Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework

Ingo Frommholz
Department of Computing Science
University of Glasgow
ingo@dcs.gla.ac.uk

Birger Larsen
Royal School of Library and Information Science
Copenhagen, Denmark
blar@db.dk

Benjamin Piwowarski
Department of Computing Science
University of Glasgow
bpiwowar@dcs.gla.ac.uk

Mounia Lalmas
Department of Computing Science
University of Glasgow
mounia@dcs.gla.ac.uk

Peter Ingwersen
Royal School of Library and Information Science
Copenhagen, Denmark
pi@db.dk

Keith van Rijsbergen
Department of Computing Science
University of Glasgow
keith@dcs.gla.ac.uk

ABSTRACT

The relevance of a document has many facets, going beyond the usual topical one, which have to be considered to satisfy a user's information need. Multiple representations of documents, like user-given reviews or the actual document content, can give evidence towards certain facets of relevance. In this respect polyrepresentation of documents, where such evidence is combined, is a crucial concept to estimate the relevance of a document. In this paper, we discuss how a geometrical retrieval framework inspired by quantum mechanics can be extended to support polyrepresentation. We show by example how different representations of a document can be modelled in a Hilbert space, similar to physical systems known from quantum mechanics. We further illustrate how these representations are combined by means of the tensor product to support polyrepresentation, and discuss the case that representations of documents are not independent from a user point of view. Besides giving a principled framework for polyrepresentation, the potential of this approach is to capture and formalise the complex interdependent relationships that the different representations can have between each other.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.
Copyright 2010 ACM 978-1-4503-0247-0/10/08 ...$10.00.

Keywords

Quantum-inspired model, Polyrepresentation

1. INTRODUCTION

When users seek for information, their decision about whether a document is useful usually depends on more dimensions than the usual IR system's estimation of topical relevance [25]. Besides its content, this decision involves different contextual aspects of a document, like for instance, writing style, understandability, authority, which are concerned with the non-topical usefulness of a document [4].

We use an example of a book store scenario to illustrate this. Consider a user whose goal is to find "good introductions to quantum mechanics". The user might visit an online bookstore, where it is common to find different representations of documents, for example, abstracts, full texts, user-given tags, editorial and user-given reviews, user-given ratings and structured bibliographical metadata (e.g. author, title, number of pages and publication date). These different representations potentially answer different facets of the information need. Title, abstract, tags or full text can provide evidence about the topicality of a book, but reviews may also be used to determine if the book is about "quantum mechanics". Another facet of the information need is that the user seeks for an "introduction" to the topic; we may get hints about this from the abstract, or the title, but to a lesser degree from the full text ("introduction" appears in many texts as a chapter heading and does not mean than the document is an introduction to a given field). Finally, the evidence about the quality of the book ("good") may come from ratings and reviews. These different representations may interplay and have different importance during search. For example, the user may judge one book relevant based on given representations (for instance the author is known to be an authority in the field), and the relevance of another book based on different representations (e.g., the document has high ratings).

The above example underlines the fact that by having and considering different representations of a document – from different contexts – we are able to address different aspects of the usefulness of a document. Naturally, we want to combine the given evidence to get a more accurate estimation

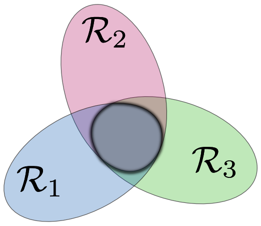


Figure 1: Different representations and cognitive overlaps. The intersection $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$ defines the *total cognitive overlap*; all other intersections between representations establish a *partial cognitive overlap*.

of the usefulness of a document. One strategy of doing so is to apply the *principle of polyrepresentation* of documents, which aims to generate and exploit the *cognitive overlap* between different representations of documents – those documents that lie in this overlap are assumed to be relevant given a user's information need [9, 10].

We regard polyrepresentation as a key principle to satisfy a user's information need. This cognitive viewpoint allows all components in IR to be considered in a single coherent and consistent theoretical framework. It is thus holistic in its intention and has a strong focus on exploiting different contexts in IR. The principle of polyrepresentation has, however, so far not been developed to a point where a strong and complete formal mathematical IR framework encompasses the principle in its entirety. One reason might be its inherent complexity with its many factors and interdependencies, which goes beyond what typical IR models cover. The goal of the paper is to develop a mathematical formalism that takes a much larger range of phenomena into consideration than current IR models. We do so by extending to polyrepresentation the geometrical IR framework presented in [21]. This framework is based on the idea of exploiting the quantum mechanics formalism for information retrieval as was suggested in [28]. Besides investigating how such a framework can be extended to support polyrepresentation, we will also show how it is possible to model interdependencies between the representations. Due to the relationship between geometry and probability theory outlined in [28], the proposed framework is also probabilistic by nature.

The remainder of the paper is structured as follows. In the next section we first give an overview of the principle of polyrepresentation and its implied requirements. In Section 3 we briefly introduce the quantum-inspired IR framework that we extend in the two next sections. In Section 4 we give examples how individual representations can be modeled. To support polyrepresentation, the representations need to be combined by creating the cogntive overlap known from the polyrepresentation principle, also considering dependencies between representations. These aspects are discussed in Section 5. Subsequently, we review related works and conclude.

To illustrate various concepts in this paper, our examples in this paper will all be inspired by the task of searching in a book store website.

2. POLYREPRESENTATION

In the principle of polyrepresentation all components in Information Retrieval (IR) are regarded as being the result of cognitive transformations of the knowledge structures of the involved actors [9, 10]. That is, documents[1] are seen as representations of their authors' ideas, retrieval models and systems as representations of their designers' ideas, as well as information behaviour including issued queries and interaction with IR systems as representations of users' needs etc. Also later interpretations of a given document by other authors, e.g., in reviews, citations, twitter feeds etc, are regarded as representations but with different cognitive origins. Figure 1 illustrates how the intersection of different representations maybe seen to create so-called *cognitive overlaps*. The principle of polyrepresentation hypothesises that documents retrieved by representations that are more different from each other in cognitive origin and time have higher probability of being relevant. Thus as an example, if Figure 1 is taken to illustrate sets of documents retrieved using three cognitively different representations, we would expect the documents in the total cognitive overlap between $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$ to have a higher probability of being relevant than those in partial overlaps, e.g., between $\mathcal{R}_2$ and $\mathcal{R}_3$, or those outside the overlaps.

