ClueWeb	**α-ndcg@10**	.093	.105	**.106***	.056	.100	.094	**.106***	.092	.063	.072
	NRBP	.032	.029	**.039***†	.025	.029	.034*†	.037*	.032	.025	.027
	IA-P@10	.033	**.041***	.038	.018	.043	.037	.044	.024	.034	.027
	s-r@10	.151	.178*	**.185***	.086	.175	.160	.184*	.113	.095	.123

Table 5.4: Overview of the results obtained over two TREC test collections when using the method of *surrogates* comparison between document representations. Each similarity or correlation function indicates an instantiation of qPRP where the corresponding function is employed to estimate quantum interference. Statistical significance over PRP is indicated by *, while † indicates statistical significant improvements over PT. The best results for each evaluation measure are reported in bold.

5.6 Summary

In this chapter we studied several aspects of quantum interference in quantum theory, in IR, and ultimately in qPRP.

First, we observed that quantum interference mathematically arises because of the associated concept of probability amplitudes, which act on the field of complex numbers. The role of complex numbers in quantum theory has been outlined in Section 5.2.1. The possible roles and uses of complex numbers in IR have then been discussed in Sections 5.2.2 and 5.2.3.

Subsequently, we focused on the interference term in qPRP. In Section 5.3 we discussed an interpretation of quantum interference in qPRP and studied how interference influences the ranking obtained with the novel ranking principle. Specifically, it was suggested that quantum interference models document dependencies at relevance level, i.e. interdependent document relevance. Then possible estimations of quantum interference within qPRP were suggested (Section 5.4). Empirical results on a series of retrieval experiments within the task of diversity retrieval have shown that good retrieval performances can be consistently obtained when employing Pearson's correlation to estimate interference in qPRP, while, the estimation based on Kullback-Leibler divergence provided the best retrieval performances overall (Section 5.5).

Chapter 6

A Comparison of Ranking Principles and Strategies

6.1 Introduction

In Chapter 3 we have examined PRP and a number of alternative ranking approaches. Moreover, in Chapter 4 we have proposed a novel ranking principle based on quantum probabilities. These approaches to document ranking can be divided into two categories:

strategies that are empirically driven and devised to cater for the limitations of PRP in tasks as diversity retrieval. Approaches that belong to this category are Maximal Marginal Relevance (MMR – Section 3.8) and Portfolio Theory (PT – Section 3.9);

principles that are theoretically driven and implicitly cater for the limitations of PRP in tasks as diversity retrieval. This category includes approaches as the interactive probability ranking principle (iPRP – Section 3.10) and the quantum probability ranking principle (qPRP – Section 4.6).

Regardless of the approach, strategy or principle, the recently proposed alternatives to the PRP mathematically deviate through the inclusion of a function that captures to some extents dependencies between documents. This function expresses the relationship between documents: depending upon how the function is set, the ranking approach promotes either document diversity or similarity. As we shall see, alternatives differ in the way dependencies are incorporated, and the extent of parameterisation of the ranking formula. Specifically, PT and qPRP are characterised by an additive ranking function, MMR by an interpolated and iPRP by a multiplicative, where PT and MMR are by definition parameterised. On the contrary, in their original formulations iPRP and qPRP do not have parameters. However, parametric instantiations may be formulated as well for qPRP and iPRP.

PRP has been formally shown to be optimal in the ad-hoc task, where all PRP's assumptions are upheld (see Section 3.5.1 and [Robertson, 1977]). Since the proposal of PRP, however, new tasks have been thought and investigated, that better describe typical or particular search scenarios[1]. For some of these tasks, PRP's assumptions are not upheld: the diversity retrieval task (Section 2.3.3) is an example of such situations.

It is therefore of interest to investigate how PRP and the alternative approaches behave when ranking documents under different circumstances. To gain insights into the document ranking process and the use of the considered ranking approaches, we compare and contrast them in the following aspects:

1. **analytically**: by examining formal relations between the ranking formulas of the approaches and by discussing under which conditions different approaches result in the same document ordering being created;

2. **empirically**: by investigating the retrieval performance of ranking approaches in two different tasks, i. e. ad-hoc retrieval and diversity retrieval, and on a number of TREC test collections;

3. **behaviourally**: by studying the kinematics of relevant documents observed on TREC collections, i. e. how documents are re-arranged by ranking approaches alternative to PRP.

Before performing this thorough investigation of ranking approaches, however, we need to set a common ground for studying them. To this aim, next we examine how to instantiate the ranking approaches in practical settings, such those considered in the retrieval tasks that form our empirical investigation.

[1] For a bird-eye view of other IR search tasks, the reader is referred to the work of Voorhees [2005] and the reports of the NTCIR and CLEF initiatives, e.g. [Joho et al., 2010] and [Agosti et al., 2010].

6.2 Instantiations of Ranking Approaches

Here we empirically instantiate the ranking approaches considered in this thesis. To allow and facilitate comparisons, we set a common ground which is then used to instantiate the ranking approaches. In particular, in all approaches we consistently employ the same estimations of the probability of relevance $P(R|q,d)$ and the same function to estimate relationships between documents[1]. To this aim, we use the Pearson's correlation between documents' term-vectors, similarly to the settings of Section 5.4.1 for the pairwise comparison of document representations. This choice is motivated by two facts. First, Pearson's correlation between document term-vectors is explicitly employed in PT, while approaches like MMR, iPRP and qPRP do not explicitly set which function shall be used to estimate document relations. Second, the empirical study of Section 5.5 performed on qPRP showed that this function provides robust retrieval performance (for qPRP), despite not being the function that provides the *best* performance.

6.2.1 Probability Ranking Principle

PRP ranks a document at rank i by following the criteria of Equation 3.1:

$$\textbf{PRP:}\ d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} P(R|q,d)$$

where $P(R|q,d)$ is the probability that document d is relevant to query q, estimated using common IR heuristics. For example, in the empirical experiments reported in this chapter, we shall consider probabilities of relevance estimated using the BM25 weighting schema (Section 2.2.3) or the Language Modelling approach (Section 2.2.4). The same probabilities of relevance estimated for PRP shall also be used for the alternative approaches. In fact, alternative approaches may be though as providing a re-ranking of the initial ordering imposed by PRP.

6.2.2 Maximal Marginal Relevance

According to MMR, a document at rank i is selected using the objective function of Equation 3.16:

[1] With the exception of PRP, where this function is not contemplated.

$$d_i = \operatorname*{arg\,max}_{d \in RE \setminus \mathcal{RL}} \left(\lambda Sim(d,q) - (1-\lambda) \max_{d' \in \mathcal{RL}} Sim(d,d') \right)$$

where $Sim(d,q)$ is a similarity function between document and query, while $Sim(d,d')$ is a function that determines the similarity between documents d and d'. If two candidate documents have the same similarity score with respect to a query, MMR will rank first the one that is least similar to any of the documents that have been ranked at previous positions. The hyper-parameter can be inferred by the user's model: $\lambda < 0.5$ characterises users with a preference for rankings where document dependencies are more important than relevance. Greater values of λ would capture the converse situation. To frame all ranking approaches on a common ground, we re-state MMR in terms of $P(R|q,d)$ as estimated for PRP and the Pearson's correlation between document d and d', i. e. $\rho(d,d')$ in place of $Sim(d,q)$ and $Sim(d,d')$, respectively:

$$\textbf{MMR: } d_i = \operatorname*{arg\,max}_{d \in RE \setminus \mathcal{RL}} \left(\lambda P(R|q,d) - (1-\lambda) \max_{d' \in \mathcal{RL}} \rho(d,d') \right)$$

6.2.3 Portfolio Theory

For each rank position i, under PT approach documents are selected according to:

$$\textbf{PT: } d_i = \operatorname*{arg\,max}_{d \in RE \setminus \mathcal{RL}} \left(P(R|d,q) - b w_d \sigma_d^2 - 2b \sum_{d' \in \mathcal{RL}} w_{d'} \sigma_d \sigma_{d'} \rho(d,d') \right)$$

where b encodes the risk propensity of the user, σ_d^2 is the variance associated to $P(R|d,q)$, and w_d is a weight that expresses the importance of the rank position of d and d'. This weight is computed as $w_{d_i} = \frac{1}{\log(1+i)}$, where i is the rank position of document d_i.

When PT has been employed in practice in previous works, σ_d has been treated as a model parameter (see for example [Wang and Zhu, 2009; Zuccon et al., 2011b]), because only a single point-wise relevance estimation is used. Moreover, this parameter had been considered constant among documents, i. e. $\sigma_d = \sigma_{d'}$, $\forall d, d'$. We follow the same route, thus considering two parameters: $\sigma_d^2 \in \mathbb{R}^+$

(i. e. positive real numbers) representing the variance associated to the relevance estimations and $b \in \mathbb{R}$, which encodes the user model.

6.2.4 Interactive PRP

By adhering to iPRP's ranking function, document d_i is ranked at position i according to Equation 6.1, i. e.

$$d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} \left(P(R|d,q) \frac{\sum_{d' \in \mathcal{RL}} Sim(d,d')}{|\mathcal{RL}|} \right)$$

where dependencies between documents are represented by a function of similarity between documents, $Sim(d,d')$, and are incorporated within the final score of a document through *multiplication*. This provides an approach completely different from the other alternatives. As for MMR, we instantiate $Sim(d,d')$ by using the Pearson's correlation between documents, thus obtaining:

$$\textbf{iPRP: } d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} \left(-P(R|d,q) \frac{\sum_{d' \in \mathcal{RL}} \rho(d,d')}{|\mathcal{RL}|} \right)$$

Although iPRP may be instantiated differently, we repute this instantiation suitable for comparing ranking approaches within the common framework considered in this chapter. An empirical investigation of this instantiation of iPRP has been reported by Zuccon et al. [2011a].

6.2.5 Quantum PRP

Under qPRP a document d is ranked at position i according to Equation 4.20, i. e.

$$d_i = \arg\max_{d \in RE \setminus \mathcal{RL}} \left(P(R|d,q) + \sum_{d' \in \mathcal{RL}} I_{d,d'} \right)$$

According to Chapter 5, quantum interference in qPRP can be approximated using a similarity or correlation function between documents. As for other approaches, we employ also for qPRP the Pearson's correlation $\rho(d,d')$ to allow some degree of similarity between approaches. Then, by setting

$$I_{d,d'} = -2\sqrt{P(R|d,q)}\sqrt{P(R|d',q)}\rho(d,d')$$

, we obtain

$$\textbf{qPRP: } d_i = \underset{d \in RE\setminus\mathcal{RL}}{\arg\max}\left(P(R|d,q) - 2\sum_{d'\in\mathcal{RL}}\sqrt{P(R|d,q)}\sqrt{P(R|d',q)}\rho(d,d')\right)$$

6.2.6 Parametric Instantiations of iPRP and qPRP

While MMR and PT are by definition characterised by the settings of their parameters, the instantiations of iPRP and qPRP of Sections 6.2.4 and 6.2.5 are not parametric. However, parametric instantiations of these principles can be given, where parameters control the impact of correlation on the ranking process. The parameter is formally introduced within the approximations of benefit (for iPRP) and quantum interference (for qPRP).

When instantiating iPRP, the benefit of ranking a document d at rank i (i. e. $b_{d,i}$) has been approximated as $-\frac{\sum_{d'\in\mathcal{RL}}\rho(d,d')}{|\mathcal{RL}|}$. A possible parametric instantiation of iPRP is obtainable by setting

$$b_{d,i} = -\beta\frac{\sum_{d'\in\mathcal{RL}}\rho(d,d')}{|\mathcal{RL}|}$$

with β being a free parameter (and $\beta \in \mathbb{R}$). Therefore, the ranking formula of iPRP becomes:

$$\textbf{iPRP(parametric): } d_i = \underset{d \in RE\setminus\mathcal{RL}}{\arg\max}\left(-\beta P(R|d,q)\frac{\sum_{d'\in\mathcal{RL}}\rho(d,d')}{|\mathcal{RL}|}\right)$$

Similarly, when operationalising qPRP, interferences have been approximated as $I_{d,d'} = -2\sqrt{P(R|d,q)}\sqrt{P(R|d',q)}\rho(d,d')$. Alternative approximations have been investigated in Section 5.4: these considered similarity functions other than Pearson's correlation for estimating interferences and no parameter was introduced. We can however consider a parametric instantiation of qPRP as well, by introducing the parameter β in the approximation of the quantum interference term, thus obtaining:

qPRP(parametric):

$$d_i = \underset{d \in RE \setminus \mathcal{RL}}{\arg\max} \left(P(R|d,q) - 2\beta \sum_{d' \in \mathcal{RL}} \sqrt{P(R|d,q)}\sqrt{P(R|d',q)}\,\rho(d,d') \right)$$

6.3 Analytical Analysis of Ranking Approaches

Each approach handles document dependencies in a characteristically different way. The question is: *How do different approaches affect document ranking?*

To answer this question, we shall consider two aspects:

1. which document is ranked first?

2. which documents are ranked next?

For all approaches, the document ranked at first position (i. e. $i = 1$) is the same. This is the document which has the highest probability of relevance. Differences between alternative approaches and PRP manifest at ranks greater than one. At $i > 1$, each alternative approach will tend to revise the original ranking such that documents which are different to those ranked previously will be promoted. To obtained deeper intuition of this phenomena for each ranking alternative, we analytically compare each approach at the functional level to determine more precisely how the ranking of documents would be affected.

To this aim, we shall consider the following example scenario, where we have two documents, d and d', with the same probability of relevance, i. e. $P(R|q,d) = P(R|q,d')$, and d has been ranked first. We are interested to determine what is likely to happen to d' given PRP, MMR, PT, iPRP, and qPRP: i. e. is it likely to be demoted or promoted with respect to other documents? We consider three further cases, where documents d, d' are:

1. virtually identical[1] and thus positively correlated, i. e. $\rho_{d,d'} = 1$;

[1] We consider the document term vectors to compute correlations (and thus dependencies): term-position does not influence correlation, while term's (weighted) presence does. Two documents containing the same exact text, but shuffled in different orders, will appear identical to the correlation function.

2. with nothing in common, and thus not correlated at all, i. e. $\rho_{d,d'} = 0$;

3. sharing the same terms, but with complete different use and frequencies, and thus anti-correlated[1], i. e. $\rho_{d,d'} = -1$.

Probability Ranking Principle The behaviour of PRP does not depend on the correlation. PRP then always ranks documents d and d' consecutively, and actually both $(d, d', ...)$ and $(d', d, ...)$ are valid rankings.

Maximal Marginal Relevance When documents are correlated (case 1), MMR assigns to d' the score $\lambda P(R|q, d') - (1 - \lambda)$, which might be a negative value. If $\lambda = 1$ then MMR reduces to PRP, while if $\lambda = 0$ document d' gets a score of 1. For $0 < \lambda < 1$, the original score of $P(R|q, d')$ is remodulated by λ and then decreased of $(1-\lambda)$. In case 2, MMR rescales the document's probability by the hyper-parameter, assigning to d' the score $\lambda P(R|q, d')$. The document score increases in the third case, i. e. when the correlation has negative value, adding to the (re-scaled) probability of the document a value proportional to $1 - \lambda$: if $\rho_{d,d'} = -1$, then the score of d' is $\lambda P(R|q, d') + 1 - \lambda$.

Portfolio Theory The score PT assigns to a document differs to the one provided by PRP of $-bw_d\sigma_d^2 - 2bw_{d'}\sigma_d\sigma_{d'}\rho_{d,d'}$. The sign of PT's variation in scores, i. e. increment or decrement, are then not only dependent upon the correlation's sign, but also upon the user's model parameter b. We focus our analysis on the situation where $b > 0$: under this circumstance PT promotes diversity in the document ranking. The initial document probability of relevance is revised of $-|b|w_d\sigma_d^2 - 2|b|w_{d'}\sigma_d\sigma_{d'}\rho_{d,d'}$. In case 1, i. e. $\rho_{d,d'} = 1$, the score of d' is decreased by $-|b|w_d\sigma_d^2 - 2|b|w_{d'}\sigma_d\sigma_{d'}$. If documents are not correlated (case 2), the initial score undergoes a limited decrement of $|b|w_d\sigma_d^2$. Finally, in case 3 (anti-correlated documents), the initial score of d' is modified by PTs's ranking formula of $-|b|w_d\sigma_d^2 + 2|b|w_{d'}\sigma_d\sigma_{d'} \approx |b|\sigma_d^2(2w_{d'} - w_d)$. The discount factor w_d is estimated through a monotonically decreasing function of the document's rank position, thus $2w_{d'} - w_d$ can be either positive or negative. If positive, d''s score gets incremented; vice versa, d' gets demoted in the document ranking. Finally, when $b = 0$ PT's ranking function reduces to the one of PRP.

[1] While in practice correlations of -1 are unlikely, there might be cases where correlations are negative because of the weighting schema used to compute document term vectors. However, for the purpose of our example, we imagine the two documents to be completely anti-correlated.

Table 6.1: Overview of the characteristics of the ranking principles and strategies.

