

Looking at Vector Space and Language Models for IR using Density Matrices

Alessandro Sordoni and Jian-Yun Nie

DIRO, Université de Montréal
Montréal, H3C 3J7, Québec

sordonia@iro.umontreal.ca, nie@iro.umontreal.ca

Abstract. In this work, we conduct a joint analysis of both Vector Space and Language Models for IR using the mathematical framework of Quantum Theory. We shed light on how both models allocate the space of density matrices. A density matrix is shown to be a general representational tool capable of leveraging capabilities of both VSM and LM representations thus paving the way for a new generation of retrieval models. We analyze the possible implications suggested by our findings.

1 Introduction

Information Retrieval (IR) has nowadays become the focus of a multidisciplinary research, combining mathematics, statistics, philosophy of language and of the mind and cognitive sciences. In addition to these, it has been recently argued that IR researchers should be looking into particular concepts borrowed from physics. Particularly, it was first evoked in 2004 in Van Rijsbergen’s pioneering manuscript “The Geometry of Information Retrieval” [15] that Quantum Theory principles could be beneficial to IR.

Despite Quantum Theory (QT) being an extremely successful theory in a number of fields, the idea of giving a quantum look to Information Retrieval could be at first classified as unjustified euphoria. However, the main motivation for this big leap should be found in the powerful mathematical framework embraced by the theory which offers a generalized view of probability measures defined on vector spaces. Events correspond to subspaces and generalized probability measures are parametrized by a special matrix, usually called *density matrix* or *density operator*. From an IR point of view, it is extremely attractive to deal with a formalism which embraces probability and geometry, those being two amongst the pillars of modern retrieval models. Even if we believe that an unification of retrieval approaches would be out-of-reach due to the intrinsic complexity of modern models, the framework of QT could give interesting overlooks and change of perspective thus fostering the design of new models. The opening lines of Van Rijsbergen manuscript perfectly reflect this interpretation: “It is about a way of looking, and it is about a formal language that can be used to describe the objects and processes in Information Retrieval” [15]. To this end, the last chapter of Van Rijsbergen’s book is mainly dedicated to a preliminary analysis

of IR models and tasks by means of the language of QT. Amongst others, the author deals with coordinate level matching and pseudo-relevance feedback.

Since then, the methods that stemmed from Van Rijsbergen’s initial intuition provided only limited evidence about the real usefulness and effectiveness of the framework for IR tasks [13, 26, 29, 20]. Several proposed approaches took inspiration from the key notions of the theory such as superposition, interference or entanglement. In [28], the authors use interference effects in order to model document dependence thus relaxing the strong assumption imposed by the probability ranking principle (PRP). An alternative solution to this problem has been proposed in [26], in which a novel reranking approach is proposed using a probabilistic model inspired by the notion of quantum measurement. In [13], the authors represent documents as subspaces and queries as density matrices. However, both documents and queries are estimated through passage-retrieval like heuristics, i.e. a document is divided into passages and is associated to a subspace spanned by the vectors corresponding to document passages. Different representations for the query density matrix are tested but none of them led to good retrieval performance. In [20], the authors work out an explicit interference formula in a topic model setting. Although marginal improvements are obtained over the baseline model, the ad-hoc application of the interference formula does not provide solid evidence towards the usefulness of the theory itself.

In order to give a stronger theoretical status to QT as a necessary or more general theory for IR, some authors step back into more theoretical considerations exposing potential improvements achievable over state-of-the-art models [23, 9, 10, 14]. In [10], the author shows how detection theory in QT offers a generalization of the Neyman-Pearson Lemma (NPL), which is shown to be strictly linked to the PRP. Dramatic potential improvements could be obtained by switching to such more general framework. Widdows [23] observed that the Vector Space Model (VSM) lacked a logic like the Boolean model. Through the formalism for quantum logic illustrated by Birkoff and Von Neumann [1], Widdows defines a geometry of word meaning by expressing word negation based on the notion of orthogonality. Recently, the work by Melucci and Van Rijsbergen [11] and Song et al. [18] offered a comprehensive review of QT methods for IR along with some insightful thoughts about possible reinterpretations of general IR methods (such as LSI [3]) from a quantum point of view. This paper shares the main purpose of the latter works.