As mentioned above, so far there is no formal model which covers the principle of polyrepresentation in its entirety. In the following we analyse some of the main features a formal model of polyrepresentation should encompass based on the extensive account in [9]. Basically, we deal with the polyrepresentation of documents in this work; some of the addressed aspects are:

- Flexible combination of representations, e.g. fusion of representations (Boolean and weighted cognitive overlaps); diffusion of representations and selection of representations given a certain context;
- Temporal aspects and dynamic changes over time in relation to representations, for instance changes/drifts in the user's information need and new interpretations of documents arising subsequently from other agents (e.g., categorisation/indexing/tagging, in-links and anchor text, social bookmarking, reviews, ratings, annotations, etc.);
- Different semantic document levels, e.g., subdocument level (akin to element/XML retrieval, logical document structure) and groups/clusters of documents;
- Although the representations may be kept separate and processed independently, it is a fact that they are interdependent and contextual to each other, which is a further aspect to consider.

The flexible combination of representations, also taking into account possible interdependencies, and temporal aspects regarding new interpretations are described in Section 5. The discussion of our basic framework and the examples of it, which can be found in sections 3 and 4, respectively, covers dynamic information need changes as well as how different semantic document levels can be addressed. As a more technical issue, a polyrepresentation framework also needs to handle several heterogeneous textual (e.g., document content, reviews and annotations) and non-textual

[1] By documents we mean "physical (digital) entities in a variety of media", that is information objects including text documents as defined in [10]

document representations (like ratings). Examples of how such representations can be modelled are given in Section 4.

The selection of features represents the ones we regard as the most important in working with polyrepresentation on the document side. Apart from these features, we can also identify others, which we do not address directly in this work, including different levels of representation of the users' information need and exploitation of the user context, and keeping track of the representation origin (origin/actor, time). A deeper discussion of these features and their integration into the framework presented here are subject to future work. Such an integration is in our view possible.

In this paper, we do not deal with the issue to relate a facet of an information need to one or more different representations. For example, for a user typing "good introductions to quantum mechanics", we would need to distinguish three different facets ("good", "introductions" and "quantum mechanics") and map them appropriately to the rating and topical spaces (e.g., title/content and comments). We therefore assume that, through either an interface or a sophisticated algorithm, we are able to analyse and assign the user's request appropriately.

3. QUANTUM-INSPIRED GEOMETRICAL IR FRAMEWORK

We describe the geometrical framework proposed in [21], upon which our work is based. In this framework, the IR system is initially uncertain about the user's information need (IN), and two dynamics modify the system view of the user's IN. Firstly, when the user interacts with the system, for instance by typing, refining a query or browsing, the *system view* of the user's IN becomes more and more specific, i.e., the uncertainty of the system about the IN is reduced. We refer to this dynamic process as **(D1)**. Secondly, the IN may change from a *user point of view*; if a user gathers more knowledge in the information seeking process, the IN may become more specific or may drift, for instance when the user's perceived information need changes. We refer to this process as **(D2)**. Supporting this process satisfies the requirement on a formal polyrepresentation model for addressing changing information needs, as stated in the previous section.

The postulate made in [21] is that this interaction can be captured using the connection between probabilities and geometry present in the quantum physics formalism, whose connection with IR has been discussed in [28, ch. 6]. This work was applied to the topical relevance facet of the information need [19]. From the next section onwards, we show how we extend it to cover different facets of relevance.

3.1 Information Need Space

The assumption underlying the framework of [21] is that there exists an Information Need space where any user's "pure" IN can be represented from an IR system point of view. "Pure" in this sense means that the user's IN is completely defined, i.e., if the IR system knew the user's pure state, then it would exactly know what the user is looking for, and return the documents that are relevant to that user's IN.

This view is motivated from quantum mechanics that postulates that associated with each physical system is a space, the *state space*. Formally, this state space is a Hilbert space $\mathcal{H}$ (a vector space with an inner product). Following the Dirac notation used in quantum mechanics, we will denote a vector as $|\varphi\rangle$. A physical system (e.g., a photon) whose state is known is completely described by a *state vector* $|\varphi\rangle$, which is a unit vector (i.e. its norm $\|\varphi\|$ is 1) in the space $\mathcal{H}$.

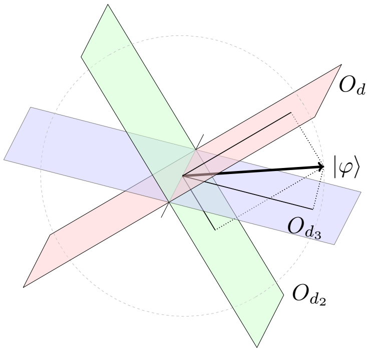


Figure 2: Projections onto subspaces in a 3-dimensional vector space with a state vector $|\varphi\rangle$

By analogy, in the IR framework proposed by [21], the state corresponds to the IN of a user, expressed in a state space which we call the *IN space*. In the following, we first describe how to compute quantities of interest for an IR system (e.g. the probability of relevance), and then show how to update the state given some observation (i.e., interaction with the user).

In the quantum formalism, any event is defined by a subspace. Hence, in the IN space, for each document d, we can define a subspace O_d corresponding to the event "the document d is relevant". If we let $|\varphi\rangle$ be the user's IN state, the probability $\Pr(R|d,\varphi)$ of the document d being relevant to the user's IN $|\varphi\rangle$ is defined as the square of the length of the projection of the vector $|\varphi\rangle$ onto the subspace O_d [28], which adheres to the definition of probabilities in quantum mechanics.

For example, let us assume we have three documents d_1, d_2 and d_3. In Figure 2 we can see the corresponding document subspaces, O_{d_1}, O_{d_2} and O_{d_3}, and the state vector $|\varphi\rangle$ representing the user's pure information need. In this state, the resulting ranking would be d_3, d_1, d_2 as the length of the projection of $|\varphi\rangle$ onto O_{d_3} is greater than onto O_{d_1}, which is greater than the length of the projection onto O_{d_2}. The actual probability of relevance for, say, d_3 is the square of the length of the projection of $|\varphi\rangle$ onto O_{d_3}.

Returning to Section 2, we discussed the need to support different semantic document levels for polyrepresentation. This can be achieved by utilising the above description of documents (respectively their relevance) as subspaces. For instance, refining a document into further subspaces is a means to reflect the logical document structure [22]. A paragraph could be a low dimensional subspace of the IN space; the subspace associated with a section the paragraph is found in would then contain the paragraph subspace. The union of document subspaces can be used to represent document groups or clusters.