Model	**Dependence**	**Parameters**	$\rho = 1$	$\rho = 0$	$\rho = -1$
PRP	-	-	$\circ$	$\circ$	$\circ$
MMR	Interpolated	λ: hyperparameter	$\downarrow$	$\sim$PRP	$\uparrow$
PT	Additive	b: user risk propensity σ: variance estimation relevance w: discount rank position	$\downarrow$ (if $b > 0$)	$\sim$PRP	$\uparrow$ (if $b > 0$)
iPRP	Multiplicative	-	$\downarrow$	0	$\uparrow$
qPRP	Additive	-	$\downarrow$	=PRP	$\uparrow$

Interactive PRP iPRP is characterised by a multiplicative ranking function. When d and d' are completely correlated (case 1), iPRP assigns to d' the score $-P(R|q,d')$, and thus the document is demoted: documents that are more relevant than others would suffer a stronger demotion. In the situation of zero-correlated documents (case 2), d' gets assigned a score of zero and is demoted in the ranking. In case 3, iPRPs assigns to d' the same score obtained with PRP, i. e. $P(R|q,d')$, and thus d' is ranked immediately after d (as for PRP).

Quantum PRP When documents correlate, as in case 1, the probability assigned to d' is revised and is modified to the value $-P(R|q,d')$: this is due to the interference term becoming $I_{d,d'} = -2\sqrt{P(R|q,d)}\sqrt{P(R|q,d')} = -2P(R|q,d')$. In this situation, as for other models, also according to qPRP d''s chances to get ranked at second position are decreased, possibly demoting it to lower positions. When d and d' are not correlated at all as in case 2, i. e. $\rho_{d,d'} = 0$, qPRP does not change PRP's estimate since the interference term is zero: there is no dependence between the actual candidate and the previous ranked document. In case 3, qPRP boost the original probability of d' to the quantity $3 \cdot P(R|q,d')$. In fact, the interference term results $I_{d,d'} = 2\sqrt{P(R|q,d)}\sqrt{P(R|q,d')} = 2P(R|q,d')$.

Summary The approaches revealed a common pattern. When promoting diversity, the initial probability estimation associated to d', i. e. $P(d')$, is revised by a quantity proportional to the correlation of d' with those documents that have been already ranked. The revision increments the initial probability estimation if documents are anti-correlated. Vice versa if documents are correlated, the document score is decreased. The case of no correlation (case 2) is handled differently by each ranking approach: for example iPRP assigns to the document a zero score, while qPRP returns the same probability estimation of PRP.

Finally, the amount of revision that the score of a document is subject to depends upon the parametrisation of the ranking function. Specifically:

- MMR weights the contribution of the correlation depending on λ; high values of λ (i. e. $\lambda \rightarrow 1$) return rankings similar to those of PRP;
- PT modulates the contribution of the correlation by the product of the parameters b and σ_d^2, and considering the importance of the rank position;

- iPRP reduces the influence of the correlation by a quantity inversely proportional to the number of documents retrieved at previous ranks;
- qPRP modulates the contribution of the correlation by the square root of the probabilities of the documents involved in the comparison.

The previous observations are summarised in Table 6.1, where the approaches are contrasted in terms of (i) the way dependency is accounted for, (ii) the parameters that characterise the ranking functions, and (ii) the behaviours of each alternative approach with respect to PRP when different values of ρ are considered.

6.3.1 Relationships between Approaches: does PT uphold qPRP?

In this section we unveil some analytical relationships between PRP PT and qPRP and between PT and qPRP. To do so, we first examine when PT and qPRP uphold PRP; then we turn our attention to examine the situation when PT upholds qPRP. The unveiled relationships are useful when examining similarities between the ranking approaches. Furthermore, we show that the relationship that is found between PT and qPRP may lead to an effective approach to estimate parameters settings for PT.

Upholding PRP Under PRP, the optimal ranking would be obtained by taking, at each rank position i, the document d that maximises $P(R|q,d)$. In relation to PT then, when the user parameter b is zero, or documents' variance is null, the additive component of Equation 3.22 is zero. In this case, PT upholds PRP. This guarantees the optimality of the ranking in tasks such as ad-hoc retrieval. But this is a trivial case. As soon as $|b|$ increases, the influence of the additive term will perturb the ranking, and PT will begin to violate PRP (the greater the $|b|$ the further PT departs from PRP)[1]. This is because documents will not be strictly ordered according to their decreasing probability of (independent) relevance as prescribed by PRP.

[1] Assuming that the other parameters of PT are non-zero.

Like PT, qPRP reduces to PRP when the interference between documents is null, i. e. documents are not interdependently related. And also like PT, qPRP is characterised by an additive ranking formula, which interpolates relevancy and document dependencies.

We have shown that PT violates PRP in non trivial circumstances. This is actually desirable in tasks such that of diversity retrieval, since PT aims to overcome PRP's assumption of independent document relevance.

Does PT uphold qPRP? To answer this, we consider a particular situation. We instantiate qPRP approximating the interference term with a function of the Pearson's correlation ρ between documents term vectors, i. e. $\cos\theta_{d,d'} = -\rho(d,d')$, as suggested in Section 6.2.5. Similarly, Pearson's correlation can be employed in PT to measure the correlation in Equation 3.22. For simplicity of exposition, we re-write the ranking formula of qPRP and PT obtained when using Pearson's correlation

$$\textbf{qPRP: } d_i = \underset{d \in RE\backslash\mathcal{RL}}{\arg\max}\left(P(R|d,q) - 2\sum_{d' \in \mathcal{RL}} \sqrt{P(R|d,q)}\sqrt{P(R|d',q)}\,\rho(d,d')\right)$$

$$\textbf{PT: } d_i = \underset{d \in RE\backslash\mathcal{RL}}{\arg\max}\left(P(R|d,q) - bw_d\sigma_d^2 - 2b\sum_{d' \in \mathcal{RL}} w_{d'}\sigma_d\sigma_{d'}\rho(d,d')\right) \tag{6.1}$$

We assume σ_d to be a constant for each document in the collection[1]; in this case Equation 6.1 can be re-stated as

$$d_i = \arg\max\left(P(R|q,d) - \sum_{d' \in RA} 2b\sigma_d^2 w_{d'}\rho(d,d')\right) \tag{6.2}$$

where w_d is dropped for rank equivalence reasons, i. e. whatever the d under consideration, w_d is constant and so is $b\sigma_d^2 w_d$. When instantiating PT in these particular circumstances, b and σ_d^2 can be treated as parameters to be tuned.

[1] This assumption is realistic in the case probabilities of relevance are estimated using weighting schemas as Okapi BM25, e.g. see Wang and Zhu [2009].

PT delivers the same ranking of qPRP, i. e. theoretical optimal performances under qPRP's assumptions, when:

$$\sum_{d' \in RA} \sqrt{P(R|q,d)}\sqrt{P(R|q,d')}\,\rho(d,d') = \sum_{d' \in RA} 2b\sigma_d^2 w_{d'}\rho(d,d') \tag{6.3}$$

or when the two quantities are proportional (this is justified by rank equivalences).

This relation can be exploited to estimate PT's parameters and thus guaranteeing optimality under qPRP. In fact, from Equation 6.3 and focusing on a particular d', optimal settings of PT according to qPRP are characterised by the pairs (b, σ_d^2) and the function $w_{d'}$ that satisfy the following equation:

$$b\sigma_d^2 = \frac{\sqrt{P(R|q,d)}\sqrt{P(R|q,d')}}{2w_{d'}} \tag{6.4}$$

While the parameterisation of PT means that the ranking strategy is more general and configurable than qPRP, this introduces the complexity and burden of having to estimate these parameters. By using this relationship between PT and qPRP, it is possible to directly estimate the parameters of PT without requiring training data and parameter estimation problems. It may also be the case that developments within qPRP, specifically how interference is estimated or approximated, could also be transferred to PT through the relationship between PT and qPRP that has been unveiled here.

6.4 Empirical Analysis of Ranking Approaches

In this section we perform two sets of empirical studies to verify the quality[1] of the document rankings provided by the principles and strategies under investigation. We study the ranking approaches in both the TREC ad-hoc retrieval task and in the TREC diversity retrieval task. In both tasks, evaluation measures are tailored to the assumptions underlying relevance.

In particular, in the ad-hoc task, document relevance is assumed to be independent, as users (or assessors in the TREC case) assess documents independently [Goffman, 1964], i. e. the relevance judgement of a document is not

[1] In terms of retrieval performance.

influenced by other documents. In this case, we study whether considering document dependencies improves retrieval under the independence assumption made within the evaluation framework. Details of this retrieval task and the associated evaluation framework are given in Sections 2.3.2 and 2.3.2.

Vice versa, the diversity retrieval task allows the relevance of documents to be assessed dependently from that of other documents [Zhai et al., 2003]. Within the TREC evaluation framework, this is simulated by defining a set of query-intents associated to a query. The relevance of documents is then assessed with respect to each intent: a document can either be relevant to one or more query-intents or be irrelevant. Evaluation measures then account for document rankings covering all the relevant query-intents, possibly preferring complete intent-coverage to sequences of relevant documents redundantly covering a small handful of intents. For more details about this retrieval task and its evaluation framework, the reader is referred to Sections 2.3.3 and 2.3.3.

6.4.1 Common Experimental Settings

In the following we shall examine the empirical results obtained when considering document rankings of length 100, i. e. rankings composed of the 100 documents with highest probability of relevance to the query topics ($P(R|q, d)$) as ordered by PRP. These documents were then re-ranked according to the alternative ranking approaches.

In preliminary studies we have found that similar performance trends were obtained with longer rankings (e.g. 200, 500, $1,000$, $10,000$), although the absolute value of the performances achieved by the alternative strategies were lower the longer the rankings.

This observation may be explained as follows. Documents retrieved at low ranks (e.g. > 100) are unlikely to be relevant, or to be on topic at all. However, these documents are likely to be well diverse from the top ranked documents. Therefore, when promoting diversity, ranking approaches may as well introduce within the top ranked documents a number of documents that are not relevant, just because they are different from the previous ranked ones. MMR appears to be the ranking approach that is most affected by this problem, when λ is small. While, qPRP appears to be the approach that is least affected: this may be because the estimations of document relationships are weighted by the square roots of the estimated probabilities of relevance. Smaller this probability, lesser the relationships between documents affect the score that is associated to the document by qPRP.

Limitations of the empirical analysis. Note that the considered ranking approaches are characterised, to different extents, by their parameters. Specifically, PT is governed by parameters σ^2 and b, MMR by λ; similarly iPRP and qPRP in their parametric instances are characterised by the choice of β. Parameter settings (i.e. estimation and tuning) may then play an important role in the assessment of approaches' effectiveness.

Within our experiments, we set parameters to their optimal values either on a query-set or on a query-by-query basis (often referred to as "oracle" setting), depending on the objectives of the specific experiment: in each experiment we shall clearly state which parameter-setting strategy was used. While alternative parameter-setting strategies may have been used (e.g. parameter estimation based on query-features, or cross validation of parameter values), we chose to study the considered ranking approaches in their best settings.

Implementation details. All ranking approaches have been implemented in C++ extending the Lemur/Indri APIs[1]. Lemur/Indri had also been used to index the document collections and for retrieving documents that matched the queries. During the indexing phase, common stop-words[2] were removed and Porter stemmer applied.

6.4.2 Empirical Evaluation: Ad-hoc Retrieval Task

In this section we consider the evaluation context of the ad-hoc retrieval task, where documents' relevance judgements are assumed to be independent and evaluation measures are tailored to this assumption.

Document rankings were evaluated in terms of Mean Average Precision (MAP), bpref, Precision at 10 (P@10), and Mean Reciprocal Rank (MRR). Details about

[1] http://www.lemurproject.org

[2] The list of stop-words used in our experiments is that distributed with Lemur/Indri.

these measures are given in Section 2.3.2. Where the ranking approaches required parameter settings, we tuned the parameters with respect to MAP. Principles and strategies were compared on both small and large test collections: Table 6.2 provides the list of adopted collections and the relative statistics.

Next, we first describe the details of the instantiations of the ranking approaches. Then we report and examine the empirical results obtained in this evaluation context.

Name	Description	# Docs	Topics
AP8889	TREC 1, 2, 3	164,597	51-200
WSJ8792	TREC 1, 2, 3	173,252	51-200
LA8990	TREC 678	131,896	301-450
TREC 2001 Web track	WT10g	1,692,096	501-550
TREC 2004 Robust	TREC 8	528,155	301-450 and 601-700 minus 672
TREC 8 ad-hoc	TREC 8	528,155	401-450

Table 6.2: Overview of the collections used for the experiments on the ad-hoc retrieval task.

6.4.2.1 Empirical Settings for the Ranking Approaches

Probability Ranking Principle. Estimations of documents' probabilities of relevance, i. e. $P(R|q,d)$, were derived using the Okapi BM25 scores assigned to query terms that appear into documents. The parameters of Okapi BM25 were set as suggested by Robertson [1992] and scores were normalised. It has been argued that these estimations are rank equivalent to the documents' probabilities of relevance. The same procedure was employed to derive documents' probabilities of relevance for the other ranking approaches. Documents were then ranked by decreasing score, according to what prescribed by PRP.

Maximal Marginal Relevance. We varied the hyper-parameter λ in the interval $[0, 1)$ with steps of 0.1. Note that $\lambda = 1$ reduces MMR into PRP and we explicitly excluded this possibility in our experiments. This choice is motivated by the fact that we want to know how good is MMR when re-ranking documents, and thus we are not interested in obtaining the same ranking of PRP because of the parameter settings. To find the best values of λ, a linear search in the parameter space was performed.

Portfolio Theory. The values of the variances related to the probability estimations are required to implement PT's ranking function. As we already discussed in Section 6.2.3, we resorted to treat variance as a parameter because Okapi BM25 provides only a point-wise estimation of a document's probability of relevance. We investigated the optimal value of variance in combination with the value of parameter b, which encodes the risk propensity of a user. We considered values of $b \in \mathbb{R}$ in the range $[-10, 1] \cup [1, 10]$ with unitary increments and values of $\sigma^2 \in \mathbb{R}^+$ in the range $[10^{-10}, 10^{-1}]$. As for MMR, the value $b = 0$ was explicitly omitted from the explored range because this setting would have reduced PT's ranking function in that of PRP. Note that a positive value of b (i. e. $b \in [1, 10]$) corresponds to the user's will of a ranking that contains diverse documents. Whereas, negative values of b (i. e. $b \in [-10, -1]$) correspond to risk averse users, who do not want diverse document rankings[1]. To find the best results obtained by the possible combinations of parameters we performed a grid search of the parameters space b by σ^2.

Interactive PRP. We considered the parametric instantiation of iPRP given in Section 6.2.6. In particular, we tested two key values of the parameter β, i. e. $+1$ and -1. Setting $\beta = 1$ recreates the standard formulation of iPRP that, at parity of probability of relevance, assigns higher scores to documents that are less correlated with those ranked, and thus more diverse. We refer to this instantiation

[1] When reporting the retrieval performance obtained by PT, we shall not distinguish among performance obtained by positive and negative values of b, i. e. when promoting diversity or not. Conversely, we shall distinguish the two situations (i. e. promote diversity or not) when considering iPRP and qPRP, as we shall discuss next. This choice has been taken for ease of exposition. In fact, as shall be clear later, PT's retrieval performance in the considered settings are dominated by the value of parameter associated with the variance estimation.

as *iPRP+*. Vice versa, when β is set to -1, at parity of probability of relevance, iPRP assignes higher scores to documents that are more correlated with those ranked at previous positions, thus favouring documents that are similar to those already ranked. We refer to this instantiation as *iPRP-*.

Quantum PRP. As for iPRP, also for qPRP we considered its parametric instantiation of Section 6.2.6. We tested to key values of the parameter β, i. e. $+1$ and -1. Similarly to the case of iPRP, the setting with $\beta = 1$ recreates the standard formulation of qPRP, where documents that are different from those already ranked (as estimated using Pearson's correlation) are favoured over other documents. We refer to this instantiation as *qPRP+*. Vice versa, when β is set to -1, the corresponding instantiation of qPRP assigns higher scores to documents that are similar to those already ranked. We refer to this instantiation as *qPRP-*.

6.4.2.2 Results and Discussion

The retrieval effectiveness of the implemented methods is reported in Tables 6.3, 6.4 and 6.5. We also report percentages of improvements achieved by the alternative ranking approaches over PRP. Statistical significant differences with respect to the retrieval performance of PRP is also reported, and is indicated with $*$. To compute statistical significance, we employed a two-tailed paired t-test, with $p < 0.05$.

Best parameters setting for MMR and PT. In the case of MMR, we report for all datasets the results obtained setting $\lambda = 0.9$. These in fact are the best MMR results in the intervals considered during the hyper-parameter tuning. For PT, several pairs of parameters values generated the best runs reported in the tables. All these runs are characterised by a low variance, i. e. $\sigma^2 \leq 10^{-8}$. We also observed that the user's model parameter b did not influence the results as the same rankings are obtained for all values of b in the investigated range, when settings of variance $\sigma^2 \leq 10^{-8}$ are considered.