In the ending section of his book, Van Rijsbergen calls for a reinterpretation of the Language Modeling (LM) approach for IR by means of the quantum framework. To our knowledge, such an interpretation has not been presented yet in the literature and this work can be considered as a first attempt to fill this gap. We provide a theoretical analysis of both LM and the VSM approach from a quantum point of view. In both models, documents and queries can be represented by means of density matrices. A density matrix is shown to be a general representational tool capable of leveraging capabilities of both VSM and LM representations thus paving the way for a new generation of retrieval models. As a conclusion, we analyze the possible implications suggested by our findings.

2 Quantum Probability and Density Matrices

In QT, the probabilistic space is naturally encapsulated in a complex vector space, specifically a Hilbert space, noted \mathbb{H}^n . We adopt the notation $|e_1\rangle, \dots, |e_n\rangle$ ¹ to denote the standard basis vectors in \mathbb{H}^n . In QT, events are no more defined as subsets but as subspaces, more specifically as projectors onto subspaces. Given a ket $|u\rangle$, the projector $|u\rangle\langle u|$ onto $|u\rangle$ is an elementary event of the quantum probability space, also called *dyad*. A dyad is always a projector onto a 1-dimensional space. Generally, a unit vector $|v\rangle = \sum_i v_i |u_i\rangle$, $v_i \in \mathbb{H}$, $\sum_i |v_i|^2 = 1$, is called a *superposition* of the $|u_i\rangle$ where $|u_1\rangle, \dots, |u_n\rangle$ form an orthonormal basis for \mathbb{H}^n .

A density matrix ρ is a symmetric positive semi-definite matrix of trace one. In QT, a density matrix defines the state of a system (a particle or an ensemble of particles) under consideration. Gleason's famous theorem [5] ensures that a density matrix is the unique way of defining quantum probability measures through the mapping $\mu_\rho(|u\rangle\langle u|) = \text{tr}(\rho|u\rangle\langle u|)$. The measure μ ensures that $\forall |u\rangle, \mu(|u\rangle\langle u|) \geq 0$. This is because, $\mu_\rho(|u\rangle\langle u|) = \langle u|\rho|u\rangle \geq 0$ because ρ is positive semi-definite. Moreover, if $|u_1\rangle, \dots, |u_n\rangle$ form an orthonormal system for \mathbb{H}^n , the probabilities for the dyads $|u_i\rangle\langle u_i|$ sum to one, i.e. they can be understood as disjoint events of a classical sample space. Given that $\sum_i |u_i\rangle\langle u_i| = I_n$, the identity matrix, we have $\sum_i \text{tr}(\rho|u_i\rangle\langle u_i|) = \text{tr}(\rho \sum_i |u_i\rangle\langle u_i|) = \text{tr}(\rho) = 1$. Therefore, for orthogonal decompositions of the vector space², a quantum probability measure μ reduces to a classical probability measure.

Any classical discrete probability distribution can be seen as a mixture over n elementary points, i.e. a parameter $\vec{\theta} = (\theta_1, \dots, \theta_n)$, $\theta_i \geq 0$, $\sum_i \theta_i = 1$. The density matrix is the straightforward generalization of this idea by considering a mixture over orthogonal dyads³, i.e. $\rho = \sum_i v_i |u_i\rangle\langle u_i|$, $v_i \geq 0$, $\sum_i v_i = 1$. Given a density matrix ρ , one can find the components dyads by taking its eigendecomposition and building a dyad for each eigenvector. We note such decomposition by $\rho = R\Lambda R^\dagger = \sum_{i=1}^n \lambda_i |r_i\rangle\langle r_i|$, where $|r_i\rangle$ are the eigenvectors and λ_i their corresponding eigenvalues. This decomposition always exists for density matrices [12]. Note that the vector of eigenvalues $\vec{\lambda} = (\lambda_1, \dots, \lambda_n)$ belongs to the simplex of classical discrete distributions over n points. If the distribution $\vec{\lambda}$ lies at a corner of the multinomial simplex, i.e. $\lambda_i = 1$ for some i , then the resulting density matrix consists of a single dyad and is called *pure state*. In the other cases, the density is called *mixed state*.