3.2 Uncertain States and User Interaction

At the beginning of the search, we cannot assume that the IR system knows exactly about the user's IN. To capture this uncertainty, we introduce a further probability distribution, the probability p_i that the system is in a state $|\varphi_i\rangle$. By doing so we allow for the IN to be in one of a set of different states with a given certain probability. In this case, we say that the user's IN is in a *mixed state*. This can be formally expressed by defining an *ensemble* $S = \{(p_i, |\varphi_i\rangle)\}$ of states $|\varphi_i\rangle$ (where each of them represents a pure IN) and their corresponding probability p_i, with $\sum_i p_i = 1$. The IR system assumes the user's IN is $|\varphi_i\rangle$ with probability p_i . When the IR system knows the user's information need with certainty, then the ensemble is reduced to only one state $|\varphi\rangle$, which in this case is called a *pure state*.

Given an ensemble S, we can compute the probability of any event, like the relevance of a document, by applying the law of total probability. For example, say that the system assumes that it is in state $|\varphi_1\rangle$ with probability p_1 or in state $|\varphi_2\rangle$ with probability p_2, so the mixed state is described by $S_1 = \{(p_1, |\varphi_1\rangle), (p_2, |\varphi_2\rangle)\}$. Then the probability that a document d is relevant given the current state is $\Pr(R|d, S_1) = p_1 \cdot \Pr(R|d, \varphi_1) + p_2 \cdot \Pr(R|d, \varphi_2)$, where $\Pr(R|d, \varphi_i)$ is again the square of the length of the projection of $|\varphi_i\rangle$ onto the subspace O_d. In general, given a mixed state described by the ensemble $S = \{(p_i, |\varphi_i\rangle)\}$, we have

$$\Pr(R|d, S) = \sum_i p_i \cdot \Pr(R|d, \varphi_i). \quad (1)$$

As said, a mixed state reflects the system's uncertainty about the user's IN if this is underspecified, which is often the case in an information seeking scenario [17]. Initially, before any user interaction has taken place, the system state is a mixture of all possible INs with a probability that depends for instance on the popularity of an IN. Upon user interaction, the system state may become more specific or react to a drift in the information need.

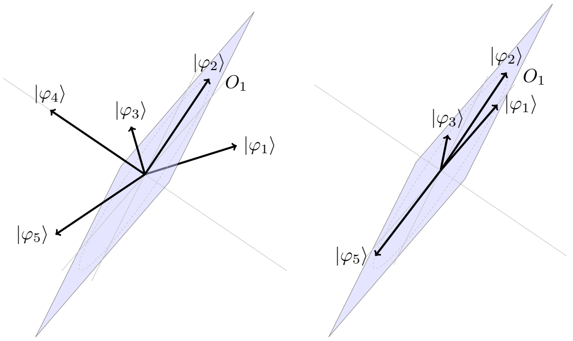
User interaction can be used to reduce this uncertainty, using another well-known concept from quantum mechanics, *measurement*, which is comparable to probabilistic conditionalisation. Measurement uses again the subspace that describes an observed event (e.g., the user has judged this document as relevant) and acts upon an ensemble S in a geometric way. Without entering into technical details, it involves projecting and renormalising each vector $|\varphi\rangle$ into the subspace defining the event, and updating the different probabilities p_i. In practice, it means that after measurement, all the vectors in the ensemble belong to the subspace that defines the observed event.

3.3 Example

Let us illustrate how measurement supports (D1) and (D2) by considering an ensemble of five possible states and an event O_1 as depicted in Figure 3a.

The state vectors of the ensemble are freely distributed in the 3-dimensional space. After measurement, the ensemble is projected onto the 2-dimensional subspace O_1, which is shown in Figure 3b. In this measurement, we can illustrate the two different dynamics. (D1) is supported because the IR system is now "less uncertain" about the user's IN due to the fact that the ensemble is now bound to the 2-dimensional plane. For example, the vector $|\varphi_4\rangle$ has been removed from the ensemble, while $|\varphi_2\rangle$ that was belonging to the subspace O_1 has been kept. (D2) is supported since $|\varphi_1\rangle$, $|\varphi_3\rangle$ and $|\varphi_5\rangle$ have been changed through projection.

The framework defined here does not make any assumption on the geometry of the underlying vector space, nor on how an "information need" is actually defined. Provided that the hypothesis of the framework are correct, that is that there exists a IN space where all the possible INs can be defined and that it is possible to define subspaces for any event of interest, this framework offers the possibility to describe geometrically the whole IR search process. In the following, we show that within this framework it is possible support a polyrepresentation of documents by modelling different aspects of relevance. We show how the quantum framework can relate in a complex manner these different aspects through a so-called tensor product of Hilbert spaces.



(a) Before measurement (b) After measurement

Figure 3: Effects of measurement. Vectors orthogonal to the subspace are eliminated, vectors already in the subspace remain unchanged, the rest is projected onto the subspace and renormalised.

4. SINGLE REPRESENTATIONS

We discuss some examples of spaces that reflect different aspects of relevance, or said otherwise, that deal with different representations of a document. As our goal is to integrate these different representations, each of them should satisfy our hypotheses about the IN space defined in Section 3. To this end, we associate with each distinct representation a *representation space*, which is a Hilbert space where a state reflects the component of the IN related to the representation.

In order to be compatible with the basic interactive framework introduced in the previous section, each space should be able to support the dynamic processes (D1) and (D2) by means of measurement. It should also allow for the calculation of the probability of relevance regarding the corresponding representation, applying Equation 1 and the projection of state vectors onto document subspaces. The geometrical description of representations can make use of the flexibility that comes with the quantum formalism (for instance by using non-orthogonality) and its relationship to probability theory.

Defining a distinct space for each representation (instead of one that covers all representations at once) is not only simpler, since we can focus on the peculiarities of each representation, but it is also necessary, since with polyrepresentation we want to be able to consider flexible combinations of representations, as stated in Section 2. Another advantage is that as we can compute the probability of relevance given any representation separately with Equation 1, we can use any strategy based on polyrepresentation to combine the different probabilities.