Collection	Measure	PRP	MMR	PT	iPRP-	iPRP+	qPRP-	qPRP+
AP8889	MAP	0.1941	0.1936 (-0.26%)	0.1941 (+0.00%)	0.1911 (+0.00%)	0.1007 (-48.12%)*	**0.1942** (+0.05%)	0.1427 (-26.48%)*
	bpref	0.2471	0.2470 (-0.04%)	**0.2472** (+0.04%)	0.2444 (-1.09%)	0.1831 (-25.90%)*	0.2454 (-0.69%)	0.2099 (-15.05%)
	P10	0.4193	0.4233 (+0.95%)	0.4193 (+0.00%)	0.4233 (+0.95%)	0.1513 (-63.92%)*	**0.4293** (+2.38%)	0.3347 (-20.18%)
	MRR	0.6172	0.6198 (+0.42%)	0.6172 (+0.00%)	0.5956 (-3.50%)	0.5496 (-10.95%)	**0.6293** (+1.96%)	0.6103 (-1.12%)
WSJ8792	MAP	0.1911	0.1892 (-0.99%)	0.1911 (+0.00%)	**0.1928** (+0.89%)	0.1012 (-47.04%)*	0.1925 (+0.73%)	0.1379 (-27.84%)*
	bpref	0.2390	0.2352 (-1.59%)	0.2391 (+0.04%)	**0.2396** (+0.25%)	0.1660 (-30.54%)*	0.2395 (+0.21%)	0.1949 (-18.45%)
	P10	**0.4393**	0.4340 (-1.21%)	0.4393 (+0.00%)	0.4280 (-2.57%)	0.1587 (-63.87%)	0.4373 (-0.46%)	0.3147 (-28.36%)
	MRR	0.6463	0.6416 (-0.73%)	0.6463 (0.00%)	0.6321 (-2.20%)	0.5653 (-12.53%)	**0.6605** (+2.20%)	0.6344 (-1.53%)

Table 6.3: Retrieval effectiveness of PRP, MMR, PT, iPRP and qPRP on the TREC collections AP8889 and WSJ8792 for the ad-hoc retrieval task. Significance at 0.01 level is calculated using t-test and is indicated with *.

Collection	Measure	PRP	MMR	PT	iPRP-	iPRP+	qPRP-	qPRP+
LA8990	MAP	0.2029	0.1939 (-4.44%)	**0.2043** (+0.69%)	0.1990 (-1.92%)	0.0944 (-53.47%)*	0.1942 (-4.29%)	0.1423 (-29.87%)*
	bpref	0.2014	0.1948 (-3.28%)	**0.2033** (+0.94%)	0.1957 (-2.83%)	0.1156 (-42.60%)*	0.1950 (-3.18%)	0.1578 (-21.65%)*
	P10	0.2567	0.2500 (-2.61%)	**0.2573** (+0.23%)	0.2547 (-0.78%)	0.0733 (-71.45%)	0.2533 (-1.32%)	0.1613 (-37.16%)
	MRR	**0.5110**	0.5026 (-1.64%)	0.5110 (+0.00%)	0.4898 (-4.15%)	0.4342 (-15.03%)	0.4960 (-2.94%)	0.4925 (-3.62%)
TREC 2001 **Web track**	MAP	0.1334	0.1359 (+1.87%)	**0.1341** (+0.52%)	0.1165 (-12.67%)	0.0564 (-57.72%)*	0.1178 (-11.69%)	0.1001 (-24.96%)*
	bpref	**0.1582**	0.1514 (-4.30%)	0.1582 (+0.00%)	0.1383 (-12.58%)	0.0675 (-57.33%)	0.1430 (-9.61%)	0.1210 (-23.51%)
	P10	0.3460	0.3440 (-0.58%)	**0.3480** (+0.58%)	0.2940 (-15.02%)	0.1040 (-64.94%)*	0.3200 (-7.51%)	0.2700 (-21.97%)*
	MRR	**0.5984**	0.5915 (-1.15%)	0.5984 (+0.00%)	0.4854 (-18.88%)	0.5175 (-13.52%)	0.5006 (-16.34%)	0.5904 (-1.34%)

Table 6.4: Retrieval effectiveness of PRP, MMR, PT, iPRP and qPRP on the TREC collections LA8990 and TREC2001 Web track for the ad-hoc retrieval task. Significance at 0.01 level is calculated using t-test and is indicated with *.

Collection	**Measure**	**PRP**	**MMR**	**PT**	**iPRP-**	**iPRP+**	**qPRP-**	**qPRP+**
TREC 2004 **Robust**	MAP	**0.1996**	0.1918 (-3.91%)	0.1996 (+0.00%)	0.1970 (-1.30%)	0.0868 (-56.51%)*	0.1943 (-2.66%)*	0.1306 (-34.57%)*
	bpref	0.2286	0.2181 (-4.59%)	**0.2287** (+0.04%)	0.2247 (-1.71%)	0.0965 (-57.79%)	0.2241 (-1.97%)	0.1466 (-35.87%)
	P10	0.4084	0.4076 (-0.20%)	**0.4088** (+0.10%)	0.3863 (-5.41%)	0.1233 (-69.81%)*	0.3871 (-5.22%)	0.2602 (-36.29%)*
	MRR	**0.6525**	0.6518 (-0.11%)	0.6525 (+0.00%)	0.6167 (-5.49%)	0.5830 (-10.65%)	0.5834 (-10.59%)	0.6416 (-1.67%)
TREC 8 **ad-hoc**	MAP	**0.1853**	0.1668 (-9.98%)*	0.1853 (+0.00%)	0.1801 (-2.81%)	0.0780 (-57.91%)*	0.1822 (-1.67%)	0.1108 (-40.21%)*
	bpref	0.2236	0.2020 (-9.66%)*	**0.2237** (+0.04%)	0.2163 (-3.26%)	0.1078 (-51.79%)*	0.2207 (-1.30%)	0.1434 (-35.87%)*
	P10	**0.4380**	0.4260 (-2.74%)	0.4380 (+0.00%)	0.4240 (-3.20%)	0.1480 (-66.21%)*	0.4320 (-1.37%)	0.2600 (-40.64%)*
	MRR	0.6110	0.6107 (-0.05%)	0.6110 (+0.00%)	0.5975 (-2.21%)	0.5343 (-12.55%)	**0.6310** (+3.27%)	0.5974 (-2.23%)

Table 6.5: Retrieval effectiveness of PRP, MMR, PT, iPRP and qPRP on the TREC collections TREC2004 Robust and TREC8 for the ad-hoc retrieval task. Significance at 0.01 level is calculated using t-test and is indicated with *.

Effectiveness of PRP. Overall, in the considered settings, PRP obtains consistently higher retrieval effectiveness than other ranking approaches. This suggests that empirically PRP is the best ranking model when a document's relevance is assessed independently from that of other documents.

Effectiveness of MMR. For MMR, significantly worse results with respect to PRP baseline are only found when examining MAP and bpref for the experiments on the TREC 8 ad-hoc collection. Few not-significant improvements with respect to PRP are recorded for MAP in TREC 2001 Web track, as well as for P10 and MRR in the AP8889 collection. Despite this, the overall results suggest that ranking search results according to MMR does not lead to improvements over PRP in the context of ad-hoc retrieval. This is not a surprising result because MMR had been devised to select at each rank position documents that, although possibly relevant, are also diverse from those retrieved in previous rank positions.

Effectiveness of iPRP and qPRP. We now turn to consider the retrieval effectiveness of iPRP and qPRP. The standard instantiations of these approaches (represented by iPRP+ and qPRP+ and obtained setting $\beta = +1$) are consistently less effective than PRP. Also, retrieval effectiveness for these settings are significantly different from those obtained by PRP, apart when examining the values measured by MRR. This suggests that diversifying document rankings does not improve the quality of the rankings in the context of the ad-hoc retrieval task, at least when considering iPRP and qPRP.

Retrieval effectiveness is different when considering the instantiations with $\beta = -1$ (i. e. iPRP- and qPRP-). In fact, in this case the effectiveness of iPRP and qPRP is not significantly worse than those of PRP (except for the TREC 2001 Web Track collection). Whereas, iPRP and qPRP achieve higher values of the evaluation measures than PRP in AP8889 (but not when considering bpref) and WSJ8792 (but not when considering P@10 and MRR, in the case of iPRP), although differences are not statistically significant. This is interesting because both iPRP's and qPRP's ranking formulas are effectively different from that of PRP in this context. In fact pairs of documents are unlikely to be perfectly not correlated, i. e. $\rho = 0$, because this would mean (in almost all the case[1]) documents are duplicated: Scholer et al. [2011] showed this situation is possible, however it can be considered rare (i. e. affecting only few pairs of documents) within the considered TREC collections. This excludes the possibility of qPRP-'s ranking formula to be reduced to PRP's. Similarly, pairs of documents in the considered TREC collections are unlikely to be all completely correlated, i. e. $\rho = 1$. This excludes the possibility of iPRP-'s ranking formula to be reduced to PRP's. Then, although iPRP- and qPRP- are characterised by ranking formulas that are not equivalent to that of PRP, they provide rankings that are substantially similar *in retrieval performance* to those of PRP.

Effectiveness of PT. PT performs as well as PRP when parameters are tuned with respect to MAP, i. e. PT's document rankings are consistently identical to PRP's in terms of retrieval performance. Small variations in the document rankings with respect to PRP's rankings are due to rounding errors and limited precision in the computation of very low scores. These occurred at the bottom of the ranking, where two documents at the lowest ranks exchange their positions. No statistically significant differences between PT's and PRP's rankings are individuated by the two-tailed paired t-test.

The findings about PT are in stark contrast with those reported by Wang and Zhu [2009], who showed that PT significantly improves upon PRP. However, note that those results are obtained when relevance information is used to estimate the parameters for each query using 5-fold cross validation. In our empirical investigation, instead, we have tuned PT by performing a grid search of the parameter space for each TREC collection, selecting those parameter-pairs that deliver the highest value of MAP given the whole set of topics (not on a query by query basis, nor on a small query-set basis).

Relationship between parameters settings and effectiveness in PT. We further investigated the relationship between parameters settings and retrieval

[1] Recall that documents are compared with respect to their term-vector representations. That is, documents are represented as bag-of-words. Two documents containing the same terms may differ because of the order the terms appear in the document: this is not captured by the document representations used in our experiment.

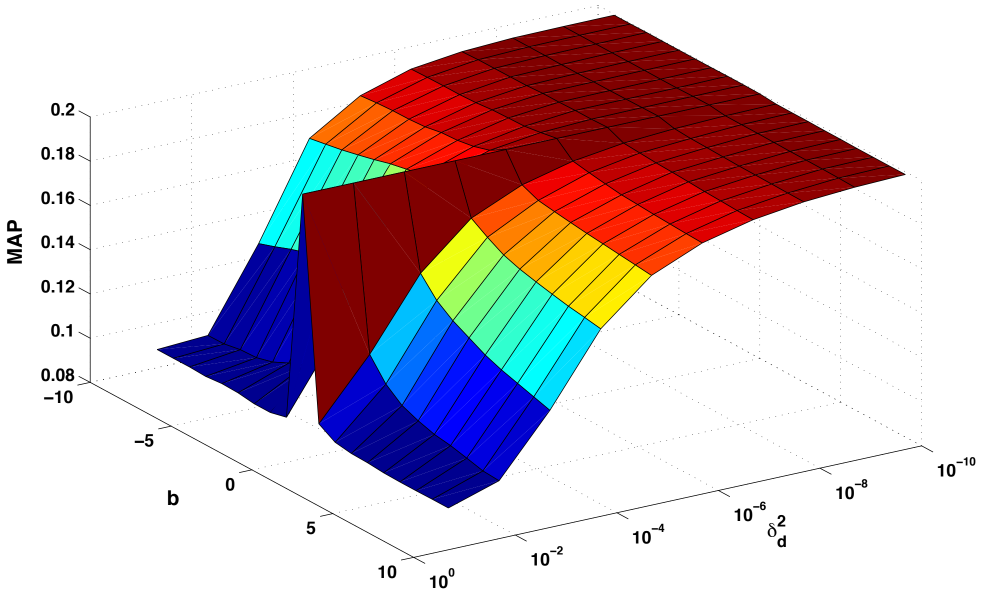


Figure 6.1: Values of MAP obtained on the TREC 2004 Robust dataset using PT and performing a grid-search exploration of the parameter space $b \times \sigma^2$.

performance for PT in our experimental settings. In Figure 6.1 we plotted the values of MAP obtained on the TREC 2004 Robust dataset by PT considering all the parameters pairs used during the grid search of the parameter space. The surface graphed in the represents PT's retrieval performance (MAP) by varying parameter values, i. e. each pair of values of parameters b and σ^2 corresponds to a point on the surface. Note that the values corresponding to $b = 0$ achieve all the same retrieval performance, regardless of the value of the variance. This is because when $b = 0$, the terms that contain the variance σ^2 are suppressed and PT's ranking formula reduces to PRP's one[1]. Similar figures can be produced for other collections. From the figure it is possible to observe that the best values of MAP[2] had been obtained for low values of variance, i. e. $\sigma^2 < 10^{-8}$, regardless of values of the user's model parameter b. Recall that when the best performing parameter settings are used, PT performs as well as PRP. This may be explained as follows. In the case of the TREC 2004 Robust dataset, documents' probabilities of relevance have a magnitude in the order of $10^{-2} \sim 10^{-3}$. The

[1] Recall however that the value $b = 0$ has not been considered in the experiment settings that produced the results of Tables 6.3 and 6.4

[2] Excluding the case $b = 0$.

lower the variance, the less likely that the ordering of two documents imposed by PRP solely on the base of the probability of relevance $P(R|q,d)$ is modified by PT because of the relations between documents, which are weighted by the low variances. As an extreme, for variance values lower than 10^{-8}, the score assigned by PT to each document in the TREC 2004 Robust dataset is substantially rank equivalent to the probabilities of relevance used by PRP for ranking. This situation explains why in such circumstances PT produces the same ranking of PRP. While parameter values are different from zero, thus not formally reducing PT's ranking formula into that of PRP, the low value of the variance estimation determines that document relations are practically ignored by PT, empirically providing document rankings identical to those of PRP.

6.4.2.3 A Follow-Up Experiment: Query-by-Query Parameter Tuning

We performed a follow-up experimentation to verify the impact of parameter tuning on retrieval performances. While previously parameters were tuned with respect to the whole query set, here parameters are tuned on a query by query basis.

Note that while for PT and MMR a wide range of parameters were explored, in the case of iPRP and qPRP we limit to select the value of β that provided the best retrieval results, among only two alternatives, i. e. $+1$ and -1. For iPRP and qPRP, this is equivalent to select for each query the best alternative between diversifying document rankings or promoting documents similar to those ranked at the top. Furthermore, in these settings, the computational costs involved in the tuning process are different for different approaches: the tuning of PT is quadratic in the number of values explored for the parameters, that of MMR is linear in number of the values explored for λ, while iPRP and qPRP have a constant tuning cost (as only two possible values of β are considered). While other settings were possible[1], our settings recreated the cost-relationship involved when using the original approaches: in PT two parameters have to be tuned, in MMR there is only one parameter to tune, while originally both iPRP and qPRP are parameter-free. We indicate the (query-by-query) tuned versions of the ranking approaches with MMRtun, PTtun, iPRPtun and qPRPtun.

[1] For example, more values may have been considered for tuning β.

We also tuned Okapi BM25's parameters b and k_1, which affect the scaling of the document's score by the document length and the scaling of the document's term frequency, respectively. For each query, we selected the pair of parameters that provided the best value of MAP through a grid search of the parameter space defined by $b \in [0.00, 1.00]$ with steps of 0.05 and $k_1 \in [0.00, 1.80]$ with steps of 0.20. The tuning cost of PRP was thus similar to that of PT, and greater than the other approaches. We refer to the tuned version of PRP with the label "PRPtun".

The results of the follow-up experiment are reported in Table 6.6.

In our settings and for *large* collections, the best values of MAP are obtained by extensively tuning PRP. For *smaller* collections, i. e. AP8889, WSJ8792 and LA8990, PRPtun is not always the best performing approach. However, in these cases PRPtun is not sensibly inferior to the best performing alternative approach (apart in LA8990). Among the approaches that attempt to cater for dependencies between document's relevance, PT achieves overall the best performances; however in the TREC 2001 Web track collection the difference between the retrieval performance of PT and qPRP is not noticeable. Furthermore, qPRP performs better than iPRP and MMR (apart in WSJ8792 and TREC 8 ad-hoc, where MMR performs better than qPRP).

The computational overhead required for the tuning of both PRP and PT grows in a quadratic way with the dimensions of the parameter space (i. e. in both cases a grid search of the parameter space is required to tune the approaches). Conversely, the tuning of qPRP, as well as iPRP, simply requires the evaluation of two runs.

The empirical results obtained throughout our experiments suggest that in the context of the ad-hoc retrieval task :

1. parameter estimation is the differentiating factor in assessing the empirical optimality of the ranking approaches,

2. PRP still remains the most performing ranking approach when the independence assumption holds in the evaluation context and model's parameters are not tuned query by query.