¹ The Dirac notation establishes that $|u\rangle$ denotes a unit norm vector in \mathbb{H}^n and $\langle u|$ its conjugate transpose.

² In a more general formulation of the theory, a quantum probability measure reduces to a classical probability measure for any set $\mathcal{M} = \{M_i\}$ of positive operators M_i such that $\sum_i M_i = I_n$. The set \mathcal{M} is called Positive-Operator Valued Measure (POVM) [12]. Therefore, the properties reported in this paper which apply to a complete set of mutually orthogonal projectors equally hold for a general POVM.

³ In general, the dyads in the mixture don't need to be orthogonal. However, in this case, the coefficients v_i cannot be easily interpreted as the probabilities assigned by the density matrix to each dyad.

Conventional probability distributions can be represented by diagonal density matrices. In this case, a classical sample space of n points corresponds to the set of projectors onto the standard basis $\{|e_1\rangle\langle e_1|, \dots, |e_n\rangle\langle e_n|\}$. Hence, the density matrix corresponding to the multinomial parameter $\vec{\theta}$ above can be represented as a mixture, $\rho_{\theta} = \text{diag}(\vec{\theta}) = \sum_i \theta_i |e_i\rangle\langle e_i|$. As an example, the density matrix ρ_{θ} below corresponds to a classical probability distribution with $n = 2$, σ is a pure state and ρ is a general quantum density, a mixed state:

$$\rho_{\theta} = \frac{1}{2}|e_a\rangle\langle e_a| + \frac{1}{2}|e_b\rangle\langle e_b| = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \sigma = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \quad \rho = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}.$$

3 Looking at Language Models

In the Language Modeling approach to IR, each document \mathbf{d} is usually assigned a unigram language model $\vec{\theta}_d = (\theta_{d1}, \dots, \theta_{dn})$, i.e. a categorical distribution over the vocabulary sample space \mathcal{V} (of size n), $w \in \mathcal{V}$, $p_{\theta_d}(w) = \theta_{dw}$ [25]. A query is represented as a sequence of terms $\{q_1, \dots, q_m\}$, sampled i.i.d. (independent and identically distributed) from the document model. The score for a document is obtained by computing the likelihood for the query to be generated by the corresponding document model:

$$\mathcal{L}(\{q_1, \dots, q_m\} | \vec{\theta}_d) = \prod_{i=1}^m p_{\theta_d}(q_i).$$

This scoring function is generally called Query Likelihood (QL). On the other hand, Kullback-Leibler (KL) divergence models can be seen as a generalization of QL models introduced in order to facilitate the use of feedback information in Language Modeling framework [25]. In KL-divergence models, both documents and queries are assigned to unigram language models. The score for a document is calculated as the negative query to document KL-divergence:

$$\mathcal{KL}(\vec{\theta}_q || \vec{\theta}_d) = - \sum_w \theta_{qw} \log \frac{\theta_{qw}}{\theta_{dw}}.$$

3.1 Query Likelihood View

As presented in Section 2, conventional probability distributions can be seen as diagonal density matrices. A straightforward quantum interpretation of the QL scoring function can be obtained by associating a diagonal density matrix to each document and consider a query as a sequence of dyads. Formally, we associate the vocabulary sample space to the orthogonal set of projectors on the standard basis, $\mathcal{E} = \{|e_1\rangle\langle e_1|, \dots, |e_n\rangle\langle e_n|\}$. The density matrix ρ for a document is a mixture over \mathcal{E} whose vector of weights corresponds to the parameter $\vec{\theta}_d$. Therefore, $\rho = \text{diag}(\vec{\theta}_d) = \sum_i \theta_{di} |e_i\rangle\langle e_i|$. It is straightforward to show that restricted to \mathcal{E} , μ_{ρ} generates the same statistics as $p_{\theta_d}(\cdot)$, i.e. $\forall w \in \mathcal{V}$:

$$\mu_{\rho}(|e_w\rangle\langle e_w|) = \text{tr}(\rho |e_w\rangle\langle e_w|) = \sum_i \theta_{di} \text{tr}(|e_i\rangle\langle e_i| |e_w\rangle\langle e_w|) = \theta_{dw} = p_{\theta_d}(w).$$