In this section, we present examples of spaces that are useful for an online book store and fulfill the above requirements. As each potential application comes with different representations of various kind, we restrict ourselves to examples. Therefore, it is the goal of this section to give the reader an impression of the potential of the quantum formalism, not to define an exhaustive set of representation spaces.

In a book store scenario we deal with different types of representations, which basically can be textual (e.g., abstract, title or comments) or non-textual (e.g., bibliographic metadata or ratings). In this section, we discuss representation spaces for both types of representations. In the case of non-textual information, we chose two important representations of different types, namely the author and the rating.

4.1 Textual Representations

The most common representations of a document in IR are textual, and given for instance by the full text of a document, the document title, annotations or comments attached to the document, but also by user-given tags. In IR, a generic way of representing textual content is in form of vectors which lie in a space where each dimension corresponds to a term; the term vector may then contain for instance the associated term weights. However, recent work indicates that it is beneficial to represent textual content by more than just one vector [3]. Furthermore, a document may be relevant to more than one information need [23], which also suggests a more fine-grained representation than with just one vector, potentially reflecting the logical document structure. The idea therefore is to represent textual content by more than one vector, i.e., a subspace.

A Hilbert space representation for documents and queries based on the term space was proposed and evaluated in [19, 20], where the term space roughly corresponds to the topical representation space of the user's IN. We briefly outline this representation here.

The (simplifying) assumption is that each document is composed of a set of excerpts, each one answering a specific (topical) IN. This is depicted in Figure 4 where a document is covered up by the excerpts. Each excerpt corresponds to a "pure" IN, it is thus possible to associate each of them with a unit vector $|u_i\rangle$ in the space. The next hypothesis made is that the relevance of a document can be represented as a subspace that spans the set of vectors $\{|u_i\rangle\}$, or said otherwise by the minimal subspace that contains all of these vectors. By doing so, it ensures that if a user is in one of the states $|u_i\rangle$, then the probability that the document is relevant is 1 (since the projection of $|u_i\rangle$ onto the subspace is $|u_i\rangle$ itself).

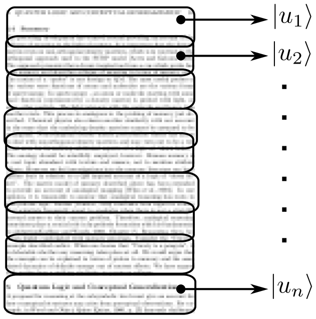


Figure 4: Extraction of IN vectors from a text

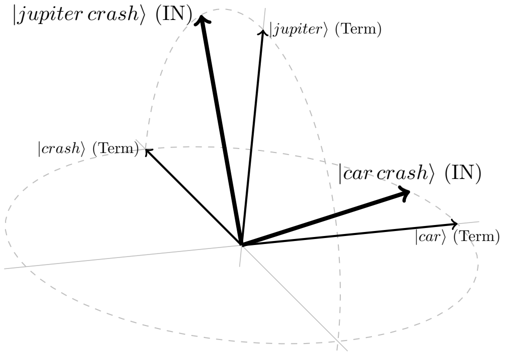


Figure 5: Textual representation – IN space superimposed on a term space

This type of construction, where we define the subspace associated with the relevance of a document as the minimal subspace containing all the INs for which this document is relevant, is a general one and we will apply it to the different representation spaces we describe in the next sections.

In the case of topical relevance, the topical IN space can be approximated by a term space where each term corresponds to one dimension [19, 20], as in the vector space model (see, e.g., [24]) widely used in IR. An IN is thus described as a set of weights, one for each term. More precisely, each vector $|u_i\rangle$ is built from the terms of the corresponding excerpt, i.e. has non null components for the terms within the excerpt.

Figure 5 illustrates the set up of this topical/term space. Here, two information needs are shown, "Jupiter crash" (describing for instance a recent comet crash on Jupiter) and "car crash". The terms "jupiter" and "crash" make up the former IN, while the latter is composed of the terms "car" and "crash". In this example, the IN vectors are not necessarily orthogonal, which motivates the use of a quantum probability framework.

A more elaborate discussion of the text/topical representation can be found in [19, 20].

4.2 Non-Textual Representations

We give examples for two important non-textual representations in a book store scenario, namely authors and ratings.

4.2.1 Author Space

Our first example deals with users searching for a book from a specific author. In this section, we propose several possibilities of increasing complexities. To simplify our discussion, we assume the user is looking for *one* specific author, not several.

A first possibility is to associate with each author a distinct dimension of the author space, where the relevance of a document is the subspace spanned by the different vectors of the book authors. In that case, a pure IN corresponds to one of the author vectors, and only documents authored by this specific person are relevant.

With this representation, we are using a standard prob-

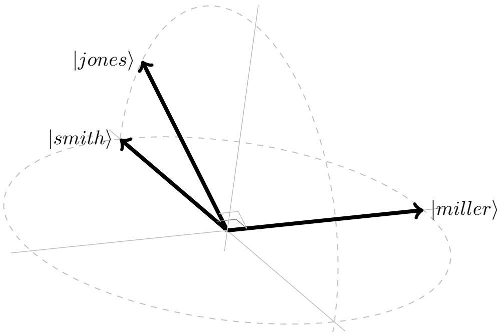


Figure 6: Author IN vectors

ability framework. However, it can be argued that a user interested in a document by a specific author may also be interested in documents by another, related author. This may for instance be the case when two authors have a co-author relationship or work in the same field. This could also be based on the style of the author (in the case of novels). The information about how close two authors should be could be constructed by using the content of authors' books, but this could be also extracted from the purchase behaviour of a book store's clients – people who bought books from an author A might often also buy books from an author B.

To allow such dependencies between authors, a second possibility is to allow vectors to be non orthogonal. This is depicted in Figure 6, where three different author vectors are shown; we assume that "Jones" and "Smith" share many characteristics and should hence be represented as non-orthogonal vectors. Let us imagine that a book has the authors Smith and Miller. This book would be (author) relevant for an IN $|smith\rangle$ (probability of 1), but differently to the first possibility, it would be also relevant to an author-IN $|jones\rangle$, although with a lesser probability. This probability can be tuned since it depends on the angle between the $|jones\rangle$ and $|smith\rangle$ vectors: The smaller the angle, the higher the probability.