Collection	**PRP**	**PRPtun**	**MMRtun**	**PTtun**	**qPRPtun**	**iPRPtun**
AP8889	0.1941	0.2062	0.2060	0.2091	**0.2062**	0.2022
		(+6.23%)	(+6.13%)	(+7.73%)	(+6.23%)	(+4.17%)
WSJ8792	0.1911	0.2070	0.2043	**0.2073**	0.2003	0.1991
		(+8.32%)	(+6.91%)	(+8.48%)	(+4.81%)	(+4.19%)
LA8990	0.2029	0.2181	0.2146	**0.2324**	0.2157	0.2147
		(+7.49%)	(+5.57%)	(+14.54%)	(+6.31%)	(+5.82%)
TREC 2001	0.1334	**0.1564**	0.1467	0.1486	0.1483	0.1409
Web track		(+17.24%)	(+9.97%)	(+11.39%)	(+11.17%)	(+5.62%)
TREC 2004	0.1996	**0.2555**	0.2032	0.2206	0.2091	0.2076
Robust		(+28.01%)	(+1.80%)	(+10.52%)	(+4.76%)	(+4.01%)
TREC 8	0.1853	**0.2058**	0.1924	0.2004	0.1921	0.1894
ad-hoc		(+11.06%)	(+3.83%)	(+8.15%)	(+3.67%)	(+0.59%)

Table 6.6: Values of MAP for PRP and the query by query tuned runs of PRPtun, MMRtun, PTtun, iPRPtun and qPRPtun on several TREC ad-hoc retrieval collections.

Table 6.7: Overview of the collections used for the experiments on the diversity retrieval task.

Name	Description	# Docs	Topics
TREC 678 **(Subtopics)**	FT1 Interactive Track	210,158	20 selected topics
ClueWeb **(Web Diversity)**	ClueWeb09 part B	50,220,423	1-50

When extensive tuning is introduced, PT performs better than any other model, at the expense of the high computational overhead that is required by the tuning procedure. This cost is comparable with the effort required for tuning the parameters of the scoring schema underlying PRP, i. e. Okapi BM25, on a query by query basis. In particular, when PRP is tuned, it delivers a better ranking than PT in large collections. Although it does not provide consistently the best retrieval performance, qPRP has shown to be a solid ranking approach when adapted to the ad-hoc retrieval task, requiring very few tuning. Moreover, the general improvements obtained by approaches other than PRP suggest that, within the ad-hoc retrieval task, some query topics might benefit by results diversification.

6.4.3 Empirical Evaluation: Diversity Retrieval Task

In this section we examine the ranking approaches on the diversity retrieval task, which introduces interdependent document relevance in the evaluation framework through the notion of query-intent coverage. This evaluation framework differs from that of the ad-hoc retrieval task, because rankings are not solely evaluated with respect to the retrieved relevant documents. In fact, the notion of query-intent coverage has a key role in the evaluation of systems in the diversity retrieval tasks: systems that rapidly provide a broad query-intent coverage shall result more effective than systems that retrieve a large amount of relevant documents but that only refer to a single query-intent.

In our empirical analysis, documents rankings were evaluated in terms of subtopics recall at rank 10 (s-r@10), subtopic precision at .2 level of subtopic recall (s-p@.2), α-NDCG at rank 10 (α-NDCG@10) with $\alpha = 0.5$[1], and subtopic mean reciprocal rank at 25% coverage (s-mrr@25%). Details about these measures are given in Section 2.3.3. Where the ranking approaches required parameter settings, we tuned the parameters with respect to α-NDCG@10. We tested the ranking principles and strategies on the TREC 678 interactive dataset and the TREC 2009 ClueWeb Category B data. Table 6.7 provides details and statistics of the employed collections.

The instantiation details of the ranking approaches are consistent to those used in the ad-hoc retrieval task described in Section 6.4.2.1. We tested both settings that intuitively would promote diversity (e.g. iPRP+, qPRP+ and PT with $b > 0$) and settings that lead to a promotion of documents similar to those ranked at top (e.g. iPRP-, qPRP- and PT with $b < 0$). This is in line with the experiment carried on in the ad-hoc retrieval task.

Next, we report and examine the empirical results obtained in the diversity retrieval task.

6.4.3.1 Results and Discussion

The retrieval effectiveness of the ranking approaches tested on the diversity retrieval task are reported in Table 6.8. We also report percentages of improvements achieved by the alternative ranking approaches over PRP. For the ClueWeb dataset, statistical significant differences with respect to the retrieval performance of PRP are also reported, and are indicated with $*$. To compute statistical significance, we employed a two-tailed paired t-test, with $p < 0.05$. Note that performing significance tests on the TREC 678 dataset is not meaningful due to the small dimension of the topic set (i. e. 20 queries), as suggested by van Rijsbergen [2004, pages 178–180].

For both collections, the results reported for MMR are obtained by setting $\lambda = 0.9$. In the case of PT on TREC 678, the best value of $\alpha - NDCG@10$ is obtained for $\sigma^2 \leq 10^{-7}$, regardless of the value of b (i. e. when $\sigma^2 \leq 10^{-7}$ all the

[1] This is consistent with the empirical evaluation performed in the TREC 2009, 2010 and 2011 Web Diversity tracks e.g. see [Clarke et al., 2009a, 2010].

tested values of b provided similar results). As for the results obtained in the ad-hoc retrieval task, this suggests that in the TREC 678 PT's ranking formula reduces to PRP's one, when parameters are tuned on the whole set of query topics. This is not the case however when considering the ClueWeb collection. In this case, the best value of α-NDCG@10 is obtained when setting $\sigma^2 = 10^{-4}$ and $b = -5$.

The best performing ranking approach depends on the collection employed. iPRP+ performs best in TREC 678 and it provides sensible increments in retrieval effectiveness with respect to PRP. Similarly, qPRP+ delivers higher effectiveness than PRP in this collection. We found that in TREC 678, PT is most effective when the parameter controlling the variance (i. e. σ^2) is kept very small: this effectively reduces the ranking provided by PT to that provided by PRP. For this reason, no differences in retrieval effectiveness are found between PT and PRP in this collection.

In ClueWeb instead, PT is the best performing model, and it sensibly improves over PRP, although the measured differences are not statistical significant. The versions of the principles tailored to the promotion of diversity (those with $\beta = +1$, i. e. iPRP+ and qPRP+) deliver mixed results in terms of retrieval effectiveness when compared to PRP. Improvements over PRP are in fact registered for s-p@.2 for both (iPRP+ and qPRP+) and α-NDCG@10 (for only qPRP+); however the two principles do not deliver improvements if considering s-r@10 and s-mrr@25%.

The versions of the principles obtained by setting $\beta = -1$ (i. e. iPRP- and qPRP-) deliver worse retrieval effectiveness than PRP in both collections. This is not surprising, because these settings of the principles would favour documents that are correlated to those ranked in the top positions.

It is somewhat surprising[1] that in our settings MMR does not provide increments in retrieval effectiveness with respect to PRP in both collections. This may be because MMR's parameter λ is selected among few alternative values and is kept constant over all queries in each collection, while alternative empirical investigations have tuned (for example employing a training and testing methodology)

[1] Although Santos et al. [2012] have recently shown that MMR is generally not effective in diversifying search results in web search.

Collection	**Measure**	**PRP**	**MMR**	**PT**	**iPRP-**	**iPRP+**	**qPRP-**	**qPRP+**
TREC 678 **(Subtopics)**	s-r@10	0.3868	0.3732 (-3.52%)	0.3868 (+0.00%)	0.1578 (-59.20%)	**0.4188** (+8.27%)	0.3771 (-2.51%)	0.3967 (+2.56%)
	s-p@.2	0.4175	0.3859 (-7.57%)	0.4174 (-0.02%)	0.2612 (-37.44%)	**0.4371** (+4.69%)	0.3964 (-5.05%)	0.4320 (+3.47%)
	α-NDCG@10	0.426	0.409 (-3.99%)	0.426 (+0.00%)	0.212 (-50.23%)	**0.461** (+8.22%)	0.408 (-4.23%)	0.433 (+1.64%)
	s-mrr@25%	0.2877	0.2605 (-9.45%)	0.2877 (+0.00%)	0.1366 (-52.52%)	**0.3196** (+11.09%)	0.2545 (-11.54%)	0.2979 (+3.55%)
ClueWeb **(Web Diversity)**	s-r@10	0.2486	0.2234 (-10.14%)	**0.2675** (+7.60%)	0.1701 (-31.58%)*	0.2236 (-10.06%)	0.2080 (-16.33%)*	0.2409 (-3.09%)
	s-p@.2	0.1775	0.1826 (+2.87%)	0.1745 (-1.69%)	0.1484 (-16.39%)*	0.1903 (+7.21%)	0.1532 (-13.69%)*	**0.1954** (+10.08%)
	α-NDCG@10	0.137	0.106 (-22.63%)*	**0.151** (+10.22%)	0.090 (-34.31%)*	0.132 (-3.65%)	0.102 (-25.55%)*	0.144 (+5.11%)
	s-mrr@25%	0.2178	0.1999 (-8.22%)	**0.2179** (+0.05%)	0.1888 (13.31-%)*	0.2037 (-6.47%)	0.1933 (-11.25%)*	0.2158 (-0.92%)

Table 6.8: Retrieval effectiveness of PRP, MMR, PT, iPRP and qPRP in the diversity document retrieval task. Significance at 0.01 level is calculated using t-test and is indicated with *. No statistically significant differences are individuated between runs in the TREC 678 due to the limited amount of query topics.

to some extent the parameter over a smaller set of queries (e.g. [Wang and Zhu, 2009]).

Note that although it does not achieve the best overall performances, qPRP+ represents a valuable alternative to the best performing model in both collections. In fact, qPRP+ is superior to iPRP+ in ClueWeb, although iPRP+ is more effective in the TREC 678 collection. Similarly, qPRP+ is superior than PT when considering the TREC 678 collection, although PT delivers the best performance in when considering the ClueWeb collection. Furthermore, in contrast with PT, qPRP does not require costly parameter tuning.

Differently from the ad-hoc retrieval case, PRP does not provide the best retrieval effectiveness, and it is often outperformed by PT (apart in TREC 678), iPRP+ and qPRP+. This is intuitive and confirms the fact that PRP may be sub-optimal when dependencies between documents' relevance assessments are considered in the evaluation context.

6.5 Behavioural Analysis of Ranking Approaches

To provide a deeper understanding of the revision process that is produced by each approach alternative to PRP, in the following we empirically explore the movement of the relevant documents across the ranking.

6.5.1 Experimental Settings

In this behavioural analysis, we employ the ClueWeb09 collection (part B only) and the TREC 2009 and 2010[1] Web Diversity topics and relevance judgements. Similarly to the other experiments in this thesis, documents and queries were stemmed and stop-words were removed: thereafter documents were indexed using the Lemur/Indri toolkit[2].

Documents were retrieved according to a unigram language model with Dirichlet smoothing, where μ was set to $2,500$. For each query, the 100 documents with highest scores were considered for ranking. The ranking of PRP was formed arranging documents in decrease order of scores. Approaches alternative to PRP

[1] This topic-set originally consisted of 50 topics. However, we removed topics 95 and 100 from the set because no relevance assessments for these topics were provided by TREC.

[2] http://lemurproject.org/

were used to re-rank documents. For PT, we regarded both the variance of the probability estimations (σ^2) and b as parameters, and we let them varying in the ranges $[10^{-7}, 10^{-2}]$ (with decimal increments) and $[-10, +10]$ (with unitary increments), respectively. MMR's hyper-parameter was varied in the range $[0, 1]$ with steps of 0.1. We considered the parametric versions of iPRP and qPRP (Equations 6.1 and 6.1), studying values of β varying in the range $[-1, 1]$ with steps of 0.1. Pearson's correlation between (normalised) term frequency representations of documents was employed in all re-ranking approaches.

For each ranking approach, we built a retrieval run by tuning the parameters with respect to α-NDCG@10[1] on a query-by-query basis: that is, for each query, we rank documents using the best parameter values for the query.

6.5.2 Methodology

While our focus is on the kinematics of documents, we report the performance of the runs, to show how the re-ranking affects performance. Specifically, the approaches obtained the following values of α-NDCG@10 [2]:

$$\text{PRP: } 0.137 < \text{qPRP: } 0.172^{*} < \text{PT: } 0.182^{*} < \text{iPRP: } 0.197^{*} < \text{MMR: } 0.205^{*}$$

To illuminate the differences in the re-ranking strategies, we focus on the kinematics of only the relevant documents. In particular, for each ranking approach, we recorded the change in the position of each relevant[3] document between the alternative ranking approach and the PRP. We thus count the number of times and the extent of the promotion or demotion of relevant documents with respect to the PRP. In Figures 6.2 and 6.3 we plot the distributions of the (relevant) document kinematics for ranking strategies and principle, respectively. In the figures, the abscissa zero on the x-axis indicates no movement of documents, greater than zero indicates that the documents have been promoted, while lesser than zero indicates the documents have been demoted. The y-axis shows the frequency of the movement. To assess the symmetry of the kinematics shapes with respect

[1] With $\alpha = 0.5$, set according to the TREC 2009 and 2010 Web Track guidelines.

[2] Where * indicates statistical significant differences with respect to the PRP as measured by a two tailed paired t-test with $p \ll 0.01$. Note that no statistical significant differences were found between the performances of PT, MMR, iPRP and qPRP.

[3] We considered a document relevant if it is relevant to at least one facet/intent.

to the zero-movement abscissa (i. e. the zero on the x-axis) we consider the area under the curve (AUC), that is given by the sum of the frequencies of promotions or demotions for a given approach. Specifically, we define as AUC left (AUCL) the sum of the frequencies for $x \in [-100, -1]$, while AUC right (AUCR) is defined as the sum of the frequencies for $x \in [+1, +100]$. We further extend the notion of AUC to a weighted version (WAUC) which weights each movement amplitude (each x value on the x-axis) by its frequency $f(x)$ and normalises this by the number of movements amplitudes different from zero contained in the considered movement range (note that for some values of x there is no movement). Formally, WAUC for a range $\mathcal{R}$ is defined as:

$$WAUC(R) = \frac{\sum_{x \in \mathcal{R}} |f(x) \cdot x|}{\sum_{x \in \mathcal{R}} \upsilon(x)} \text{ , where } \upsilon(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

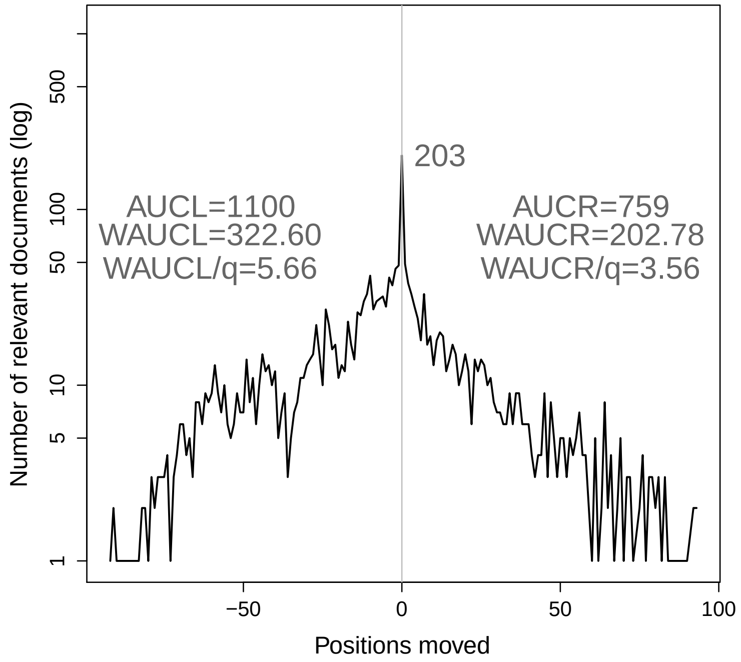
In particular, in the following we consider WAUCL for $x \in [-100, -1]$ (the area on the left of the zero-movement abscissa) and WAUCR for $x \in [+1, +100]$ (the area on the right of the zero-movement abscissa).

6.5.3 Findings

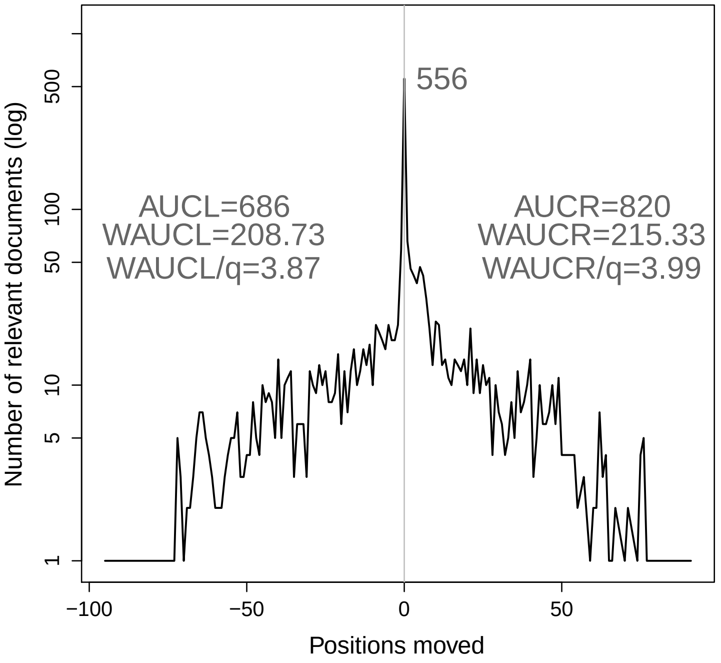
Values of (W)AUCL and (W)AUCR for each ranking approach are reported in Figures 6.2 and 6.3, together with the frequency of the zero-movement (i. e. $f(x = 0)$).