In the query likelihood view, the query is represented as an i.i.d. sample of word events. As word events correspond to projectors onto the standard basis, we represent a query as a sequence of i.i.d.⁴ quantum events belonging to \mathcal{E} , $\{|e_{q_1}\rangle\langle e_{q_1}|, \dots, |e_{q_m}\rangle\langle e_{q_m}|\}$. Therefore, the score for a document is computed by the following product:

$$\mathcal{L}(\{|e_{q_1}\rangle\langle e_{q_1}|, \dots, |e_{q_m}\rangle\langle e_{q_m}|\}|\rho) = \prod_{i=1}^m \mu_\rho(|e_{q_i}\rangle\langle e_{q_i}|) = \prod_{i=1}^m p_{\theta_d}(q_i), \quad (1)$$

which indeed corresponds to the classical QL scoring function. However, we shall stress out an important point about the equation above. If the projectors included in the query sequence are mutually orthogonal (as above), the calculation above behaves as a proper classical likelihood, i.e. the sum of the likelihoods of all possible samples of length m is one. On the contrary, the product cannot be considered as a classical likelihood because quantum probabilities for arbitrary events does not need to sum to one. Further considerations on these issues will be made in Section 6.

3.2 Divergence View

The KL scoring function computes a divergence between a query language model $\vec{\theta}_q$ and document language model $\vec{\theta}_d$. In QT, the KL-divergence is a special case of a more general divergence function acting on density matrices called Von-Neumann (VN) Divergence. Note $\rho = \sum_i \lambda_i |r_i\rangle\langle r_i|$, and $\sigma = \sum_i \zeta_i |s_i\rangle\langle s_i|$ the eigendecompositions of two arbitrary density matrices. In the following, the log function applied to a matrix refers to the matrix logarithm, i.e. the natural logarithm applied to the matrix eigenvalues, $\log \rho = \sum_i \log \lambda_i |r_i\rangle\langle r_i|$. The VN divergence writes as:

$$\mathcal{VN}(\rho\|\sigma) = \text{tr}(\rho(\log \rho - \log \sigma)) = \sum_i \lambda_i \log \lambda_i - \sum_{i,j} \lambda_i \log \zeta_j |\langle r_i | s_j \rangle|^2.$$

This divergence quantifies the difference in the eigenvalues as well as in the eigenvectors of the two density matrices [21].

In order to see how the classical KL retrieval framework is recovered, we assign a density matrix to the query very similarly to what has been done for a document. Precisely, ρ_q and ρ_d are diagonal density matrices such that $\rho_q = \sum_i \theta_{qi} |e_i\rangle\langle e_i|$ and $\rho_d = \sum_i \theta_{di} |e_i\rangle\langle e_i|$. As ρ_q (ρ_d) is diagonal in the standard basis, its eigenvalues correspond to $\vec{\theta}_q$ ($\vec{\theta}_d$), thus:

$$\mathcal{VN}(\rho_q\|\rho_d) = \sum_i \theta_{qi} \log \theta_{qi} - \sum_{i,j} \theta_{qi} \log \theta_{dj} |\langle e_i | e_j \rangle|^2 = \sum_i \theta_{qi} \log \frac{\theta_{qi}}{\theta_{di}}, \quad (2)$$

⁴ In quantum physics, the meaning of i.i.d. can be associated to the physical notion of measurement. If a density matrix ρ represents the state of a system, an i.i.d. set of m quantum events is obtained by performing a measurement on m different copies of ρ and by recording the outcomes.

which corresponds to the KL divergence. As conventional probability distributions correspond to diagonal density matrices, their eigensystem is fixed to be the identity matrix. Intuitively, KL divergence captures the dissimilarities in the way they distribute the probability mass on that eigensystem, i.e. by their eigenvalues.