The relationship between Jones and Smith might be more complex than the one described above. Say, for example, Jones is interested in probabilistic logics and also in interactive retrieval, whereas Smith' interests are probabilistic logics on the one hand and theoretic models in information retrieval on the other hand. Smith may be a former PhD student of Jones with probabilistic logics as PhD topic, so that both share many publications on this topic, and both may have published further articles on probabilistic logics, although not as co-authors. So Smith and Jones are very related when it comes to probabilistic logics, and users looking for documents about probabilistic logics by Jones may likely be interested in the further work performed by Smith on that topic. On the other hand, when a user seeks for documents about interactive retrieval by Jones, Smith' publications are likely to be not relevant.

This motivates our third and last possibility of representation, where an author is represented as a subspace. More precisely, an author can be associated with a set of author INs. With our Jones/Smith example, a possible representation is depicted in Figure 7. In this figure, a book authored by Jones can be either an answer to a user looking for the logic writings of Jones or his interactive IR writings. In the latter case (interactive IR), a book written by Smith would have a zero probability of being relevant, whereas in the former one (logics) it would have a non zero probability.

4.2.2 Rating Space

In e-stores, ratings reflect the user's opinion about the quality of an object. They are often given on a rating scale, ranging for instance from zero star ("very bad") to 5 stars ("very good"). It is common to compute an average value of the ratings and categorise the average onto the given rating scale or a more finer grained discrete one (for example, the average could be "3 1/2 stars").

We assume that users generally prefer higher-rated documents, and that the ratings can be mapped on a set of ordinal scaled labels $\{l_i\}$. For example, in an online store it may be possible to give, say, 0 to 2 stars, where 0 stars are mapped to the label *bad*, 1 star is mapped to the label *average* and 2 stars means *good*. We can then establish an order on these labels: *good* > *average* > *bad*.

As in the case of independent authors, we can set up a Hilbert space for such a representation. In the above example, this would result in a 3-dimensional vector space with the orthogonal vectors $|good\rangle$, $|average\rangle$ and $|bad\rangle$, where each vector corresponds to the minimum rating the user wants for a book.

To reflect the order of labels, a "good" book would thus be represented as the whole 3-dimensional subspace (as it is relevant to any rating-IN), an "average" book would be the 2-dimensional subspace spanned by $|average\rangle$ and $|bad\rangle$, and a "bad" book would be the 1-dimensional subspace corresponding to $|bad\rangle$. A pure rating-IN can then be interpreted as a threshold. For example, if the rating-IN is $|average\rangle$, this means that for the user average and good books are interesting; the $|average\rangle$ vector is contained in both the subspaces for "good" and "average" books.

As for authors, it might be interesting to make the different vectors non-orthogonal, since if the user rating-IN is $|good\rangle$, an average book is better than a bad book. In practice (not described here), it is possible to set a probability that a user is satisfied with an average/bad book given its rating-IN, and to compute the non-orthogonal $|good\rangle$, $|average\rangle$ and $|bad\rangle$ vectors in a three dimensional space that satisfy those probabilities.

5. COMBINING THE EVIDENCE

In the last section we gave examples of single representation spaces. In each of these spaces, a state can change according to the dynamics defined in Section 3, and it is

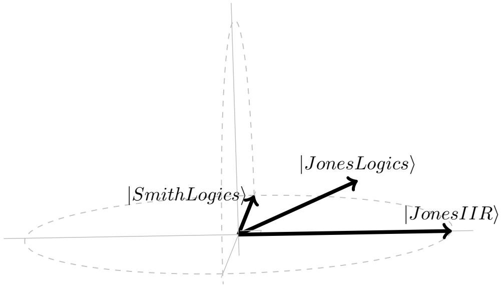


Figure 7: Author/Topic Space

possible to compute a probability of relevance. However, we have not yet described how to combine the evidence of the different representations into a single framework, which is central to every polyrepresentation approach.

At first, the underlying assumption is that the representations are independent. As the geometrical framework in Section 3 has a probabilistic interpretation given by Equation 1, we can describe the creation of the (total) cognitive overlap from these representations purely probabilistic. In order to support partial overlaps and weighted representations, a requirement stated in Section 2, we need to introduce an extra dimension to our representation spaces. All this is discussed in Section 5.1. Then, in Section 5.2, we show how we can go beyond this simple probabilistic model by exploiting further the quantum formalism, allowing us to drop the assumption that representations are independent.

5.1 Total and Partial Cognitive Overlaps

The total cognitive overlap introduced in Section 2 requires that the different representations of a relevant document should all be relevant to the user's IN.

If we suppose that the representations are independently influencing relevance, then we can apply the probabilistic interpretation of our framework; following Griffiths [8], we define the probability of a document to be relevant as the product of the probabilities of the document to be relevant *in each representation.* Formally, we write

$$\Pr(R|d) = \prod_i \Pr(R|d, S_i) \tag{2}$$

where $\Pr(R|d, S_i)$ is the probability of relevance in the i^{th} representation space and is computed with Equation 1.

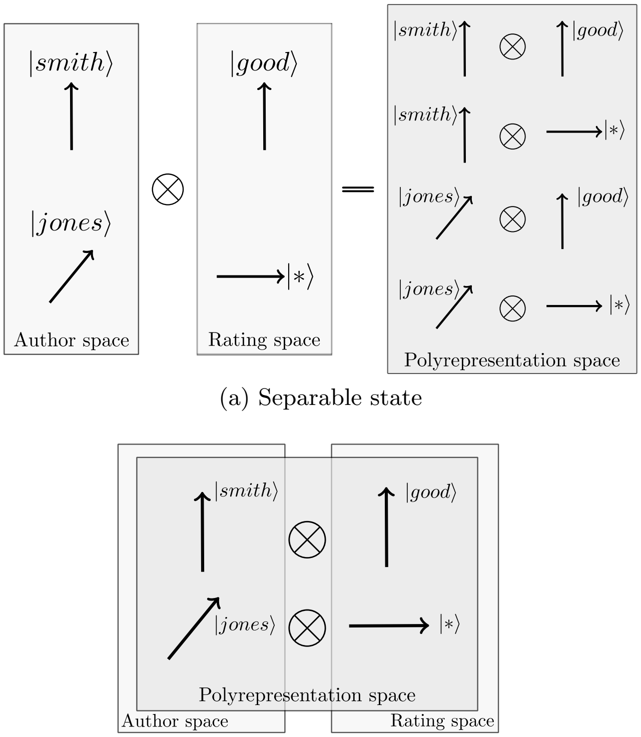
The creation of the total cognitive overlap has some shortcomings, as a document can easily have a zero probability of relevance. More precisely, it is sufficient that the document it not relevant in one representation to get a value of zero. This problem is stressed when we dynamically add new representation spaces. Furthermore, some kinds of representations seem to be more promising than others, and the representations should be carefully mixed and weighted [26].