Retrieval strategies (i. e. PT and MMR, Figures 6.3(a) and 6.3(b)) are characterised by wider kinematics shapes that the ones of the principles (i. e. iPRP and qPRP, Figures 6.4(a) and 6.4(b)). MMR appears to be the approach that most revises the position of relevant documents, as it is characterised by the lowest frequency of zero-movements among all approaches. This might be mainly due to the fact that for 57 out of the 98 queries of the TREC 2009-2010 dataset the best performing value of the parameter λ is different from 1: that is, MMR's ranking function effectively provides a ranking different than that of PRP, while for the remaining 41 queries MMR's ranking function reduces to PRP's one (since $\lambda = 1$ for these queries). The movement of relevant documents that is witnessed in Figure 6.3(a) is therefore generated by a high number of queries. While, movements that form the kinematics shapes of other approaches involve a lower number of

Figure 6.2: Kinematics, with respect to the PRP, imposed to the relevant documents by the ranking strategies, i. e. MMR and PT. We also report the values of AUC, WAUC and WAUC/q. Finally, in correspondence to $x = 0$, we report the frequency of zero-movements, i. e. $f(x = 0)$.

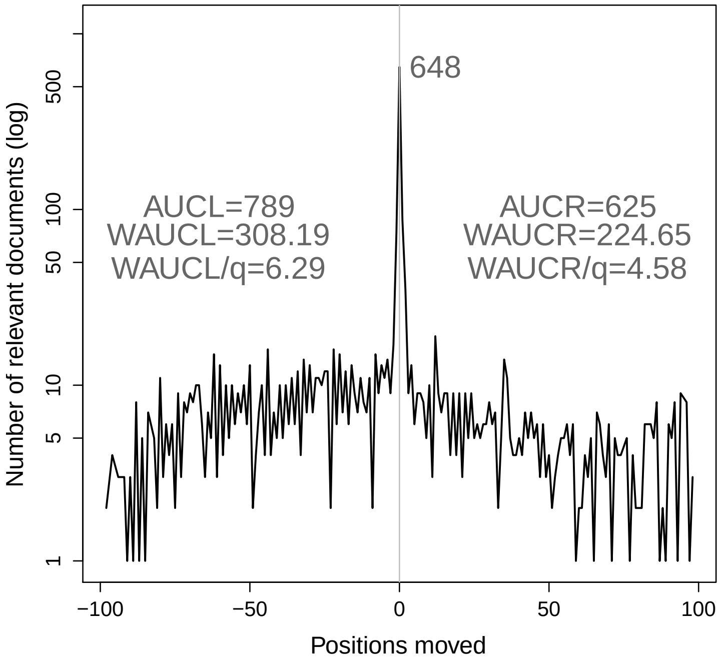


(a) Kinematics for MMR.

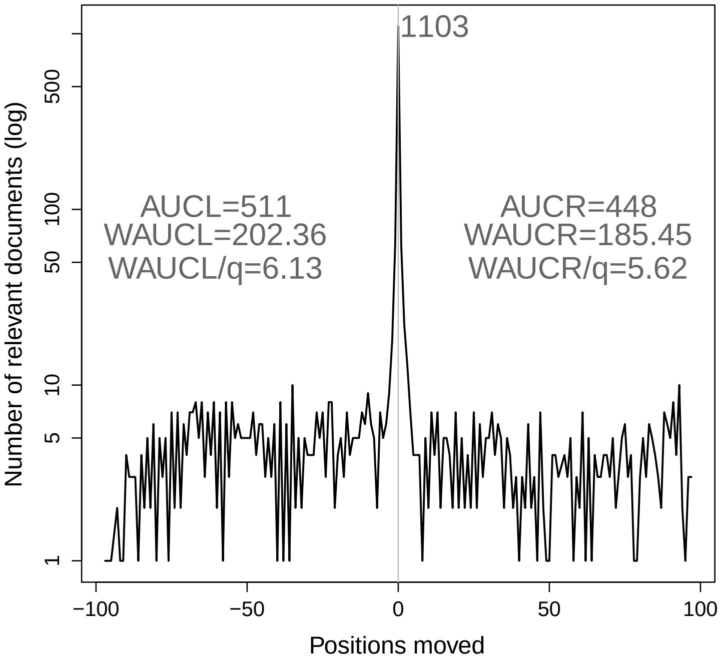


(b) Kinematics for PT.

Figure 6.3: Kinematics, with respect to the PRP, imposed to the relevant documents by the ranking principles, i. e. iPRP and qPRP. We also report the values of AUC, WAUC and WAUC/q. Finally, in correspondence to $x = 0$, we report the frequency of zero-movements, i. e. $f(x = 0)$.



(a) Kinematics for iPRP.



(b) Kinematics for qPRP.

queries. Specifically, the number of queries for which the best performing parameters do not reduce the ranking functions to that of PRP are 54 for PT, 49 for iPRP, 33 for qPRP.

The shape of MMR's kinematics is asymmetric and unbalanced towards the left side of the x-axis. The AUC of MMR confirms this impression: AUCL amounts to 1100, while the AUCR amounts to 759. This suggests that relevant documents are demoted more times than what are promoted. If compared to the kinematics shapes of other approaches, that of MMR can be regarded as being the most unbalanced towards the left side of the x-axis. Nevertheless, MMR achieves the highest value of α-NDCG@10 in our experiments: this might be because the relevant documents that are most demoted are those that are also most redundant, while the relevant documents that get promoted are novel with respect to the ones ranked at previous positions.

The shape of PT's kinematics is similar to the one of MMR's, although PT moves less relevant documents than MMR (higher zero-movement frequency) and its kinematics "ends" sooner than MMR's: no relevant documents are moved of more than 90 positions up or down the ranking. Furthermore, the kinematics of PT seems to favour the promotion of relevant documents over their demotion, as the kinematics shape is slightly unbalanced towards the right of the x-axis. This is confirmed by the difference between AUCR and AUCL; note that PT is the only approach for which AUCR $>$ AUCL. However, the difference between the area under the curve for the left and the right range decreases if WAUC is considered (i. e. WAUCL= 208.73, WAUCR= 215.33): this means that PT promotes relevant documents of fewer positions more than the ones it demotes.

The kinematics of the ranking principles (i. e. iPRP and qPRP) have a common shape. The kinematics are characterised by a high spike in correspondence of the zero-movement coordinate and a fast flattering out shape when movements involve more than half a dozen rank positions (note that the y-axis is in log-scale). The central spike represents no movement of relevant documents with respect to PRP: more relevant documents are moved by iPRP than qPRP. As for MMR, this observation is in line with the number of queries for which iPRP and qPRP provide a ranking different than PRP's one: this happens 49 times (out of 98 queries – i. e. for the 50% of the cases) for iPRP, while only 33 times for qPRP.

For both principles the shapes are asymmetric and slightly unbalanced towards left (AUCL > AUCR).

By comparing the WAUC of the approaches' kinematics, we can understand which strategy promotes or demotes relevant documents of more positions. Note however that a higher WAUC might not be due only to a propensity to promote or demote relevant documents of more positions, but might be as well biased by the number of queries that generated the kinematics. A better indication might be provided by the WAUC-to-query ratio (reported in Figures 6.2 and 6.3), where WAUC is divided by the number of queries for which there has been an effective movement of relevant documents with respect to the PRP. For example, while WAUCR of PT (215.33) is higher than the one of qPRP (185.45), WAUCR-to-query ratio of PT (3.99) is lower than the correspondent value for qPRP (5.62).

Notably, the lowest WAUC-to-query ratio is achieved by MMR with respect to documents that are promoted up the ranking (see WAUCR/q ratio of MMR), suggesting that overall MMR is the approach that less promotes relevant documents. However, MMR is not the approach that most demotes relevant documents, as the WAUCL-to-query ratios of iPRP (6.29) and qPRP (6.13) are higher than that of MMR. The highest promotion of relevant documents is achieved by qPRP (WAUCR/q = 5.62): however this positive characteristic does not seem to find a parallel in the retrieval performances (at least in terms of α-NDCG@10). This might be due to the fact that (i) promoted relevant documents are redundant with respect to those ranked at previous positions, and/or (ii) promotions of relevant documents do not take place within the first 10 rank positions.

The previous analysis clearly shows how each ranking approach moves relevant documents within the ranking. As a further note, we can observe that if little movement transpires then the retrieval results are similar to PRP, while more movement results in greater or lower performance.

6.6 Summary

In this chapter we compared and contrasted analytically, empirically, and behaviourally the ranking approaches we considered in this thesis: PRP, MMR, PT, iPRP, and in particular qPRP.

We have first provided empirical instantiations of each approach suitable for performing ranking experiments on standard TREC collections, both for the ad-hoc and diversity task (Section 6.2). In particular, to perform a controlled evaluation of the ranking approaches we employed the same methods across all ranking approaches for estimating document relevance and diversity consistently.

In our analytical study (Section 6.3), we have shown that links exist between ranking approaches. Furthermore, we were able to unveil formal relationships between PT and qPRP: these may be exploited for setting PT's parameter in absence of training data (Section 6.3.1).

In our empirical analysis (Section 6.4), we compared the approaches within two different evaluation contexts: ad-hoc retrieval (Section 6.4.2) and diversity retrieval (Section 6.4.3). We have shown that in the ad-hoc retrieval task, despite approaches alternative to PRP being overfitted, PRP delivers the best retrieval performance consistently across a number of TREC test collections. This suggests that PRP still remains the most effective ranking approach when the independence assumption holds in the evaluation context. A follow-up study (Section 6.4.2.3) showed that the alternative approaches may potentially deliver performance better than PRP, but parameter estimation is found to be the differentiating factor in assessing the empirical optimality of the ranking approaches in the context of ad-hoc retrieval. From the results of the experiments concerning the alternative evaluation context of diversity retrieval, we were able to conclude that in this situation PRP does not provide the best retrieval effectiveness. In fact, PRP is consistently outperformed by PT, iPRP and qPRP. This finding confirms that PRP is sub-optimal when dependent document relevance is considered in the evaluation context.

Finally, in our behavioural analysis of the ranking approaches (Section 6.5), we studied how approaches differ empirically when deciding whether promote or demote a document given previously ranked documents. To do so, we examined the relevant document kinematics with respect to PRP that the alternative approaches impose on the ranking, and contrasted the obtained kinematics shapes across approaches. To the best of our knowledge, this is the first work that investigates this aspect of ranking approaches. A first observation that transpired from the kinematics analysis was that if little "document-movement" are found (i.e. documents moved of only few ranks up and down the ranking), then the retrieval results are similar to those of PRP. On the other hand, wider kinematics are a sign of higher divergence from PRP, which may lead to a greater or lower effectiveness than that of PRP itself. Secondly, it was observed that ranking strategies as identified in Section 6.1 (i.e. MMR and PT) and ranking principles (i.e. iPRP and qPRP) have different kinematics shapes. In fact, ranking strategies were found to exhibit kinematics where many documents are moved (with respect to PRP) of just few rank positions, and the kinematics itself diminishes rapidly (Figure 6.2). On the other hand, ranking principles were found to have a kinematics that contemplates swaps of documents of many positions (e.g. observe the long tail in the kinematics' shape), although in general these principles move less documents across the ranking than the considered strategies (Figure 6.3).

Chapter 7

Conclusions and Further Work

7.1 Summary of Work and Discussion

The premise of this thesis was that quantum theory and in particular quantum probability theory can be used to model information retrieval more effectively than current approaches.

We focused on document ranking and studied the dominant ranking strategy, the probability ranking principle (Chapter 3). We examined PRP's assumptions and showed how these are related to the optimality of the ranking approach. We highlighted that if the assumptions are not upheld PRP may be sub-optimal (Section 3.6). With this respect, the evaluation context of diversity retrieval is a particular example of an IR task where PRP's assumptions are not upheld. We analysed several ranking approaches, considered alternatives to PRP (Sections 3.7–3.10). These approaches are valid when PRP's assumptions are not met. Specifically, we examined two general parametric strategies, i. e. maximal marginal relevance and portfolio theory for IR. These strategies relax one of the PRP's most stringent assumptions, that of independence between relevance assessments.

Furthermore, we considered an alternative ranking principle, the interactive PRP proposed by Fuhr [2008], which provides a criterion for ranking documents in the context of interactive IR. We showed how iPRP can be instantiated for ranking documents in the first pass of retrieval, i. e. when no user interactions have transpired apart from the submission of a (initial) query.

We subsequently turned our attention to examine how quantum probability theory can be employed to rank documents (RQ1 in Chapter 1). To facilitate this task, we proposed and analysed an analogy between the document ranking scenario and a physical scenario, that of the double slit experiment. This was shown in Section 4.3. We were able to express PRP in terms of our analogy (Section 4.5). Specifically, PRP corresponds to using Kolmogorovian probability theory in the double slit experiment. However, we argued that Kolmogorovian probability theory is not adequate to model the experimental observations that can be collected in the double slit experiment. Experiments carried out in Physics had shown that there exists a substantial discrepancy between what is empirically measured in the double slit experiment and the predictions produced by Kolmogorovian probability theory [Feynman, 1963]. In particular, the Kolmogorovian rule of additivity between probabilities associated with disjoint events is found to fail in the double slit experiment. Conversely, quantum probability theory has been shown to provide predictions that are consistent with the values experimentally measured in the double slit settings. That is, quantum probability theory appears to be *more adequate* to model the double slit experiment than Kolmogorovian probability. Because of our analogy between the double slit experiment and document ranking, we hypothesised that this observation applies to the case of document ranking too (RQ2 in Chapter 1).

The previous observation provided us with the motivation for the development of a ranking principle alternative to PRP and based on quantum probabilities (RQ3 in Chapter 1). Moreover, our analogy provided us with a powerful tool for deriving such a new ranking principle, i. e. the quantum probability ranking principle (Section 4.6). Mathematically, the ranking criterion underlying qPRP is described as

$$\underset{d_B \in \mathcal{B}}{\operatorname{argmax}}\big(p_{d_B} + \sum_{d_A \in \mathcal{A}} I_{d_A d_B}\big) \tag{7.1}$$

where $I_{d_A d_B}$ is the quantum interference that occurs between documents d_A and d_B. We then examined the optimality of qPRP and argued that, *from a theoretical perspective*, ranking according to qPRP results in higher retrieval effectiveness than using alternative approaches (i. e. PRP). This was shown in Section 4.7.

We further turned to examine how quantum interference can be interpreted in information retrieval, and in particular within qPRP (Chapter 5). We argued that interference appears when measuring the relevance of a document in the context of a ranking of other documents. We posited that interference models the extent to which documents share relationships at relevance level, i. e. the extent of their dependencies in terms of relevance (Section 5.3).

We also considered what is the source of the quantum interference term from the perspective of the mathematical formalism we employed (Section 5.2). We defined the notion of complex probability amplitudes, which are essential to correctly describe and model the observations made in the double slit experiment; probabilities are the square of these complex amplitudes. We noticed that the interference term arises when computing the sum of the probabilities of disjoint events. In fact, this sum corresponds to the square of the sum of the correspondent complex probability amplitudes, resulting in the sum of the disjoint probabilities plus a third component, the interference term. We unveiled that the interference term is ultimately dependent upon phase differences between complex amplitudes.

From an empirical perspective, we showed that the retrieval effectiveness qPRP provides depends upon how quantum interference is estimated or approximated (RQ4 in Chapter 1). In fact, at the current stage no method had been devised to formally calculate quantum interference in information retrieval. However, we put the basis for future work on this issue by examining possible approaches to derive a complex-valued representation of documents, which can serve as a tool for deriving quantum interference (Section 5.2.3.2). We also proposed approaches to estimate quantum interference within qPRP that have been shown to be effective in terms of retrieval performance[1] (Sections 5.4 and 5.5). In fact, although such estimations do not guarantee the optimality of qPRP, they allowed us to empirically instantiate the principle so as to validate our proposal.

To validate qPRP and to gain more insights about approaches for document ranking, we then:

1. analysed PRP, qPRP, and other ranking approaches,

[1] When compared against PRP and approaches alternative to PRP.

2. empirically compared the three ranking principles (i. e. PRP, iPRP, and qPRP) and two popular ranking strategies (i. e. MMR and PT) in two retrieval scenarios, those of ad-hoc retrieval and diversity retrieval,

3. analytically contrasted the ranking criteria underlying the examined ranking approaches, exposing similarities and differences,

4. studied the ranking behaviours of approaches alternative to PRP in terms of the kinematics over relevant documents, i. e. by considering the extent and direction of the movements of relevant documents across the ranking that are recorded when comparing PRP with its alternatives.

These aspects were addressed in Chapter 6.

The findings of our empirical investigation in the two document ranking tasks showed that the effectiveness of the examined approaches depends upon the evaluation context. In the traditional evaluation context of ad-hoc retrieval, PRP is shown to be empirically better or comparable to alternative ranking approaches. However, underlying the evaluation context of ad-hoc retrieval is the assumption that the relevance of a document is independent from those of other documents. While PRP complies with this assumption, we showed that alternative approaches are devised for situations where the assumption does not hold.