4 Looking at the Vector Space Model

In this section, we are attempting to look at the VSM [17] in a new way. In its original formulation, no probabilistic interpretation could be given because of the lack of an explicit link between vector spaces and probability theory [24]. In the model, documents and queries are represented in the non-negative part of the vector space \mathbb{R}_+^n , where n is the number of terms in the collection vocabulary. In VSM, each term corresponds to a standard basis vector. The location of each object in the term space is defined by term weights (i.e. *tf*, *idf*, *tf-idf*) on each dimension. Similarity between documents and queries are computed through a vector similarity score $\vec{q}^\top \vec{d}$, where \vec{q}, \vec{d} are the vector representations of the query and the document. In [17], the authors show that normalizing document vectors is important to reduce bias introduced by variance on document lengths. By normalizing both document vector and query vector, the similarity score reduces to the cosine similarity between the two vectors, which is an effective similarity measure in the model [27]. From now on, we consider $|q\rangle, |d\rangle \in \mathbb{R}_+^n$, the normalized ($\|\cdot\|_2$) query vectors. Documents can thus be safely ranked by decreasing cosine $\langle q|d\rangle \in [0, 1]$, which cannot be negative because the ambient space is \mathbb{R}_+^n .⁵

4.1 Query Likelihood View

In this interpretation of the VSM, each document is associated to a probabilistic “model” in the same spirit of the Language Modeling approach. We define a density matrix ρ for the document as $\rho_d = |d\rangle\langle d|$, which is a pure state, i.e. its mixture weights are concentrated onto the projector $|d\rangle\langle d|$. Note that this density matrix does not have a statistical meaning. It has been determined by merely normalizing heuristic weighing schemes and it cannot be related to a statistical estimators such as Maximum Likelihood (MLE).

A query can be represented as the quantum event corresponding to the subspace spanned by $|q\rangle$. This subspace naturally corresponds to the dyad $|q\rangle\langle q|$.

⁵ In this paper, we do not explicitly take into account situations in which the vectors could contain negative entries. For example, this could easily happen after the application of Rocchio’s algorithm [16] in feedback situations or by reducing the dimensionality of the vector space by LSI [3]. Besides the historically encountered difficulties in the interpretation of such negative entries [6], in these particular cases, the rank equivalence situations discussed here could not hold. However, we argue that ignoring these situations causes no harm to the generality of our conclusions on the need of an enlarged representation space.

Hence, a query can be seen as the sequence of quantum events of length one $\{|q\rangle\langle q|\}$. In this setting, the likelihood given the document model is calculated by:

$$\mathcal{L}(\{|q\rangle\langle q|\}|\rho_d) = \mu_{\rho_d}(|q\rangle\langle q|) = \text{tr}(\rho_d|q\rangle\langle q|) = \text{tr}(\langle q|d\rangle\langle d|q\rangle) = |\langle q|d\rangle|^2, \quad (3)$$

The above calculation shows that the quantum “likelihood” assigned to the event $|q\rangle\langle q|$ by the density ρ_d is the square of the cosine similarity between the query and the document. When restricted to the non-negative domain, the square function is a monotonic, increasing transformation. This means that $\mu_{\rho_d}(|q\rangle\langle q|) \stackrel{rank}{=} \langle q|d\rangle$, i.e. the two formulations lead to the same document ranking.