Another important argument comes from the user side – a user may be interested in some representations more than in others, and may change her mind at another point in time. This suggests that the relative importance of the representation should be flexible enough so that it can evolve with time.

It is therefore desirable to loosen the total cognitive overlap requirement, and to ask for a level of coordination between representations: documents should also be retrieved even if they are not in the total cognitive overlap, but failing to be relevant w.r.t. certain representations.

We discuss a possible solution to introduce weights for representations into our framework. Our basic idea is that if the user does not care about a certain representation, all documents should be fully relevant in this representation. This can be achieved by adding an extra dimension, that we call the "*don't care" dimension,* in each representation space. The associated vector is denoted $|*\rangle$, and it is orthogonal to all other IN vectors and subspaces.

In order to introduce a weight, we can simply state that the user's IN is in a mixed state, and that with a probability α the user's IN is "don't care". We can then distribute the remaining $1 - \alpha$ probability among the possible states of the user's IN. If we take our author space for example, the system initially does not know what author the user is looking at and assumes a priori that the user does not consider authors at all with probability α. Having Jones, Smith and Miller as possible authors, the initial ensemble would be $\left\{\left(\frac{1-\alpha}{3}, |jones\rangle\right), \left(\frac{1-\alpha}{3}, |smith\rangle\right), \left(\frac{1-\alpha}{3}, |miller\rangle\right), (\alpha, |*\rangle)\right\}$.



(a) Separable state

(b) Non-separable state

Figure 8: Separable and non-separable state example

Regarding document subspaces, we require that each such subspace includes the "don't care" dimension, which makes each document "relevant" to the don't care "need". This has the following effect. If the "don't care" dimension has the weight 1 and the representation is therefore in the pure state $|*\rangle$, we would have $\Pr(R|d, S = |*\rangle) = 1$ for every document d, because $|*\rangle$ is always contained in the document subspace. According to Equation 2, this would mean that the representation is completely ignored.

Note that we can go back to a total cognitive overlap by giving a zero probability, i.e. $\alpha = 0$, to the "don't care" IN in all the representations. All other cases mean that we gradually relax the obligation of the document to be in overlaps containing the particular representation by assigning increasingly higher probabilities to the "don't care" dimension.

Since $|*\rangle$ is a valid state vector, through interaction it is possible (e.g., using projections as described in Section 3) to adapt automatically the weight α associated to this state. It is beyond the scope of this paper to explain how.

5.2 Interdependent Representations

So far we assumed that the different representations are independent, but from a user point of view, this assumption does not necessarily hold. For example, a user might think that documents written by Jones are always of high quality and therefore does not care about how they were rated. At the same time, the user does not know the author Smith, and thus tends to rely more on the ratings. So in one case the ratings are considered, while in the other they are not, showing that the author and rating representations are not independent of each other from the user's perspective. This is of course just one example of possible relationships between representations as they can occur in a book store scenario.

Before we continue the discussion, we need to briefly introduce another important geometrical concept, namely tensor products and tensor spaces. We refer the reader to [18, ch. 2] for a more in-depth discussion. The *tensor product* (denoted $\otimes$) is a way to combine different Hilbert spaces into a (larger) Hilbert space. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are two Hilbert spaces of respective dimensions n and m, the *tensor space* $\mathcal{H}_1 \otimes \mathcal{H}_2$ is an $n \cdot m$-dimensional Hilbert space. If $|\lambda_1\rangle$ is a vector in $\mathcal{H}_1$ and $|\lambda_2\rangle$ is a vector in $\mathcal{H}_2$, then $|\lambda_1\rangle \otimes |\lambda_2\rangle$ is a vector in $\mathcal{H}_1 \otimes \mathcal{H}_2$. Furthermore, if A and B are subspaces in $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, then $A \otimes B$ is a subspace in $\mathcal{H}_1 \otimes \mathcal{H}_2$. Finally, the projection of $|\lambda_1\rangle \otimes |\lambda_2\rangle$ onto $A \otimes B$ is simply the tensor product of the two projected vectors (if one of the projections is null, then the result is the null vector in $\mathcal{H}_1 \otimes \mathcal{H}_2$). Eventually, the norm $||.||$ of a vector $|\lambda_1\rangle \otimes |\lambda_2\rangle$ is the product of the single norms. The definition of the projection and the norm in the tensor space allows us to compute the probability of any event as described in Section 3. Note that these operations can be extended to tensor products of more than two spaces.

In quantum mechanics, tensor spaces and the tensor product are used to create composite spaces out of single component spaces. In a tensor product of spaces, we can distinguish *separable* and *non-separable* (or *entangled*) states. Separable states describe a state where the component states are independent of each other (that is, knowing something on a component state does not give any information about the other component states), whereas non-separable states are states where the composite state cannot be decomposed anymore into independent component states.

Figure 8 shows an example of separable and non-separable states which we will describe below. In Figure 8a, the system assumes that the user wants a document to be either authored by Smith or by Jones, and the rating to be good or that the user doesn't care about the rating. The possible states in the author-rating tensor product, which we call the *polyrepresentation space*, are given in the right hand side of the figure. As illustrated, we can isolate the IN of both representations, by operating in each representation separately and obtaining the corresponding probability of relevance. This is equivalent to computing the probability of relevance as in Equation 2 – the representation states are independent.

To cope with above situation where ratings and authors are interdependent, we cannot regard our single representations in isolation anymore. We must consider the composite polyrepresentation space. Figure 8b depicts a state which exactly reflects the situation given in the example. The combined representation ensemble consists of two states this time: the state s_1 defined as $|smith\rangle \otimes |good\rangle$ (good documents by Smith), and the state s_2 defined as $|jones\rangle \otimes |*\rangle$ (the user does not care about the ratings when the document is written by Jones). The peculiarity of this state is that we cannot isolate a state for each representation anymore. Both states shown in Figure 8 involve the same author and rating vectors, but only the state in Fig. 8a can be broken down to separate states in the author and rating space.