When we turned to examine evaluation contexts that account for interdependent document relevance, i. e. when the relevance of a document is assessed also with respect to other retrieved documents, as it is the case in the diversity retrieval scenario, then the use of quantum probability theory, and specifically the use of qPRP for ranking documents, has been shown to improve retrieval and ranking effectiveness. In particular, the improvements that we witnessed in our empirical investigation in this evaluation context were comparable to those of previous approaches, such as PT, that differently from qPRP are highly dependent on parameter settings and tunings.

7.2 Contributions

Several contributions emerge within this thesis:

- In Chapter 4 we have proposed an *alternative view of document ranking*, inspired by quantum theory and realised through an analogy with the double slit experiment. Within this analogy, measuring the relevance of a document in context with the other retrieved documents is akin to measure the arrival of a particle on a measuring screen when a screen with several slits is placed between the particle emitter and the measuring screen itself. Because the latter measurements are better modelled by quantum probability theory than Kolmogorovian probability theory, we posit that the same observation may be true in the case of document ranking, under the assumption of interdependent document relevance.

- In Chapter 4 we also proposed a *novel ranking principle*, the quantum probability ranking principle (qPRP), that extends PRP to situations where interdependent document relevance is admitted. This principle is directly derived from the analogy we instructed between the double slit experiment and document retrieval and indeed it prescribes to rank documents according to quantum probabilities.

- In Chapter 5 we provided an *interpretation of quantum interference in information retrieval*. Specifically, we argued that the notion of quantum interference is central to quantum probability theory and to qPRP. In the context of the double slit experiment, quantum interference can be mathematically described as the statistical deviation of the observed measurements from the predictions obtained by the Kolmogorovian rule of additivity of probabilities of disjoint events. In the context of information retrieval, and specifically of document ranking, we suggested that quantum interference is related to interdependent document relevance. In such context, we in fact interpreted quantum interference as the extent to which the relevance of a document is affected by those of other retrieved documents. To make the operationalisation of qPRP in IR possible, we have also proposed a number of *approaches to estimate or approximate the value of quantum interference* when ranking documents.

- Throughout Chapters 4, 5, and 6 we proposed, analysed, and evaluated a number of *empirical instantiations of qPRP for ranking documents in two different retrieval scenarios*, i. e. ad-hoc and diversity retrieval. Our instantiations are based on the hypothesis that quantum interference can be estimated from similarities and differences between documents with respect to the terms they contain and their statistical distributions.

- In Chapter 6 we *compared and contrasted PRP, qPRP and other ranking approaches* on three levels: (i) empirically, (ii) analytically, (iii) behaviourally.

- In Chapter 3 we proposed the first *empirical instantiation of the interactive PRP* [Fuhr, 2008] in the case of first passage retrieval. This instantiation is shown to be empirically comparable with other alternative ranking approaches to PRP in Chapter 6.

- In Chapter 5 we provided *insights on the proposal of using complex numbers in information retrieval* (originally put forward by van Rijsbergen [2004]). In particular, we unveiled that the presence of quantum interference is intimately related to the use of complex probability amplitudes (and indeed complex numbers) for characterising quantum probabilities. We also suggested how complex numbers may be interpreted in the context of the *Quantum Information Retrieval framework* proposed by Piwowarski et al. [2010] and the technique of Latent Semantic Analysis [Landauer, 2006].

- In Chapter 6 we *analytically examined the relationships that stands between the ranking criteria underlying PRP, qPRP and PT*. Specifically, we stated the conditions required by PT to uphold qPRP. From an empirical perspective, such conditions may be used as an approach to bootstrap the parameter instantiation of PT.

- Finally, the empirical results reported in Chapter 6 suggest that the application of quantum theory to problems within information retrieval, as proposed by van Rijsbergen [2004], can lead to improvements in retrieval effectiveness (and in particular in the context of ranking documents under dependent relevance assessments). This consideration is in accordance to the arguments that have been recently put forward by Piwowarski et al. [2010] and Melucci [2011].

7.3 Further Work

Based on the work contained in this thesis, we identified several avenues for future research: we discuss them in the following.

Alternative estimations of interference. In Sections 5.4 and 5.5 we proposed and empirically evaluated a number of approaches to estimate quantum interference within qPRP. These approaches are characterised by two aspects: (i) documents are represented as vectors of terms, and (ii) similarity functions are used to measure the extent of the interference.

Alternative approaches to estimate interference can be sought. For example, one may consider to retain the use of the similarity function to measure the extent of interference, but use a different document representation approach. Instead of considering the whole document, alternative approaches may consider more concise representations, for example based on the most informative terms of each document, or the terms that co-occur with the query-terms within each document. More sophisticated approaches may be based on representing documents as subspaces, as proposed by Zuccon et al. [2009a]. Following this approach, each document would be represented by a multi-dimensional subspace of the whole information space, rather than a single vector (i. e. a one-dimension subspace) as in the current approach. This would necessarily require the use of different metrics for measuring the extent of the interference, which arises when observing two documents. A suitable metric may be that explored in [Zuccon et al., 2009a].

Derivation of a representation based on complex numbers. In Section 5.2 we analysed initial proposals for deriving a complex-valued representation of documents suitable for quantum-like models of IR. We empirically showed that the proposal of encoding term frequency and inverse document frequency counts within magnitudes and phases of complex numbers does not provide a sensible and effective document representation.

However, the use of complex numbers within quantum-like models for IR is undoubtedly an intriguing avenue of research. The group structure that complex numbers exhibit, and specifically the periodic behaviour of phases, may be exploited to encode specific query-intent. For example, similar query intents may be associated with relatively close values of phases, while query intents that consistently differ may be associated with opposite phases, i. e. phases that differ of an additive π factor. According to this representation, the cosine between the representations of documents addressing similar intents would result in a value close to 1; while, if two documents address different intents, then their cosine would be valued -1.

Application of qPRP to tasks other than document ranking. In this thesis we empirically tested qPRP in the context of document retrieval. The ranking approach may be also applied to other domains where interdependent relevance is central. For example, it seems likely that qPRP can be applied to the task of novelty detection in sentence retrieval, which consists of finding relevant and novel sentences in a ranking of documents given a query (e.g. [Allan et al., 2003; Fernández and Losada, 2008; Soboroff and Harman, 2005]). This would require adapting the strategy that is used to estimate interference to the sentence retrieval context, where the unit of retrieval is of a shorter length than when considering documents (i. e. a sentence contains just few terms, while a document is itself usually composed of a number of sentences).

Furthermore, strategies derived from qPRP may be successfully applied to tasks other than ranking. In this case, the key aspect that would be "borrowed" from qPRP is that of interference and how this is manipulated within the mathematical formulation of the ranking principle. With this respect, it is interesting to note that the same intuition underlying qPRP has been recently used in domains other than IR: for instance Busemeyer et al. [2011] developed a model of decision making based on the same quantum-like tools and concepts qPRP employes. Similarly, Franco and Zuccon [2010] formulated a model of tag combination, which was tested in the Delicious[1] taxonomy. However, these works employed quantum probabilities and interference to *describe and explain* some

[1] http://www.del.icio.us/

of the empirical data that is observed when considering human decision making or tag combination, respectively. In particular, both approaches fit the observed data to the quantum-like model, and derive the value of interference (or equivalently of phase difference) that best describes the data. On the other hand, in our approach we aim to *predict* future observations. Thus we built a model that from the observed evidence provides an estimation of the extent of interference that will be measured after some action is taken (e.g. rank a document in a specific position).

Examine complex aspects of the document kinematics. In this thesis we examined an alternative approach to the analysis of the empirical results obtained in ranking tasks: the investigation of the document kinematics. We produced such kinematics by considering the extent to which relevant documents are moved by an approach alternative to PRP when compared to the original document ranking obtained by PRP itself. This analysis allowed us to make conclusive statements about which ranking approach overall promotes relevant documents of more positions, and which demotes them more. Observing the kinematics also allowed us to understand the extent of the movements and that for example iPRP moves (with respect to PRP) relevant documents of more positions than PT as iPRP's kinematics is more extended than PT's.

The investigation of the document kinematics with respect to PRP can be however further extended. In fact, we focused our attention on where relevant documents are moved to by the alternative approaches to PRP. This analysis can be extended by including in the kinematics investigation the relationships between the relevant documents, that is, the fact that each document covers particular query-intents. This would allow to produce an analysis on the document kinematics based on the coverage of the query intent. Producing such kinematics and the relative metrics to quantitatively assess the document movements is however not straightforward. This is because successful approaches to measure such aspects of the document kinematics would have to consider both the extent to which documents overlap and differ in terms of the intents they address.

Appendix A

Notation and Conventions

Notation

- $\mathbf{x}$: a vector
- $\underline{x}_i$: the component i of vector $\mathbf{x}$
- $|\mathbf{x}\rangle$: vector $\mathbf{x}$ (also called ket) expressed in Dirac notation
- $\langle\mathbf{x}|\mathbf{y}\rangle$: an inner product expressed in Dirac notation
- † : Hermitian conjugation
- $\bar{x}$: complex conjugation of the complex number x
- $\bar{\mathbf{x}}$: complex conjugation of the complex valued vector $\mathbf{x}$
- $\mathcal{H}$: a Hilbert space
- $\mathcal{A} = \{d_{A_1}, \ldots, d_{A_n}\}$: a set of n documents
- $\mathcal{A} = \{A_1, \ldots, A_n\}$: a set of n slits
- $< d_1, \ldots, d_n >$: ranking containing d_1 in the first position and d_n in the n-th position
- θ : an angle
- $P(.)$: a probability function
- $p_{.}$: a probability

- $P(R|q,d)$: a conditional probability function (probability of relevance given a query q and a document d)

- $\rho(.,.)$: a correlation function (usually Pearson's correlation

- σ^2 : variance

- $|b|$: the absolute value of a number b

- RE : a set

- $\mathcal{RL}$: a list

- $|RE|$: the size of set RE

- $|\mathcal{RL}|$: the size of list $\mathcal{RL}$

Abbreviations

- IR: Information Retrieval

- PRP: Probability Ranking Principle

- qPRP: quantum Probability Ranking Principle

- MMR: Maximal Marginal Relevance

- PT: Portfolio Theory of Information Retrieval

- iPRP: interactive Probability Ranking Principle

- AP: Average Precision

- RR: Reciprocal Rank

- MRR: Mean Reciprocal Rank

- MAP: Mean Average Precision

- MAP-IA: Intent Aware Mean Average Precision

- DCG: Discounted Cumulative Gain

- nDCG: normalised Discounted Cumulative Gain
- nDCG-IA: Intent Aware normalised Discounted Cumulative Gain
- ERR-IA: Intent Aware Expected Reciprocal Rank
- NRBP: Novelty- and Rank- Biased Precision
- TF: Term Frequency
- IDF: Inverse Document Frequency

Appendix B

Details of Experiments of Section 3.6.4: Variance in Relevance Estimations

In this appendix we provide the details of the experimental procedure used for investigating the variance in relevance estimations across systems participating to the TREC 2009 and 2010 Web diversity track (see Section 3.6.4).

TREC 2009	TREC 2010
input.arsc09web	input.DFalah2010
input.yhooumd09BFM	input.THUIR10Str
input.watrrfw	input.UAMSA10d2a8
input.uvamrftop	input.UAMSA10mSF30
input.UMHOOqlIF	input.UCDSIFTMAP
input.UMHOObm25IF	input.UCDSIFTProb
input.UCDSIFTprob	
input.UamsAw7an3	
input.SIEL09	
input.scutrun3	
input.scutrun2	
input.scutrun1	

Figure B.1: Runs submitted at the TREC 2009 and 2010 Web diversity track and considered in the experiments of Section 3.6.4.

From all runs submitted to TREC, we selected those listed in Figure B.1. These runs were selected because the scores recorded in the ranking-files could be directly translated into probability estimations. In the majority of these ranking-files, in fact, scores represented the log-probability of documents being relevant to queries. Other runs were characterised by scores that could be converted into probabilities through normalisation. On the other hand, other runs were excluded because of missing scores for the retrieved documents, or because scores were obfuscated[1].

[1] e.g. by assigning to each document a score equivalent to the difference between the highest rank at which a document was retrieved and the actual rank of that document.

Appendix C

The Double Slit Experiment in Hilbert Spaces

In this appendix we describe the double slit experiment in terms of Hilbert spaces. To this aim, we first shall introduce the mathematical notation used in this appendix (i.e. the Dirac notation) and then formally describe what a Hilbert space is.

C.1 Dirac Notation

The Dirac notation, also known as the bra-ket notation, has been introduced by Dirac [1939] and is often used within Quantum Theory (as well as in Quantum Information and Quantum Computation) to represent vectors and operators.

In this notation, a vector $\mathbf{x}$ is represented by $|\mathbf{x}\rangle$, i.e. the letter associated with the vector and enclosed within a vertical bar and a right-angled bar. This vector representation is usually called ket. Similarly, the Hermitian conjugate of vector $|\mathbf{x}\rangle$ (obtained from $\mathbf{x}$ by taking its transpose and then applying the complex[1] conjugation on each component, i.e. changing the sign of the imaginary part of the complex valued components) is called a bra, and is represented by $\langle\mathbf{x}|$:

$$|\mathbf{x}\rangle^{\dagger} = \langle\mathbf{x}| \tag{C.1}$$

where † represents the operation of Hermitian conjugation.

In Dirac notation, the inner product between vectors $|\mathbf{x}\rangle$ and $|\mathbf{y}\rangle$ is represented by $\langle\mathbf{x}|\mathbf{y}\rangle$. Similarly, the external product between the same two vectors is

[1] Recall that in quantum theory, vectors are defined over a complex vector space

represented by $|\mathbf{x}\rangle\langle\mathbf{y}|$. Finally, the norm of a vector $\mathbf{x}$, indicated by $||\mathbf{x}||$, relates to the inner product according to:

$$||\mathbf{x}|| = \sqrt{\langle\mathbf{x}|\mathbf{x}\rangle} \tag{C.2}$$

C.2 Hilbert Space

Formally, a Hilbert space $\mathcal{H}$ is a *complex*[1] vector space for which is defined an inner product $\langle\mathbf{x}|\mathbf{y}\rangle : \mathbf{x}, \mathbf{y} \rightarrow \mathbb{C}$ (and where $\mathbf{x}$ and $\mathbf{y}$ are complex-valued vectors) that satisfies the following properties:

- $\langle\mathbf{y}, \mathbf{x}\rangle$ is the complex conjugate[2] of $\langle\mathbf{x}, \mathbf{y}\rangle$. In particular, the product $\langle\mathbf{x}, \mathbf{y}\rangle$ is non-commutative, i.e. $\langle\mathbf{x}, \mathbf{y}\rangle \neq \langle\mathbf{y}, \mathbf{x}\rangle$;
- the inner product is distributive with respect to the sum, i.e. $\langle\mathbf{x}|(|\mathbf{z}\rangle+|\mathbf{y}\rangle) = \langle\mathbf{x}|\mathbf{z}+\mathbf{y}\rangle = \langle\mathbf{x}|\mathbf{z}\rangle + \langle\mathbf{x}|\mathbf{y}\rangle$;
- it is possible to rescale the arguments of an inner product with complex scalars (i.e. the inner product is associative with respect to multiplication by a number), e.g. $\langle\lambda\mathbf{x}, \mathbf{y}\rangle = \lambda\langle\mathbf{x}, \mathbf{y}\rangle$, with $\lambda \in \mathbb{C}$;
- $\langle\mathbf{x}, \mathbf{x}\rangle$ is positive definite, i.e. $\langle\mathbf{x}, \mathbf{x}\rangle > 0, \forall\mathbf{x} \in \mathcal{H}$, and $\langle\mathbf{x}, \mathbf{x}\rangle = 0$ iff. $\mathbf{x} = \mathbf{0}$, i.e. the null vector.

Note that real-valued vector spaces[3], such those we are used in IR for representing terms and documents (see the vector space model for IR, discussed in Section 2.2.2), are a subset of complex-valued vector spaces. Similarities and differences between the vector space model used in IR and the use of Hilbert spaces as suggested by van Rijsbergen [2004] have been (briefly) examined by Li and Cunningham [2008].

[1] That is, the components of the vectors belonging to the space $\mathcal{H}$ are complex values numbers $\mathbb{C}$.

[2] We indicate the complex conjugate of a complex quantity z (i.e. a number, a vector, etc.) with the notation $\bar{z}$. In the case of the inner product, we indicate with $\overline{\langle\mathbf{x}, \mathbf{y}\rangle}$ the complex conjugate of $\langle\mathbf{x}, \mathbf{y}\rangle$.

[3] i.e. the components of the vectors belonging to the space are real values numbers, i.e. $\mathbf{x} \in \mathbb{R}$.

C.3 The Double Slit Experiment

Consider the settings of Figure 4.1. Let $|\mathbf{s}\rangle$ represent the state of a particle as it leaves the source. Similarly, let $|\mathbf{A}\rangle$ and $|\mathbf{B}\rangle$ represent the event of a particle passing through slit A when B is closed, and vice versa. Note that[1] $|\mathbf{A}\rangle\langle\mathbf{A}| + |\mathbf{B}\rangle\langle\mathbf{B}|$ is equivalent to the identity operator in the case depicted by Figure 4.1, i.e. in the configuration with only two slits.

