Exploring a Multidimensional Representation of Documents and Queries

Benjamin Piwowarski
University of Glasgow, UK
benjamin@bpiwowar.net

Ingo Frommholz
University of Glasgow, UK
ingo@dcs.gla.ac.uk

Mounia Lalmas
University of Glasgow, UK
mounia@acm.org

Keith van Rijsbergen
University of Glasgow, UK
keith@dcs.gla.ac.uk

ABSTRACT

In Information Retrieval (IR), whether implicitly or explicitly, queries and documents are often represented as vectors. However, it may be more beneficial to consider documents and/or queries as multidimensional objects. Our belief is this would allow building "truly" interactive IR systems, i.e., where interaction is fully incorporated in the IR framework.

The probabilistic formalism of quantum physics represents events and densities as multidimensional objects. This paper presents our first step towards building an interactive IR framework upon this formalism, by stating how the first interaction of the retrieval process, when the user types a query, can be formalised.

Categories and Subject Descriptors

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

General Terms

Theory

Keywords

Model, Quantum Theory

1. INTRODUCTION

Most information retrieval (IR) models, including probabilistic and vector ones, use the same underlying one-dimensional representation of documents and queries, i.e., as vectors defined in a vector space, typically a term space. However, this representation has some limits when dealing with more complex IR aspects like interaction, diversity and novelty[1]. Indeed, recent research showed that these complex aspects of the retrieval process benefit from more sophisticated representations of documents and queries [11, 3], in particular those providing for more powerful geometric manipulations of IR components.

The representation of documents and queries in IR should evolve so the user interaction can be incorporated in a natural and principled way in the IR process [10]. Our claim is that representing documents and queries as *multidimensional* objects (e.g. subspaces in a vector space) allows for not only a novel but also a more powerful way to tackle this challenge. This representation is particularly interesting from a theoretical point of view because it is possible to use a principled interpretation of the probabilities associated with such multidimensional objects, which comes from quantum physics [10] – the so-called "quantum probabilities" framework. This representation is also interesting from an intuitive point of view because it relies on a geometric representation of documents and queries in a vector space, which has proved successful in IR [2]. This representation reveals also a strong connection between orthogonality (in the vector space) and non-relevance, which has been successfully used to represent term negation in queries [12].

In [8], a framework for interactive IR that relies on such a multidimensional representation of documents and queries was proposed. In this framework, the user's information need (IN) is represented by a set of weighted vectors that evolve with the user's interaction. A probability of relevance of a document (for that IN) is computed with respect to this set. Although the components of our framework were described, they remained abstract. In particular, no explicit document and query representations were proposed. The next step is to operationalise the framework, which is the focus of this paper. We show how document and query representations are computed to then allow estimating the probability of relevance of the document to a given IN.

With respect to related work, multidimensional representations, respectively, of queries were used in [11] to model negative user feedback, and of documents were investigated in [3] in an ad hoc setting. Our work encompasses those since it provides a principled and probabilistic way to work with multidimensional objects. Finally, two lines of research explored, respectively, a subspace representation of documents [5] and of a user's IN [5]. In our work, we go further and show that both documents and INs can be represented as multidimensional objects, and propose a principled methodology to construct these representations.

The outline of this paper is as follows. We first briefly

[1]In our research, we are particularly interested in these aspects of the IR process.

RIAO'10, 2010, Paris, France.

introduce our framework and describe how the probability of relevance is computed within the quantum probability framework (Section 2). Next we show how we construct the query and document representations, and introduce several parameters for these representations (Sections 3 and 4). Experimental results (along with more experimental parameters) are reported in the full version of this paper [6].

2. A QUANTUM-INSPIRED VIEW FOR IR

Our IR framework is built upon [8], which is based on quantum probabilities and where we assume that there exists a vector space of *pure*[2] *information needs* (INs), where each vector corresponds to an IN that completely characterises a possible user's IN – by analogy with quantum physics where a vector completly characterises a physical system. Knowing a user's pure IN would determine which documents the IR system should return to that user. From a geometric perspective, a pure IN is answered by a document with a probability that depends on the length of the projection of the pure IN vector onto the document subspace. Because of the uncertainty attached to the IR search process, we suppose that the information being searched by a user can be represented by a set of such pure INs, one for each possible *pure* IN that composes a user's IN.

To compute a probability of relevance of a document to a user's IN, we make use of the generalisation of probabilities developed in quantum physics, which is strongly connected to the geometry of the space used to represent events and densities. A probabilistic event is represented as a subspace (denoted S) in a Hilbert space[3]. Let us assume that S is the event "the document is relevant". A probability can first be defined for a pure IN, represented as a *unit* vector φ, by computing the length of the projection of the vector φ onto the subspace S, that is by computing the value $\left\|\widehat{S}\varphi\right\|^2$ where $\widehat{S}$ is the projector onto the subspace S. This value is the probability that the document is relevant with respect to the pure IN[4].