State	Projection	Probability	Non-sep.
$\|smith\rangle \otimes \|good\rangle$	null	0	■
$\|jones\rangle \otimes \|good\rangle$	null	0	
$\|smith\rangle \otimes \|*\rangle$	$\|smith\rangle \otimes \|*\rangle$	1	
$\|jones\rangle \otimes \|*\rangle$	$\|jones'\rangle \otimes \|*\rangle$	$\|\|\|jones'\rangle\|\|^2$	■

Table 1: Probabilities of relevance for O_{d_1} regarding the single state vectors in the author-rating polyrepresentation space. $|x'\rangle$is the projection of $|x\rangle$ onto the author subspace $O^{(a)}_{d_1}$. The column "Non-sep." shows the states that are part of the non-separable mixed state (denoted by a black square).

Let us see how this translates to the calculation of the probability of relevance in the polyrepresentation space. In Table 1, we see in the left column the states from Figure 8. Two of them are part of the non-separable mixed state, as shown in the right column. The relevance of a document w.r.t. the author and ratings is defined by a subspace $O_d = O^{(a)}_d \otimes O^{(r)}_d$, where $O^{(a)}_d$ is the subspace in the author space and $O^{(r)}_d$ the one in the rating space. Consider a document d_1 authored by Smith and rated "bad". $O^{(a)}_{d_1}$ is then a 2-dimensional subspace in the author space spanned by $|smith\rangle$ and $|*\rangle$, and from the way we model relevance in the rating space discussed in Section 4.2.2, $O^{(r)}_{d_1}$ would be the 2-dimensional subspace spanned by $|bad\rangle$ and $|*\rangle$. We can see that in the case of a separable state (Fig. 8a), document d_1 has a probability of 1 to be relevant to the state $|smith\rangle \otimes |*\rangle$ (3rd row). The probability that d_1 is relevant w.r.t. $|jones\rangle \otimes |*\rangle$ (4th row) is determined by the projection of $|jones\rangle$ onto O_{d_1} (remember that in this example we assumed that $|jones\rangle$ and $|smith\rangle$ are non-orthogonal). The final probability of relevance for a bad book by Smith is thus determined by the states in row 3 and 4 in the separable case, while in the non-separable case, only the probability in row 4 determines the relevance – due to the interdependence between authors and ratings, the state in row 3 was ruled out. The bad book by Smith would get a higher probability of relevance in the separate case than in the non-separate one. In the latter case, the probability of relevance just depends on the relationship between the authors Smith and Jones and the fact that the user does not care about the ratings for books by Jones.

With regards to the cognitive overlap, non-separable states and the interdependencies going along with them give us finer control what conditions between representations must be satisfied for a document to be in the cognitive overlap.

The remaining question is how a state in the polyrepresentation space can become non-separable. We can assume that we initially have a separable state, for instance a tensor product of the initial states of the single representations. A user might then state relationships like the ones above directly, or we may extract such interdependence from other sources, for instance by automatically creating association rules (like "author=Smith ⇒ rating=good"), known from data mining, from documents the user judged relevant during the session [29]. Such rules may directly be translated into subspaces inducing non-separable states like those in Fig. 8b by means of measurement.

6. RELATED WORK

There are basically three classes of related work: models, applications and the evaluation of the polyrepresentation principle.

Regarding (quantum) models supporting polyrepresentation, Melucci [16] proposes a dual approach to our framework where a subspace is used to describe a user's information need and a vector to represent documents. The probability that a document is relevant to a user's information need is determined by the projection of the document vector representation onto the corresponding IN subspace. Similar to ours, this approach also utilises the relationship between geometry and probability theory. In contrast to Melucci's idea and following the notion of state vectors and dynamics as applied in quantum mechanics, we interchanged the role of document and user's information need in our framework. This is motivated by the fact that the user's information need should be represented as a dynamic component, as advocated in e.g. [10]. The approach in [16] does not consider polyrepresentation per se, but in [2] an approach is proposed to use Melucci's framework for combining multiple sources of evidence. A document is still modelled as one vector in a vector space, but the same document can be described using different representations, where a document vector can be generated by a different vector space basis. In this work, though dealing with multiple evidence coming from different sources, no explicit relationship to the polyrepresentation principle is discussed.

Beckers [1] discusses the possible application of polyrepresentation of documents to support information seeking strategies, motivated with a book store example as in this paper. Other works considering different document contexts, like annotation-based retrieval [7], can be interpreted as an application of polyrepresented documents.

Since the introduction of polyrepresentation, for instance in [9], several experiments have been performed to validate the effectiveness of this principle regarding its various aspects. One aspect, which we focus on in our work, is the polyrepresentation of documents. The results reported in [26, 27] support the principle of polyrepresentation of documents and also show that assigning weights to different representations (higher weights for those with higher precision) can be crucial to gain better effectiveness, motivating the introduction of different representation weights into our framework through the "don't care" dimension. A second aspect of polyrepresentation considers the different representations of a user's information need, which includes among others the work task, the perceived information need, the experience, the domain knowledge and different query facets [11, 12, 5, 6, 14]. A third form of polyrepresentation sees different search engines as different reflections of the cognitive view of their designers on the retrieval problem [13, 15]. The main conclusion from evaluating all these facets of polyrepresentation is that the more positive evidence is coming from different representations, the more likely is the document in the cognitive overlap relevant to a given information need. This strongly supports the principle of polyrepresentation in general and also the idea to create a retrieval framework that explicitly considers polyrepresentation.

7. CONCLUSION AND OUTLOOK

In this paper, we discussed how to introduce the principle of polyrepresentation to a geometrical IR framework inspired by the quantum mechanics formalism, addressing some of the main requirements on a model for polyrepresentation. First, we recap our motivation for amalgamating a geometrical quantum formalism and polyrepresentation. After presenting the basic framework, we showed by example how textual and non-textual representations could be expressed within the framework, by defining vector spaces for the representation and subspaces of it for document representations. We have described how these representations benefited from the connection between geometry and probability present in the quantum formalism. To calculate the cognitive overlap and rank documents based on the probability that they lie in this overlap, we can combine the probabilities of relevance coming from the single representations to compute the probability that a document is in the total cognitive overlap. A "don't care" dimension, expressing the fact that the user does not use a representation to determine relevance, is introduced to each representation space to relax the strict obligation that an information object lies in the total cognitive overlap and assign weights to each representation. Finally, we discussed how we could handle situations in which the single representations are not independent any more from a user's point of view. To do so, the notion of non-separability and quantum entanglement is introduced in our framework, giving us the possibility to reflect representation interdependencies when creating the cognitive overlap.