The probability amplitude associated with a particle being emitted by the source and passing through slit A is given by $\langle\mathbf{A}|\mathbf{s}\rangle$; vice versa, $\langle\mathbf{B}|\mathbf{s}\rangle$ represents the amplitude associated with passing through slit B. Similarly, $\langle\mathbf{x}|\mathbf{A}\rangle$ is the probability amplitude associate with a particle passing through slit A and hitting the measuring screen at a location x. Vice versa, when the particle passes through B, the probability amplitude $\langle\mathbf{x}|\mathbf{B}\rangle$ is obtained.

The probability amplitude of measuring a particle at location x when emitted by the source s and with both slits open is indicated by $\langle\mathbf{x}|\mathbf{s}\rangle$ and is calculated using the following equation:

$$\langle\mathbf{x}|\mathbf{s}\rangle = \langle\mathbf{x}|\mathbf{A}\rangle\langle\mathbf{A}|\mathbf{s}\rangle + \langle\mathbf{x}|\mathbf{B}\rangle\langle\mathbf{B}|\mathbf{s}\rangle \tag{C.3}$$

Following Equation 4.2 which establishes the relation between probability amplitudes and probabilities, the probability associated with the event of measuring at location x a particle emitted by source s when both slits are open can be derived from the probability amplitude $\langle\mathbf{x}|\mathbf{s}\rangle$ as follows:

$$\begin{aligned} P(\mathbf{x}|\mathbf{s}) &= |\langle\mathbf{x}|\mathbf{s}\rangle|^2 \\ &= |\langle\mathbf{x}|\mathbf{A}\rangle\langle\mathbf{A}|\mathbf{s}\rangle|^2 + |\langle\mathbf{x}|\mathbf{B}\rangle\langle\mathbf{B}|\mathbf{s}\rangle|^2 + 2Re[\langle\mathbf{x}|\mathbf{A}\rangle\langle\mathbf{A}|\mathbf{s}\rangle\langle\mathbf{x}|\mathbf{B}\rangle\langle\mathbf{B}|\mathbf{s}\rangle] \end{aligned} \tag{C.4}$$

where $Re[\langle\mathbf{x}|\mathbf{A}\rangle\langle\mathbf{A}|\mathbf{s}\rangle\langle\mathbf{x}|\mathbf{B}\rangle\langle\mathbf{B}|\mathbf{s}\rangle]$ is the real part of the complex number given by the multiplication of the four probability amplitudes associated with the events of being emitted by s and passing through a slit (either A or B) and of being detected at location x once passed through a slit (either A or B).

[1] $|\mathbf{A}\rangle\langle\mathbf{A}|$ is the projection operator that projects on the state of slit A.

Appendix D

Proofs of Equations 4.4-4.7

D.1 Proof of Equation 4.4

For any complex number z

$$z\bar{z} = |z|^2$$

Proof Let $z = a + ib$. Then,

$$z\bar{z} = (a + ib) \cdot (a - ib) = a^2 - iab + iab - i^2b^2 = a^2 + b^2 = |z|^2$$

□

D.2 Proof of Equation 4.5

For any complex number z_1 and z_2

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$$

Proof Let $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$. Then,

$$\begin{aligned}\overline{z_1 + z_2} &= \overline{(a_1 + a_2) + i(b_1 + b_2)} = (a_1 + a_2) - i(b_1 + b_2) \\ &= (a_1 - ib_1) + (a_2 - ib_2) = \overline{z_1} + \overline{z_2}\end{aligned}$$

□

D.3 Proof of Equation 4.6

For any complex number z

$$\bar{z} = re^{-i\theta}$$

Proof Let $z = r(\cos\theta + i\sin\theta)$. Then,

$$\bar{z} = r(\cos\theta - i\sin\theta) \tag{D.1}$$

Also, recall the following trigonometric relations:

$$\begin{aligned} \sin(-\theta) &= -\sin(\theta) && \text{(D.2)} \\ \cos(-\theta) &= \cos(\theta) && \text{(D.3)} \end{aligned}$$

Therefore, using Equations D.2 and D.3 in Equation D.1, we obtain:

$$\bar{z} = r(\cos(-\theta) + i\sin(-\theta)) = re^{-i\theta}$$

□

D.4 Proof of Equation 4.7

For any complex number z

$$|z|^2 = r^2$$

Proof Let $z = r(\cos\theta + i\sin\theta)$ and recall the result demonstrated in D.1. Then

$$\begin{aligned} |z|^2 &= z\bar{z} = r \cdot r(\cos\theta + i\sin\theta) \cdot (\cos\theta - i\sin\theta) \\ &= r^2(\cos^2\theta - i\sin\theta\cos\theta + i\sin\theta\cos\theta - i^2\sin^2\theta) = r^2(\cos^2\theta + \sin^2\theta) \\ &= r^2 \end{aligned}$$

because $\cos^2\theta + \sin^2\theta = 1$.

□

References

L. Accardi. Some Trends and Problems in Quantum Probability. In L. Accardi, A. Frigerio, and V. Gorini, editors, *Quantum Probability and Applications to the Quantum Theory of Irreversible Processes*, volume 1055 of *Lecture Notes in Mathematics*, pages 1–19. Springer, 1984. 103

M. Agosti, N. Ferro, C. Peters, M. de Rijke, and A.F. Smeaton. *Multilingual and Multimodal Information Access Evaluation: International Conference of the Cross-Language Evaluation Forum*, volume CLEF 2010. Springer, 2010. 128

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM'09)*, pages 5–14. ACM, 2009. 26, 29, 60

J. Allan. HARD Track Overview in TREC 2003: High Accuracy Retrieval from Documents. In *Proceedings of the 12th TExt Retrieval Conference*, pages 24–37, Gaithersburg, Maryland, 2003. 21

J. Allan, C. Wade, and A. Bolivar. Retrieval and Novelty Detection at the Sentence Level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, pages 314–321, 2003. 174

R. Aly, A. Doherty, D. Hiemstra, and A.F. Smeaton. Beyond Shot Retrieval: Searching for Broadcast News Items Using Language Models of Concepts. In *Advances in Information Retrieval (ECIR'10)*, volume 5993 of *Lecture Notes in Computer Science*, pages 241–252. Springer, 2010. 67

G. Amati and C. J. van Rijsbergen. Probabilistic Models of Information Retrieval based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems (TOIS)*, 20:357–389, October 2002. 16, 57

B. Andreopoulos, X. Huang, A. An, D. Labudde, and Q. Hu. Promoting Diversity in Top Hits for Biomedical Passage Retrieval. In Z. Ras and A. Dardzinska, editors, *Advances in Data Management*, volume 223 of *Studies in Computational Intelligence*, pages 371–393. Springer, 2009. 60

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, volume 463. ACM, 1999. 13

R.K. Belew. *Finding out About*. Cambridge University Press, 2000. 21

N.J. Belkin. Intelligent Information Retrieval: Whose Intelligence. In *Proceedings of the Fifth International Symposium for Information Science (ISI'96)*, volume 96, pages 25–31, 1996. 68

J.S. Bell. On the Einstein-Podolsky-Rosen Paradox. *Physics*, 1(3):195–200, 1964. 104

S. J. Benkoski, M. G. Monticino, and J. R. Weisinger. A Survey of the Search Theory Literature. *Naval Research Logistics (NRL)*, 38(4):469–494, 1991. 37

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003. 29, 63, 104

A. Bookstein. Information Retrieval: A Sequential Learning Process. *Journal of the American Society for Information Science*, 34(5):331–342, 1983. 55

A. Bookstein. Outline of a General Probabilistic Retrieval Model. *Journal of Documentation*, 39(2):63–72, 1993. 39, 42

A. Bookstein and D.R. Swanson. Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information science*, 25(5):312–316, 1974. 15

P. Borlund. The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research*, 8(3):8–3, 2003. 101

M.M. Bouamrane, C. Macdonald, I. Ounis, and F. Mair. Protocol-driven Searches for Medical and Health-sciences Systematic Reviews. In *Advances in Information Retrieval Theory (ICTIR '11)*, pages 188–200. Springer, 2011. 21

G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78, 1950. 46

Sergey Brin and Lawrence Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer networks and ISDN systems*, 30:107–117, April 1998. 19

P. Bruza, K. Kitto, D. Nelson, and C. McEvoy. Is there something Quantum-like about the Human Mental Lexicon? *Journal of Mathematical Psychology*, 53 (5):362 – 377, 2009a. 104

P. Bruza, D. Sofge, W. F. Lawless, C. J. van Rijsbergen, and M. Klusch, editors. *Proceedings of Quantum Interaction, Third International Symposium (QI'2009)*, volume 5494 of *Lecture Notes in Computer Science*, 2009b. Springer. 108

P. D. Bruza and R. Cole. Quantum Logic of Semantic Space: An Exploratory Investigation of Context Effects in Practical Reasoning. In S. Artemov, H. Barringer, A.S. d'Avila Garcez, and J.H. Woods, editors, *We Will Show Them: Essays in Honour of Dov Gabbay*, pages 339–361. College Publications, 2005. 5

J. R. Busemeyer, Z. Wang, and J. T. Townsend. Quantum Dynamics of Human Decision Making. *Journal of Mathematical Psychology*, 50(3):220 – 241, 2006. 5

J.R. Busemeyer, E.M. Pothos, R. Franco, and J.S. Trueblood. A quantum Theoretical Explanation for Probability Judgment Errors. *Psychological Review*, 118(2):193, 2011. 174

S. Calegari and G. Pasi. Gronto: A Granular Ontology for Diversifying Search Results. In *Proceedings of the Second Italian Information Retrieval Workshop (IIR'10)*, volume 560, pages 59–63. CEUR Workshop Proceedings, 2010. 71

J. Carbonell and J. Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, SIGIR '98, pages 335–336. ACM, 1998. 2, 60, 62

B. Carterette. An Analysis of NP-completeness in Novelty and Diversity Ranking. *Information Retrieval*, 14:89–106, 2011. 27, 28, 61

P. Chandar and B. Carterette. Diversification of Search Results using Webgraphs. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10)*, pages 869–870. ACM, 2010. 71

O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM '09)*, pages 621–630. ACM, 2009. 26

H. Chen and D. R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 429–436. ACM, 2006. 2, 55, 56, 60, 113

O. Al. Choustova. Quantum Bohmian Model for Financial Market. *Physica A: Statistical Mechanics and its Applications*, 374(1):304 – 314, 2007. 5

O. Al. Choustova. Quantum Probability and Financial Market. *Information Sciences*, 179(5):478 – 484, 2009. 5

C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, Z. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st annual international ACM SIGIR* *conference on Research and development in information retrieval (SIGIR'08)*, pages 659–666. ACM, 2008. 25, 26

C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of the 18th Text Retrieval Conference*, 2009a. 24, 155

C. L. A. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Advances in Information Retrieval Theory (ICTIR'09)*, volume 5766 of *Lecture Notes in Computer Science*, pages 188–199. Springer, 2009b. 26

C. L. A.. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the trec 2010 web track. In *Proceedings of the 19th text retrieval conference, Gaithersburg, Maryland*, 2010. 24, 155

C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM'11)*, pages 75–84. ACM, 2011. 27

J.F. Clauser, M.A. Horne, A. Shimony, and R.A. Holt. Proposed Experiment to Test Local Hidden-variable Theories. *Physical Review Letters*, 23(15):880–884, 1969. 104

C. W. Cleverdon. The Significance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'91)*, pages 3–12. ACM, 1991. 20

D.W. Cohen. *An Introduction to Hilbert Space and Quantum Logic*. Springer, 1989. 72

W. S. Cooper. Expected Search Length: A Single Measure of Retrieval Effectiveness based on the Weak Ordering Action of Retrieval Systems. *American Documentation*, 19(1):30–41, 1968. 40

W. S. Cooper. The Inadequacy of Probability of Usefulness as a Ranking Criterion for Retrieval System Output. Unpublished working paper, School of Librarianship, University of California, Berkeley, October 1972. 38

E. Cosijn and P. Ingwersen. Dimensions of Relevance. *Information Processing & Management*, 36(4):533–550, 2000. 35, 101

B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 2009. 29

W. B. Croft and D. J. Harper. Using Probabilistic Models of Document Retrieval without Relevance Information. *Journal of documentation*, 35(4):285–295, 1979. 17

W. Bruce Croft. Combining Approaches to Information Retrieval. In *Advances in Information Retrieval*, volume 7 of *The Kluwer International Series on Information Retrieval*, pages 1–36. Springer, 2002. 58

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional Relevance: A New Aggregation Criterion. In *Advances in Information Retrieval (ECIR'09)*, volume 5478 of *Lecture Notes in Computer Science*, pages 264–275. Springer, 2009. 101

A. Das Sarma, S. Gollapudi, and S. Ieong. Bypass Rates: Reducing Query Abandonment using Negative Inferences. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08)*, pages 177–185. ACM, 2008. 32

L. De Vine and P. Bruza. Semantic Oscillations: Encoding Context and Structure in Complex Valued Holographic Vectors. In *Quantum Informatics for Cognitive, Social, and Semantic Processes (QI'10)*, 2010. 108

E. Di Buccio, M. Melucci, and D. Song. Towards Predicting Relevance Using a Quantum-Like Framework. In *Advances in Information Retrieval (ECIR'11)*, volume 6611 of *Lecture Notes in Computer Science*, pages 755–758. Springer, 2011. 103, 114

P. A. M. Dirac. *The Principles of Quantum Mechanics*. The International Series of Monographs on Physics. Clarendon Press, Oxford, second edition edition, 1939. 181

J. M. Dobbie. Search Theory: A Sequential Approach. *Naval Research Logistics Quarterly*, 10(1):323–334, 1963. 37

S. M. Dowdy, S. Wearden, and D. M. Chilko. *Statistics for Research*. Wiley-Blackwell, 2004. 32

F. Dubois. On Voting Process and Quantum Mechanics. In P. Bruza, D. Sofge, W. F. Lawless, C. J. van Rijsbergen, and M. Klusch, editors, *Proceedings of Quantum Interaction, Third International Symposium (QI'2009)*, volume 5494 of *Lecture Notes in Computer Science*, pages 200–210. Springer, 2009. 5

H. G. Dyke. A Figure-of-Merit Ordering System for a Search Output. *American Documentation*, 10(1):85–86, 1959. 38

A. Einstein, B. Podolsky, and N. Rosen. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47:777–780, May 1935. 104

M. Eisenberg and C. Barry. Order Effects: A Study of the Possible Influence of Presentation Order on User Judgments of Document Relevance. *Journal of the American Society for Information Science*, 39(5):293–300, 1988. ISSN 1097-4571. 2, 54

R. T. Fernández and D. E. Losada. Novelty as a Form of Contextual Re-ranking: Efficient KLD Models and Mixture Models. In *Proceedings of the second international symposium on Information interaction in context (IIiX'08)*, pages 27–34, 2008. 174

Richard Feynman. *The Feynman Lectures on Physics*, volume 3 of *The Feynman Lectures on Physics*. Addison-Wesley, Boston, 1963. 81, 168

R.P. Feynman. The Concept of Probability in Quantum Mechanics. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 533–541, 1951. 74, 78

W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc., 1992. 1

R. Franco. Quantum Amplitude Amplification Algorithm: An Explanation of Availability Bias. In P. Bruza, D. Sofge, W. Lawless, C. J. van Rijsbergen, and M. Klusch, editors, *Proceedings of Quantum Interaction, Third International Symposium (QI'2009)*, volume 5494 of *Lecture Notes in Computer Science*, pages 84–96. Springer, 2009. 5

R. Franco and G. Zuccon. Social Tagging, Guppy Effect and the Role of Interference: A Quantum-inspired Model for Tags Combination. In *Modelling the Non-Separability of a Very Complex World*, 2010. 174

D. Frank Hsu and Isak Taksa. Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval. *Information Retrieval*, 8:449–480, 2005. 58

N. Fuhr. Probabilistic Models in Information Retrieval. *The Computer Journal*, 35(3):243, 1992. 16, 33, 40

Norbert Fuhr. A Probability Ranking Principle for Interactive Information retrieval. *Information Retrieval*, 11(3):251–265, 2008. 7, 31, 68, 69, 167, 172

A. Fujii, M. Iwayama, and N. Kando. The Patent Retrieval Task in the Fourth NTCIR Workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'04)*, pages 560–561. ACM, 2004. 21

L. Gabora and D. Aerts. Contextualizing Concepts using a Mathematical Generalization of the Quantum Formalism. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(4):327–358, 2002. 5

F. Gebhardt. A Simple Probabilistic Model for the Relevance Assessment of Documents. *Information Processing & Management*, 11(1-2):59 – 65, 1975. 33

W. Goffman. On Relevance as a Measure. *Information Storage and Retrieval*, 2 (3):201 – 203, 1964. 21, 54, 139

W. Goffman. An Indirect Method of Information Retrieval. *Information Storage and Retrieval*, 4(4):361 – 373, 1968. 46

M. D. Gordon and P. Lenk. A Utility Theoretic Examination of the Probability Ranking Principle in Information Retrieval. *Journal of the American Society for Information Science*, 42(10):703–714, 1991. 2, 37, 39, 40, 45