4.2 Divergence View

According to the original VSM, queries and documents should share the same representation and the scoring function should be a distance measure between these representations. In the previous formalization, this initial paradigm seems apparently lost. The following alternative quantum interpretation of the VSM is perhaps closer to the original vision of the model. We associate a density matrix both to the document and to the query. Specifically, those density matrices would be pure states, projectors onto the corresponding vectors, i.e. $\rho_d = |d\rangle\langle d|$, $\rho_q = |q\rangle\langle q|$. It turns out that computing the Fidelity measure [12] between density matrices produces a ranking function equivalent to cosine similarity:

$$\mathcal{F}(\rho_q, \rho_d) = \text{tr}(\sqrt{\sqrt{\rho_q}\rho_d\sqrt{\rho_q}}) = \text{tr}(\sqrt{|q\rangle\langle q|d\rangle\langle d|q\rangle\langle q|}) = |\langle q|d\rangle|\text{tr}(\rho_q) = |\langle q|d\rangle|, \quad (4)$$

obtained by noting that ρ_q is a projector thus $\sqrt{\rho_q} = \rho_q$, and $\text{tr}(\rho_q) = 1$. As $|q\rangle, |d\rangle \in \mathbb{R}_+^n$, ranking by Fidelity measure is equivalent to ranking by cosine similarity, thus $\mathcal{F}(\rho_q, \rho_d) \stackrel{rank}{=} \langle q|d\rangle$.

5 A joint analysis

In this section, we will try to summarize the commonalities and the differences arising from the quantum formalizations of the two models given in the preceding sections. The following analysis is succinctly reported in Table 1. As a starting point, we shall note that the ambient space for both models is the Hilbert space \mathbb{H}^n , where n is the size of the collection vocabulary. Each standard basis vector $\mathcal{E} = \{|e_1\rangle, \dots, |e_n\rangle\}$ is associated to a word event. Therefore, the vocabulary sample space corresponds to the set of projectors onto the standard basis vectors $\{|e_i\rangle\langle e_i|\}_{i=1}^n$.

5.1 Query Likelihood View

In query likelihood interpretations, the query is represented as a sequence of i.i.d. dyads. In the VSM, the sequence contains one dyad corresponding to the

Table 1. Summary of the representations for documents and queries and the scoring functions of the two studied methods.

Query Likelihood View			
	Query	Document	Scoring
VSM	$\{ q\rangle\langle q \}$	$\rho_d = d\rangle\langle d $	$\mu_\rho(q\rangle\langle q)$
LM	$\{ e_{q_1}\rangle\langle e_{q_1} , \dots, e_{q_m}\rangle\langle e_{q_m} \}$	$\rho_d = \sum_w \theta_{dw} e_w\rangle\langle e_w $	$\prod_i \mu_{\rho_q}(e_{q_i}\rangle\langle e_{q_i})$
Divergence View			
	Query	Document	Scoring
VSM	$\rho_q = q\rangle\langle q $	$\rho_d = d\rangle\langle d $	$\mathcal{F}(\rho_q, \rho_d)$
LM	$\rho_q = \sum_w \theta_{qw} e_w\rangle\langle e_w $	$\rho_d = \sum_w \theta_{dw} e_w\rangle\langle e_w $	$-\mathcal{VN}(\rho_q \ \rho_d)$

projector onto the query vector $\{|q\rangle\langle q|\}$. On the contrary, in the LM approach the sequence contains a dyad for each classical word event, i.e. $\{|e_{q_i}\rangle\langle e_{q_i}|\}_{i=1}^m$.

Besides the number of dyads included in the sequence, a major difference distinguishes the two formalizations. Contrary to probabilistic retrieval models such as LM, a query is not considered as a sequence of independent classical word events but as a single event and a particular kind thereof. The query event is a *superposition* of word events. This can be seen because the vector $|q\rangle$ can be expressed, up to normalization, as $|q\rangle = \sum_w f(w) |e_w\rangle$ where $f(w)$ is the weight for term w in the query vector. This kind of event cannot be expressed using set theoretic operations neither it has a clear classical probabilistic interpretation: it does not belong to \mathcal{E} thus it can only be justified in the quantum probabilistic space. Arguing further, we would say that, in the case of VSM, term weighting methods aim at estimating the “best” query event, i.e. the event which is the most representative for the information need of the user. Intuitively, if a single choice would be given to us on what to observe, we would rather be observing in the “direction” of important words in the query.