When a user starts interacting with an IR system by, for instance, typing a query[5], we first compute (see Section 3) an initial set of weighted pure IN vectors, where each weight is the probability that the pure IN corresponds to the actual user's IN. This captures the uncertainty typical to IR where firstly, the representation is only an approximation of the user's IN, and, secondly, the query may be ambiguous. The goal of an IR system is to reduce this undeterminism through interaction.

More formally, we assume that each pure IN vector φ_i is associated with a probability p_i (the weight). We define the probability of the event S by using the usual total probability theorem (across all possible pure INs)[6]:

$$\Pr(S) = \sum_i p_i \Pr(S|\varphi_i) = \sum_i p_i \varphi_i^\top \widehat{S}\varphi_i = \mathrm{tr}\left(\rho\widehat{S}\right) \quad (1)$$

where tr is the trace operator [10, p. 83] and $\rho = \sum_i p_i \varphi_i \varphi_i^\top$ is called a *density operator*[7] and corresponds to a (probabilistic) *mixture* of the pure INs φ_i. In general, any operator ρ characterised by the fact that it is both positive-semi-definite[8] and of trace 1 defines a probability distribution over the subspaces, i.e. it is possible to interpret $\Pr(S) = \mathrm{tr}\left(\rho\widehat{S}\right)$ as a probability [10].

For each document d, we compute a projector $\widehat{S}_d$ (Section 3) and, for a query q, the IN density ρ is approximated by ρ_q (Section 4). Using the projector $\widehat{S}_d$ and the density ρ_q, the probability that a document d is relevant to the query q is then given by $\mathrm{tr}\left(\rho_q\widehat{S}_d\right)$.

In our work, we assume that the vector space of pure INs is the term space, where each dimension corresponds to a term. A pure IN is hence described by a series of weighted terms. A (simplified) example is shown in Figure 1, where the pure IN "pop music" (one unit vector) is represented by the terms "music", "chart" and "hit" of the term space. We show now how document and query representations are computed in this term space.

3. CREATING THE DOCUMENT SUBSPACE

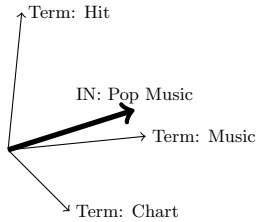


Figure 1: A pure IN in a term space

It is reasonable to assume that a typical document answers various (pure) INs, since it is likely to contain answers (be relevant) to several queries. Moreover, [9] have shown in the context of XML retrieval, that answers to topics (statements of INs) usually correspond to document fragments and not full documents. Building on this, we assume that for each document there is a mapping between its (possibly overlapping and non-contiguous) fragments and a set of pure INs.

A document is thus associated with a set $\mathcal{U}_d$ of vectors in the IN space. We hypothesise that a document is *fully* relevant to a pure IN if the latter can be written as a linear combination of the vectors of $\mathcal{U}_d$, that is, if it is contained in the subspace S_d defined as the span of the vectors in $\mathcal{U}_d$. The document will be *partially* relevant to a pure IN with a probability that depends on the length of the projection of the pure IN vector onto the subspace S_d. The subspace S_d can be interpreted as a geometric representation of the event "the document is relevant". This construction process was validated in a document filtering task [7].

4. CREATING THE QUERY DENSITY

[2]The concept of "pure" IN is new and central to our framework. In this paper, we use "pure IN" to distinguish it from "IN", where the latter refers to information need in its usual sense in IR, e.g., see [4].

[3]Hilbert spaces (roughly, vector spaces with complex scalars) are a central mathematical concept in quantum physics.

[4]We have $\left\|\widehat{S}\varphi\right\|^2 \in [0,1]$ since $\|\varphi\| = 1$.

[5]Queries are what (usually) users provide to an IR system, as means to express their INs [4].

[6]As in quantum physics, we assume different φ_i correspond to different systems and are thus mutually exclusive.

[7]We will omit the term "operator" in the remaining of the paper.

[8]This means $v^\top \rho v \geq 0$ for any vector v.

We now focus on the primary contribution of the paper, namely, the construction of the IN density ρ_q for a given query q.

As a query in its simplest form consists of a set of terms, we are first interested in building the query representation for a query composed of a single term, t. We described how a document is represented as a set of pure IN vectors corresponding to different fragments of the document. We extend this idea, and suppose that a query term t can be represented as the set $\mathcal{U}_t$ of pure IN vectors that correspond to document fragments containing the term t. That is, we use the immediate surroundings of the term occurrences in the documents of the collection being searched to build that term representation. This is similar to pseudo-relevance feedback using passages from retrieved documents containing the query terms [1]. The difference is that we use all the passages to build the query representation as we want to consider all possible pure INs associated with the term t.