M. D. Gordon and P. Lenk. When is the Probability Ranking Principle Suboptimal? *Journal of the American Society for Information Science*, 43(1):1–14, 1992. ISSN 1097-4571. 2, 37, 39, 44, 45, 46, 55

D. Harman. *Ranking Algorithms*, chapter 14. Prentice-Hall, Inc., 1992. 32

S.P. Harter. A Probabilistic Approach to Automatic Keyword Indexing. Part II. An Algorithm for Probabilistic Indexing. *Journal of the American Society for Information Science*, 26(5):280–289, 1975. 16

D. Hawking. Challenges in Enterprise Search. In *Proceedings of the 15th Australasian database conference (ADC'04)*, volume 27, pages 15–24. Australian Computer Society, Inc., 2004. 21

J. He, E. Meij, and M. de Rijke. Result Diversification Based on Query-specific Cluster Ranking. *Journal of the American Society for Information Science and Technology*, 62(3):550–571, 2011. 63

D. L. Hemmick. *Hidden Variables and Nonlocality in Quantum Mechanics*. PhD thesis, Rutgers University, May 1997. 104

F. Herbut. Quantum Interference Viewed in the Framework of Probability Theory. *American Journal of Physicsournal of physics*, 60:146, 1992. 105, 107

D. Hiemstra. A Linguistically Motivated Probabilistic Model of Information Retrieval. In *Research and Advanced Technology for Digital Libraries (ECDL'98)*, volume 1513 of *Lecture Notes in Computer Science*, pages 515–515. Springer, 1998. 18

D. Hiemstra. *Using Language Models for Information Retrieval.* PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001. 2, 18, 19, 36

Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'99)*, pages 50–57. ACM, 1999. 29, 63

P. Ingwersen. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of documentation*, 52(1):3–50, 1993. 68

K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20:422–446, October 2002. 22

H. Joho, T. Kato, N. Kando, K. Kishida, and M. Yoshioka, editors. *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, 2010. 128

C. Jönsson. Electron Diffraction at Multiple Slits. *American Journal of Physics*, 42:4–11, January 1974. 74

D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009. 20

K. Kitto, B. Ramm, P. Bruza, and L. Sitbon. Testing for the Non-Separability of Bi-ambiguous Compounds. In *Quantum Informatics for Cognitive, Social, and Semantic Processes: Papers from the AAAI Fall Symposium*, pages 62–69, 2010. 104

K. Kitto, B. Ramm, L. Sitbon, and P. Bruza. Quantum Theory Beyond the Physical: Information in Context. *Axiomathes*, 21:331–345, 2011. 104

Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'02)*, pages 27–34. ACM, 2002. 19

F. Laloë. Do we really Understand Quantum Mechanics? Strange Correlations, Paradoxes, and Theorems. *American Journal of Physics*, 69:655, 2001. 104

T. K. Landauer. *Latent Semantic Analysis.* John Wiley & Sons Ltd, 2006. 110, 172

Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127. ACM, 2001. ISBN 1-58113-331-6. doi: http://doi.acm.org/10.1145/383952.383972. URL `http://doi.acm.org/10.1145/383952.383972`. 2

L. Lee. Measures of Distributional Similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pages 25–32, 1999. 118, 120

T. Leelanupab, G. Zuccon, and J. M. Jose. Revisiting Sub-topic Retrieval in the ImageCLEF 2009 Photo Retrieval Task. In H. Muller, P. Clough, T. Deselaers, and B. Caputo, editors, *ImageCLEF - experimental evaluation in image retrieval*, volume 32 of *The Information Retrieval Series*, pages 277–294. Springer, 2010a. 63

T. Leelanupab, G. Zuccon, and J. M. Jose. When Two Is Better Than One: A Study of Ranking Paradigms and Their Integrations for Subtopic Retrieval. In *Information Retrieval Technology - The 6th Asia Information Retrieval Societies Conference (AIRS'10)*, volume 6458, pages 162–172. Springer, 2010b. 29, 63, 71

T. Leelanupab, G. Zuccon, and J. M. Jose. A Query-Basis Approach to Parametrizing Novelty-Biased Cumulative Gain. In *Advances in Information Retrieval Theory (ICTIR'11)*, volume 6931 of *Lecture Notes in Computer Science*, pages 327–331. Springer, 2011. 26

Y. Li and H. Cunningham. Geometric and Quantum Methods for Information Retrieval. *SIGIR Forum*, 42(2):22–32, 2008. 182

H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of research and development*, 1:309–317, October 1957. 12

C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 1, 14, 15, 33

T. V. Marcella. Quantum Interference with Slits. *European Journal of Physics*, 23(6):615, 2002. 102

H. M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Wiley, 2nd edition, 1991. 58, 64

M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM (JACM)*, 7:216–244, July 1960. 35, 38

M. Melucci. A Basis for Information Retrieval in Context. *ACM Transactions on Information Systems (TOIS)*, 26:14:1–14:41, June 2008. 5

M. Melucci. An Investigation of Quantum Interference in Information Retrieval. In H. Cunningham, A. Hanbury, and S. Rüger, editors, *Advances in Multidisciplinary Retrieval (IRFC'10)*, volume 6107 of *Lecture Notes in Computer Science*, pages 136–151. Springer, 2010. 114

M. Melucci. Can Information Retrieval Systems be Improved using Quantum Probability? In *Advances in Information Retrieval Theory (ICTIR'11)*, pages 139–150. Springer, 2011. 173

S. Mizzaro. Relevance: the Whole History. *Journal of the American Society for Information Science*, 48(9):810–832, 1997. 35, 101

A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27:2:1–2:27, December 2008. 27

Y. Moshfeghi, G. Zuccon, and J. M. Jose. Using Emotion to Diversify Document Rankings. In *Advances in Information Retrieval Theory (ICTIR'11)*, pages 337–341, 2011. 63

H. Muller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer, 2010. 60

J. Mulvihill and E. H. Brenner. Ranking Boolean Search Output. *American Documentation*, 19(2):204–205, 1968. 38

S.A. Munson, D.X. Zhou, and P. Resnick. Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators. In *Third International AAAI Conference on Weblogs and Social Media*, 2009. 60

L. A. F. Park, K. Ramamohanarao, and M. Palaniswami. A Novel Document Retrieval Method using the Discrete Wavelet Transform. *ACM Transactions on Information Systems (TOIS)*, 23(3):267–298, 2005. 109

B. Piwowarski, I. Frommholz, M. Lalmas, and C. J. van Rijsbergen. What can Quantum Theory Bring to Information Retrieval? In *Proceedings 19th International Conference on Information and Knowledge Management (CIKM'10)*, October 2010. 110, 172, 173

J. M. Ponte and W. B. Croft. A language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 275–281. ACM, 1998. 2, 18, 36, 57

F. Radlinski and S. Dumais. Improving Personalized Web Search using Result Diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 691–692. ACM, 2006. 32, 60

F. Radlinski, P.N. Bennett, B. Carterette, and T. Joachims. Redundancy, Diversity and Interdependent Document Relevance. *ACM SIGIR Forum*, 43:46–52, December 2009. 23, 30, 56

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 43–52, New York, NY, USA, 2008. ACM. 1

S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977. 1, 2, 36, 38, 39, 40, 128

S. E. Robertson. Okapi at TREC-4. In *Proceeding of the Fourth Text REtrieval Conference (TREC-4)*, 1992. 142

S. E. Robertson and K. Sparck-Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. 2, 15, 17, 19, 36

S. E. Robertson and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'94)*, pages 232–241. ACM, 1994. 15, 18, 19, 57

S. E. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, April 2009. 18

S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic Models of Indexing and Searching. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR'80)*, pages 35–56. Butterworth & Co., 1981. 17

S. E. Robertson, M. Maron, and W. Cooper. Probability of Relevance: A Unification of two Competing Models for Document Retrieval. *Information technology: research and development*, pages 1–21, 1982. 36

S.E. Robertson and N.J. Belkin. Ranking in Principle. *Journal of Documentation*, 34(2):93 – 100, June 1978. 35, 36

T. Sakai and R. Song. Evaluating Diversified Search Results Using per-intent Graded Relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information (SIGIR'11)*, pages 1043–1052. ACM, 2011. 27

T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.Y. Lin. Simple Evaluation Metrics for Diversified Search Results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)*, 2010. 26, 27

G. Salton. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971. 14

G. Salton and C. Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information processing & management*, 24(5):513–523, 1988. 14

G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986. 14

G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18:613–620, November 1975. 5, 14

G. Salton, E. A. Fox, and H. Wu. Extended Boolean Information Retrieval. *Communications of the ACM*, 26:1022–1036, November 1983. 33

M. Sanderson. Ambiguous Queries: Test Collections Need More Sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'08)*, pages 499–506. ACM, 2008. 56

Rodrygo Santos, Craig Macdonald, and Iadh Ounis. On the role of novelty for search result diversification. *Information Retrieval*, pages 1–25, 2012. 63, 156

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, pages 881–890, 2010. 29

T. Saracevic. The Stratified Model of Information Retrieval Interaction: Extension and Applications. In *Proceedings of the Annual Meeting - American Sociery for Information Science*, volume 34, pages 313–327, 1997. 68

L. Schamber, M. Eisenberg, and M. S. Nilan. A Re-examination of Relevance: toward a Dynamic, Situational Definition. *Information processing & management*, 26(6):755–776, 1990. 35

F. Scholer, A. Turpin, and M. Sanderson. Quantifying Test Collection Quality based on the Consistency of Relevance Judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information (SIGIR'11)*, pages 1063–1072. ACM, 2011. 149

U. Sinha, C. Couteau, T. Jennewein, R. Laflamme, and G. Weihs. Ruling out Multi-order Interference in Quantum Mechanics. *Science*, 329(5990):418, 2010. 96

A. F. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval (MIR'06)*, pages 321–330. ACM, 2006. 67

I. Soboroff and D. Harman. Novelty Detection: the TREC Experience. In *roceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, pages 105–112, 2005. 174

K. Spärck-Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of documentation*, 28(1):11–21, 1993. 57

K. Spärck-Jones and C. J. van Rijsbergen. Report on the Need for and Provision of an "Ideal" Judgements Retrieval Test Collection. Technical Report 5266, Computer Laboratory, University of Cambridge, 1975. 20

K. Spärck-Jones, S. Walker, and S.E. Robertson. A Probabilistic Model of Information Retrieval: Development and Status. *Information Processing and Management: an International Journal*, 1998. 5

K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous Requests: Implications for Retrieval Tests, Systems and Theories. *SIGIR Forum*, 41(2): 8–17, 2007. 35, 56

A. Spink. Study of Interactive Feedback during Mediated Information Retrieval. *Journal of the American Society for Information Science*, 48(5):382–394, 1997. 68

A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the Web: The Public and their Queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001. 55

K. H. Stirling. *The Effect of Document Ranking on Retrieval System Performance: A Search for an Optimal Ranking Rule.* PhD thesis, University of California, 1977. 2, 28, 35, 44, 73

A. Turpin and F. Scholer. User Performance versus Precision Measures for Simple Search Tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 11–18. ACM, 2006. 21

C. J. van Rijsbergen. *Information Retrieval.* Butterworth, 2nd edition, 1979. 1, 20, 123

C. J. van Rijsbergen. *A Non-classical Logic for Information Retrieval*, volume Readings in Information Retrieval, pages 268–272. Morgan Kaufmann Publishers Inc, 1997. 5

C. J. v. van Rijsbergen. *The Geometry of Information Retrieval.* Cambridge University Press, August 2004. 3, 5, 7, 105, 108, 109, 111, 155, 172, 182

H. R. Varian. Economics and Search. *SIGIR Forum*, 33:1–5, September 1999. 53, 54, 58

V. V. Vazirani. *Approximation Algorithms.* Springer, 2001. 28

J. Verhoeff, W. Goffman, and J. Belzer. Inefficiency of the Use of Boolean Functions for Information Retrieval Systems. *Communications of the ACM*, 4:557–558, December 1961. 38

E. M. Voorhees and D. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, September 2005. 20, 21

Ellen M. Voorhees. TREC: Improving Information Access through Evaluation. *Bulletin of the American Society for Information Science and Technology*, 32 (1):16–21, 2005. 20, 128

H. Walach and N. von Stillfried. Generalised Quantum Theory – Basic Idea and General Intuition: A Background Story and Overview. *Axiomathes*, pages 1–25, 2011. 103

J. Wang. Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval (ECIR'09)*, volume 5478 of *Lecture Notes in Computer Science*, pages 4–16. Springer, 2009. 58, 64, 67

J. Wang and J. Zhu. Portfolio Theory of Information Retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09)*, pages 115–122. ACM, 2009. 2, 58, 64, 66, 67, 73, 130, 138, 149, 158

J. Wang, D. Song, P. Zhang, H. Yuexian, and P. Bruza. Explanation of Relevance Judgement Discrepancy with Quantum Interference. In *Proceedings Quantum Interaction 2010*, pages 117–124, 2010. 103, 114

M. L. Weitzman. Optimal Search for the Best Alternative. *Econometrica*, 47(3): 641–654, 1979. 54

D. Widdows. *Geometry and Meaning*. Center for the Study of Language and Information/SRI, 2004. 5

Y. C. Xu and Z. Chen. Relevance Judgment: What do Information Users Consider beyond Topicality? *Journal of the American Society for Information Science*, 57(7):961–973, 2006. 35, 101

C. X. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'01)*, pages 334–342. ACM, 2001. 19

C. X. Zhai and J. Lafferty. A Risk Minimization Framework for Information Retrieval. *Information Processing & Management*, 42:31–55, January 2006. 2, 63

C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, pages 10–17. ACM, 2003. 24, 25, 67, 73, 113, 122, 140

J. Zhu, J. Wang, I. J. Cox, and M. J. Taylor. Risky Business: Modeling and Exploiting Uncertainty in Information Retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09)*, pages 99–106. ACM, 2009. 58, 67

G. Zuccon. An Analogy between the Double Slit Experiment and Document Ranking. In *The 3rd IRSG Symposium: Future Directions in Information Access 2009 (FDIA 2009)*, 2009. 9

G. Zuccon and L. Azzopardi. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and C. J. van Rijsbergen, editors, *Advances in Information Retrieval (ECIR'10)*, volume 5993 of *Lecture Notes in Computer Science*, pages 357–369. Springer, 2010a. 10

G. Zuccon and L. Azzopardi. Developing the Quantum Probability Ranking Principle. In M. Melucci, S. Mizzaro, and G. Pasi, editors, *Proceedings of* *the First Italian Information Retrieval Workshop (IIR'10)*, volume 560, pages 21–22. CEUR-WS.org, 2010b. 10

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. A Formalization of Logical Imaging for Information Retrieval Using Quantum Theory. In *DEXA Workshop on Textual Information Retrieval (TIR'08)*, pages 3–8. IEEE Computer Society, September 2008. 5, 9

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Semantic Spaces: Measuring the Distance between Different Subspaces. In P. Bruza, D. Sofge, W. Lawless, C. J. van Rijsbergen, and M. Klusch, editors, *Proceedings of the Third International Quantum Interaction Symposium (QI'2009)*, volume 5494 of *Lecture Notes in Computer Science*, pages 225–236. Springer, 2009a. 9, 173

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Revisiting Logical Imaging for Information Retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09)*, pages 766–767. ACM, 2009b. 9

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. The Quantum Probability Ranking Principle for Information Retrieval. In *Advances in Information Retrieval Theory (ICTIR'09)*, volume 5766 of *Lecture Notes in Computer Science*, pages 232–240. Springer, 2009c. 9

G. Zuccon, L. Azzopardi, C. Hauff, and C. J. van Rijsbergen. Estimating Interference in the QPRP for Subtopic Retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10)*, pages 741–742. ACM, 2010a. 10

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Has Portfolio Theory got any Principles? In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10)*, pages 755–756. ACM, 2010b. 10

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. The Interactive PRP for Diversifying Document Rankings. In *Proceeding of the 34th international ACM*

SIGIR conference on Research and development in information retrieval (SIGIR'11), pages 1227–1228. ACM, 2011a. 10, 68, 131

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Back to the Roots: Mean-Variance Analysis of Relevance Estimations. In *Advances in Information Retrieval (ECIR'11)*, volume 6611 of *Lecture Notes in Computer Science*, pages 716–720. Springer, 2011b. 10, 58, 68, 130

G. Zuccon, L. Azzopardi, and C. J. Van Rijsbergen. An Analysis of Ranking Principles and Retrieval Strategies. In *Proceedings of the Third international conference on Advances in information retrieval theory (ICTIR'11)*, pages 151–163. Springer, 2011c. 11

G. Zuccon, B. Piwowarski, and L. Azzopardi. On the Use of Complex Numbers in Quantum Models for Information Retrieval. In *Proceedings of the Third international conference on Advances in information retrieval theory (ICTIR'11)*, pages 346–350. Springer, 2011d. 10

G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k Retrieval using Facility Location Analysis. In *Advances in Information Retrieval (ECIR'12)*, volume 7224 of Lecture Notes in Computer Science, pages 305–316. Springer, 2012. 11, 30, 34

Index