It follows from the considerations above that VSM creates query representations by accessing the whole projective space through appropriate choices of $f(w)$. On the contrary, LM “sees”, and consequently can handle, only events from the classical sample space \mathcal{E} . However, the principled probabilistic foundations of the model give the flexibility of adding an arbitrary number of such events in the sequence, thus refining query representation⁶. In the next section, this kind of duality between VSM and LM approaches will be strengthened by analyzing the properties of the density matrices used in the two models.

Before continuing, we shall make one last consideration about the “likelihood” written in Eq. 1. This equation and its corresponding maximization algorithm have already been proposed by Lvovsky et al. [7] in Quantum Tomography applications in order to achieve a Maximum Likelihood Estimation (MLE) of a density matrix. As we have already pointed out, \mathcal{L} reduces to a classical likelihood if and only if the projectors in the sequence are picked from the same

⁶ This is indeed the practice of Query Expansion (QE), see for example [2].

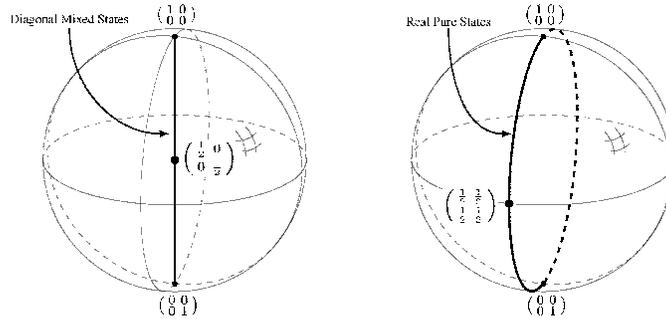


Fig. 1. The set \mathcal{D}^2 visualized using the Bloch sphere parametrization [12]. Highlighted in black are the region of \mathcal{D}^2 used by LM (to the left) and VSM (to the right).

eigensystem. Therefore, the product in its general form cannot be understood as a proper likelihood. We believe that it would be interesting to focus future research in finding a proper likelihood formulation in the quantum case that would enable principled statistical estimation and Bayesian inference (see [22] for a recent attempt in formulating a Bayesian calculus for density matrices).

5.2 Divergence View

In the divergence view, a density matrix is associated both to the document and to the query and the scoring function is a divergence defined on the set \mathcal{D}^n of $n \times n$ density matrices. Valuable insights can be provided by noting that the models gain access to different regions within \mathcal{D}^n . As an example, in Figure 1, we plot the set \mathcal{D}^2 using the well known Bloch parametrization [12]. Highlighted in black are the regions of the space used by LM (to the left) and VSM (to the right). Distinct regions are likely to denote different representational capabilities.

In the case of LM, density matrices are restricted to be diagonal, i.e. mixtures over the identity eigensystem. For two density matrices to be different, one has to modify the distribution of the eigenvalues. Therefore, LM ranks based upon differences in the eigenvalues between density matrices. The picture of the VSM approach appears as the perfect dual of the preceding situation. Query and documents are represented by *pure states*, i.e. dyads. Whatever the dimensionality of the Hilbert space, the mixture weights of these density matrices are concentrated onto a single projector. In order to be different, density matrices must be defined over different eigensystems. Therefore, VSM ranks based on the difference in the eigensystem between query and document density matrices.

The set of diagonal density matrices is represented in Figure 1 (left). Any two antipodal points on the surface of the sphere correspond to a particular eigensystem. Diagonal density matrices are restricted to the identity eigensystem. However, they can delve inside the sphere by spreading the probability mass across their eigenvalues. The black circle in Figure 1 (right) highlights pure states with real positive entries. These naturally lie on the surface of the Bloch sphere.