As we have *a priori* no way to distinguish between the different vectors in $\mathcal{U}_t$, we assume that each vector is equally likely to be a pure IN composing the user's actual IN. Hence, a document is relevant to the user's IN if it is relevant to any of the vectors of $\mathcal{U}_t$, where the vectors are drawn with a uniform probability. The corresponding density is then written as:

$$\rho_t = \frac{1}{N_t} \sum_{\varphi \in \mathcal{U}_t} \varphi\varphi^\top \tag{2}$$

where N_t is the number of vectors associated with term t (the cardinality of $\mathcal{U}_t$). This definition of ρ_t has all the required properties of a density (see Section 2). In practice, this representation of a single-term query t means that, the more vectors φ from $\mathcal{U}_t$ lie in the document subspace, the higher the relevance of the document to the query. This query representation hence favours documents containing different "aspects" of the IN, each of them as represented by one of the pure INs in $\mathcal{U}_t$ associated with a query term t. We discuss next the representation of a query composed of several terms.

Query construction (mixture). The above query representation (Equation 2) can be generalised to a query composed of several terms. We assume that a relevant document should equally answer all pure INs associated with each query term. To compute the probability of relevance of a document d, we first select a term from the query (with a probability w_t), and then one of the vectors in $\mathcal{U}_t$. With this vector, we compute the probability of document d to be relevant to this pure IN. We repeat the process and average over all the possible combinations. This defines the probability of relevance of document d given the query. Formally, this corresponds to a density defined as a *mixture* of all the pure IN vectors associated with the query terms. This density is built from the individual query term densities ρ_t (Equation 2):

$$\rho_q^{(m)} = \sum_{t \in q} \sum_{\varphi \in \mathcal{U}_t} \frac{w_t}{N_t} \varphi\varphi^\top = \sum_{t \in q} w_t \rho_t \tag{3}$$

We present a second query construction process, inspired from IR and quantum theory. In vectorial IR, a query is represented by a vector that corresponds to a linear combination of the vectors associated with the query terms. In quantum theory, a normalised linear combination corresponds to the principle of superposition, where the description of a system state can be *superposed* to describe a new system state.

In our case, the system state corresponds to the user's pure IN, and we use the superposition principle to build new pure INs from existing ones, as illustrated with the example shown in Figure 2. Let φ_p, $\varphi_{c/uk}$ and $\varphi_{c/usa}$ be three vectors in a three-dimensional IN space that, respectively, represent the INs "I want a pizza", "I want it to be delivered in Cambridge (UK)" and "I want it to be delivered in Cambridge (USA)". The pure IN vector "Pizza delivered in Cambridge (UK)" would be represented by a (normalised) linear combination (or superposition) of φ_p and $\varphi_{c/uk}$, as depicted in Figure 2(a). We can similarly build the IN for Cambridge (USA). To represent the ambiguous query "pizza delivery in Cambridge" where we do not know whether Cambridge is in the USA or the UK, and assuming there is no other source of ambiguity, we would use a mixture of the two possible superposed INs, as depicted by the two vectors of the mixture in Figure 2(b), which brings us to another variant of query construction, the mixture of superpositions.

Query construction (mixture of superpositions). To compute the probability of relevance, for each term t of the query, we randomly select a vector from the set $\mathcal{U}_t$. We then superpose (i.e., compute a linear combination) the selected vectors (one for each term), where the weight in the linear combination is $\sqrt{w_t}$ (see below for why we use a square root). From this vector, we compute the probability of the document to be relevant to this IN made from the superposition of IN vectors (one per query term). With respect to our example, the set $\mathcal{U}_{pizza}$ would be just one vector ("*I want a pizza to be delivered*"), and $\mathcal{U}_{Cambridge}$ would contain two vectors (one for UK, one for USA).

As with the simple mixture approach, the above process can be repeated for all the possible selections of vectors and the corresponding query density is:

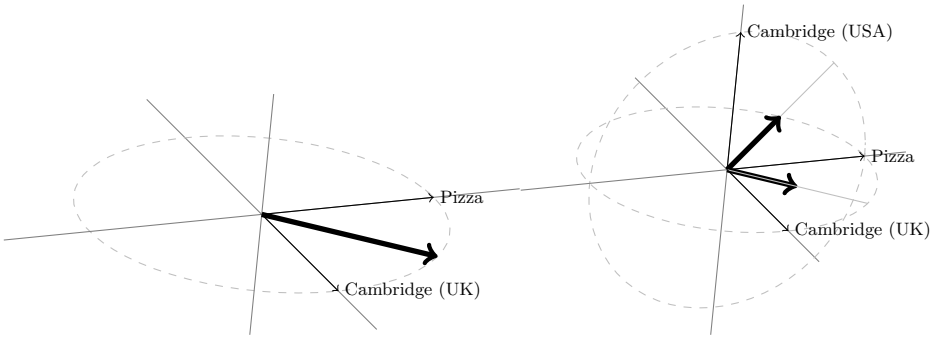
$$\rho_q^{(ms)} = \frac{1}{Z_q} \sum_{\varphi_1 \in \mathcal{U}_{t_1}} \cdots \sum_{\varphi_n \in \mathcal{U}_{t_n}} \left(\sum_{i=1}^{n} \sqrt{\frac{w_{t_i}}{N_{t_i}}} \varphi_i \right) \left(\sum_{i=1}^{n} \sqrt{\frac{w_{t_i}}{N_{t_i}}} \varphi_i \right)^\top \tag{4}$$

where Z_q is a normalisation coefficient, and $t_i \, (i = 1 \ldots n)$ are the n query terms. We use N_t to ensure that each term contribution is equally important, and square roots because both N_t and w_t appear two times in the above formula. In theory the vector $\sum_i \sqrt{\frac{w_{t_i}}{N_{t_i}}} \varphi_i$ should be normalised but to obtain a computable formula we did not do so [9].

Note that for one-term queries, the two described query constructions (mixture and mixture of superpositions) give the same result. Another important point from a computational perspective is that in both cases, the query can be estimated from single term densities (not demonstrated for Equation 4). We hence pre-compute the densities ρ_t for each term t, and use them at query time to compute $\rho_q^{(m)}$ and $\rho_q^{(ms)}$.

5. CONCLUSION AND FUTURE WORK

[9]The effect will be to give higher importance to superpositions of vectors φ_i who are similar, i.e., whose cosine is closer to 1.



(a) A superposition of two INs (b) A mixture of two INs

Figure 2: Combining INs

In this paper, we presented a methodology to build multidimensional representations of documents and queries. These representations are inspired from the geometric/probabilistic framework of quantum physics. The latter allows us to compute probabilities of relevance based on a more complex representations of documents than a simple bag of words, namely, a multidimensional one based on document fragments. We believe that such a multidimensional representation is key to a successful framework for exploiting user's interaction [10].

In the full version of this paper [6], we performed experiments to explore various parameters influencing the effectiveness of our representations. Our findings show that sentences are the best fragments, and that different weighting schemes for vectors perform similarly. With respect to the query representation, we show that queries whose terms define a concept and those whose terms are more independent are better handled by two different methods, respectively, the mixture of superpositions and the (simple) mixture. This suggests that we can gain further improvements if both strategies are applied together in an adaptive manner. This is part of our future work.

As our representation of queries and documents aims at tackling interactive IR, this works validates our framework for the most common first interaction step between a user and an IR system – a user typing a query. Exploiting further interaction steps (for example viewing or saving a document), is also part of our future work.

Acknowledgements. This research was supported by an Engineering and Physical Sciences Research Council grant (Grant Number EP/F015984/2). Mounia Lalmas is currently funded by Microsoft Research/Royal Academy of Engineering.

6. REFERENCES

[1] J. Allan. Relevance feedback with too much data. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *18th ACM SIGIR conference*, pages 337–343, Seattle, Washington, United States, 1995. ACM.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, USA, 1999.

[3] L. Che, J. Zen, and N. Tokud. A "stereo" document representation for textual information retrieval. *JASIST*, 5:768–777, April 2006.

[4] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag, Secaucus, NJ, USA, 2005.

[5] M. Melucci. A basis for information retrieval in context. *ACM TOIS*, 26(3):1–41, June 2008.

[6] B. Piwowarski, I. Frommholz, M. Lalmas, and K. van Rijsbergen. Exploring a multidimensional representation of documents and queries (extended version). *ArXiv e-prints*, (1002.3238), 2010.

[7] B. Piwowarski, I. Frommholz, Y. Moshfeghi, M. Lalmas, and K. van Rijsbergen. Filtering documents with subspaces. In *Proceedings of the 32nd ECIR Conference*, 2010. Poster.

[8] B. Piwowarski and M. Lalmas. A Quantum-based Model for Interactive Information Retrieval (extended version). *ArXiv e-prints*, (0906.4026), September 2009.

[9] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessments for XML retrieval. *ACM TOIS*, 27(1), jan 2009.

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

[11] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR*, pages 219–226, New York, NY, USA, July 2008. ACM.

[12] D. Widdows. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st ACL conference*, pages 136–143, Morristown, NJ, USA, 2003. Association for Computational Linguistics.