We have succeeded in building a model that encompasses quantum mechanics and polyrepresentation and which is a strong platform for future work and further developments. As a next step, we plan to experiment with the framework. In particular, we would use the data collected in 2009-10 by the interactive track of the INitative for the Evaluation of XML Retrieval (INEX), where participants ran user experiments using a collection consisting of a crawl of over 2 million records (bibliographic metadata, reviews and ratings) from the online bookseller Amazon, enriched with reviews, ratings and tags coming from the cooperative book cataloguing tool LibraryThing[2]. The continuing user experiments yield log files capturing several kinds of user interaction (queries and query reformulations, relevance judgements) as well as insights which representations were actually considered by the users. These representations from different contexts give us the possibility to evaluate our framework and also polyrepresentation w.r.t. retrieval effectiveness, using simulated user interaction coming from these log files. The test data can also potentially be used to validate some of the assumptions underlying our framework, especially the non-independence of representations from a user point of view.

8. REFERENCES

[1] Thomas Beckers. Supporting Polyrepresentation and Information Seeking Strategies. In *Proceedings of the 3rd Symposium on Future Directions in Information Access (FDIA)*, pages 56–61, 2009.

[2] Emanuele Di Buccio, Mounia Lalmas, and Massimo Melucci. From Entities to Geometry: Towards

[2] See `http://www.inex.otago.ac.nz/tracks/interactive/interactive.asp` for further information

exploiting Multiple Sources to Predict Relevance. In *Proc. of the first Italian Information Retrieval Workshop (IIR'10)*, pages 35–39.

[3] Liang Chen, Jia Zeng, and Naoyuki Tokuda. A Stereo Document Representation for Textual Information Retrieval. *Journal of the American Society for Information Science*, 57(6):768–774, 2006.

[4] C. Cool, N. J. Belkin, O. Frieder, and P. Kantor. Characteristics of texts affecting relevance judgments. In *Proceedings of the 14th National Online Meeting*, pages 77–84, 1993.

[5] Abdigani Diriye, Ann Blandford, and Anastasios Tombros. A polyrepresentational approach to interactive query expansion. In *International Conference on Digital Libraries*, pages 217–220, 2009.

[6] Miles Efron and Megan Winget. Query Polyrepresentation for Ranking Retrieval Systems Without Relevance Judgments. *Journal of the American Society for Information Science and Technology.*

[7] Ingo Frommholz and Norbert Fuhr. Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. In Michael Nelson, Cathy Marshall, and Gary Marchionini, editors, *Proc. of the 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2006)*, pages 55–64, New York, 2006. ACM.

[8] Robert Budington Griffiths. *Consistent quantum theory.* Cambridge University Press, Cambridge; New York, 2002.

[9] Peter Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52:3–50, 1996.

[10] Peter Ingwersen and Kalvero Järvelin. *The turn: integration of information seeking and retrieval in context.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[11] Diane Kelly, Vijay Deepak Dollu, and Xin Fu. The loquacious user: a document-independent source of terms for query expansion. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 457–464, New York, NY, USA, 2005. ACM.

[12] Diane Kelly and Xin Fu. Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, 43(1):30–46, 2007.

[13] Birger Larsen, Peter Ingwersen, and Berit Lund. Data fusion according to the principle of polyrepresentation. *Journal of the American Society for Information Science and Technology*, 60:646–654, 2009.

[14] Christina Lioma, Birger Larsen, Hinrich Schütze, and Peter Ingwersen. A Subjective Logic Formalisation of the Principle of Polyrepresentation for Information Needs. In *Proceedings of the 3rd Symposium on Information Interaction in Information Retrieval - IIiX 2010 (forthcoming)*, 2010.

[15] Berit Lund, Jesper Schneider, and Peter Ingwersen. Impact of relevance intensity in test topics on IR performance in polyrepresentative exploratory search systems. In *Evaluating Exploratory Search Systems, Proceedings of the SIGIR 2006 EESS Workshop*, pages 42–46, 2006.

[16] Massimo Melucci. A basis for information retrieval in context. *ACM Transactions on Information Systems (TOIS)*, 26(3), 2008.

[17] Stefano Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, 10:303–320, 1998.

[18] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Interaction.* Cambridge University Press, Cambridge, UK, 2000.

[19] Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen. Exploring a Multidimensional Representation of Documents and Queries. In *Proceedings of the 9th RIAO Conferences (RIAO 2010)*, Paris, France, 2010.

[20] Benjamin Piwowarski, Ingo Frommholz, Yashar Moshfeghi, Mounia Lalmas, and Keith van Rijsbergen. Filtering documents with subspaces. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)*, 2010.

[21] Benjamin Piwowarski and Mounia Lalmas. A Quantum-based Model for Interactive Information Retrieval. In *Proceeedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR 2009)*, pages 224–231, Cambridge, UK, 2009.

[22] Benjamin Piwowarski and Mounia Lalmas. Structured Information Retrieval and Quantum Theory. In Peter Bruza, Donald Sofge, William Lawless, Keith van Rijsbergen, and Matthias Klusch, editors, *Proceedings of the Third International Symposium on Quantum Interaction (QI 2009)*, Lecture Notes in Computer Science, pages 289–298, Heidelberg et al., March 2009. Springer.

[23] Benjamin Piwowarski, Andrew Trotman, and Mounia Lalmas. Sound and complete relevance assessments for XML retrieval. *ACM TOIS*, 27(1):1–37, 2008.

[24] Gerard Salton and Michael J McGill. The SMART and SIRE experimental retrieval systems. *Document retrieval systems*, pages 192–229, 1988.

[25] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–346, 1975.

[26] Mette Skov, Birger Larsen, and Peter Ingwersen. Inter and intra-document contexts applied in polyrepresentation. In *IIiX: Proceedings of the 1st international conference on Information interaction in context*, pages 97–101, New York, NY, USA, 2006. ACM.

[27] Mette Skov, Birger Larsen, and Peter Ingwersen. Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing & Management*, 44(5):1673–1683, 2008.

[28] Keith van Rijsbergen. *The Geometry of Information Retrieval.* Cambridge University Press, New York, NY, USA, 2004.

[29] Ian H Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.