In summary, the VSM restriction to pure states leaves free choice on the eigensystem while fixing the eigenvalues. Conversely, by restricting density matrices to be diagonal, i.e. classical probability distributions, LM leaves free choice on the eigenvalues while fixing the eigensystem. Leveraging both degrees of freedom by employing the machinery of density matrices seems to be a natural step in order to achieve more precise representation for documents and queries. VSM and LM also differ in the choice of scoring functions. The former uses the Fidelity measure which is a metric on \mathcal{D}^n . The latter uses an asymmetric divergence on \mathcal{D}^n . More insights into these differences are given in the next section, where we try to contextualize our considerations by referring to common IR issues and concepts.

6 A Joint Interpretation and Perspectives

In [25], the author presents KL divergence models as “essentially similar to the vector-space model except that text representation is based on probability distributions rather than heuristically weighted term vectors”. The analysis done in the previous section extends this remark and highlights how VSM and LM leverage very different degrees of freedom by allocating different regions in \mathcal{D}^n . However, no clue is given about what should be the meaning of the eigensystems and the eigenvalues from an IR point of view, nor why controlling both could be useful for IR. We will try to give some perspective for the potential usefulness of the enlarged representation space.

In basic bag-of-words retrieval models such as LM or VSM, terms are assumed to be unrelated, in the sense that each term is considered to be an atomic unit of information. To enforce this view, LM associates to each term a sample point and the VSM a dimension in a vector space. Our analysis showed that sample points correspond to dimensions in a vector space. The heritage left by LSI [3] suggests that a natural interpretation for such dimensions is to consider them as *concepts*. In this work, we interpret projectors onto directions as concepts. Because terms are considered as unrelated, the projectors onto the standard basis $|e_1\rangle\langle e_1|, \dots, |e_n\rangle\langle e_n|$ in \mathbb{H}^n form a *conceptual basis* in which each term labels its own underlying concept.⁷

From this point of view, LM builds representations of queries and documents by expressing uncertainty on which concept chosen from the standard basis represents the information need. On the contrary, VSM does not have the flexibility of spreading probability weights. However, it can represent documents and queries by a unique but arbitrary concept. In VSM, the similarity score is computed by comparing how similar the query concept is to the document concept. In this picture, the cosine similarity reveals to be a measure of relatedness

⁷ In [8], *each* basis of a vector space is considered as describing a *contextual property* and the vectors in the basis as *contextual factors*. We prefer not to adopt such interpretation for two reasons: (1) in this paper, classical sample spaces are exclusively associated to orthonormal basis and (2) we believe that referring to concepts leads to a more general formulation, better tailored to our needs.

between concepts. In LM, the score is not at all computed on concept similarity, but by considering how the query and the document spread uncertainty on the same conceptual basis.

In order to see how this all could be instantiated, let us suppose that compound phrases such as “*computer architecture*” express a different concept than “*computer*” and “*architecture*” taken separately. Modeling interactions between terms has been a longstanding problem in IR (for example, see [4]). We conjecture that a very natural way to handle such cases stems from our analysis. Assume that both “*computer*” and “*architecture*” are associated to their corresponding single term concepts, i.e. $|e_c\rangle\langle e_c|$, $|e_a\rangle\langle e_a|$. The concept expressed by the compound could be associated to a superposition event $|k_{ca}\rangle\langle k_{ca}|$ where $|k_{ca}\rangle = f(c)|e_c\rangle + f(a)|e_a\rangle$ and f is a weight function (assuming normalization) expressing how compound and single term concepts are related. In this setting, the enlarged representation space turns out to be the perfect fit in order to express uncertainty on this set of concepts. One could build a density matrix associated both to a query and to a document assigning uncertainty to both single term concepts $|e_c\rangle\langle e_c|$, $|e_a\rangle\langle e_a|$ and compound concepts $|k_{ca}\rangle\langle k_{ca}|$. This could be done, for example, by leveraging quantum estimation methods such as described in [7]. As we have pointed out before, the VN divergence could be the right scoring function in order to take into account both divergences in uncertainty distribution and concept similarities. The practical instantiation of such general framework can be found in [19]. Experiments on several TREC collections show that the model leads to higher retrieval effectiveness than the existing models (in particular, LM).

As a last remark, we shall point out that the considerations made until now do not necessitate of the whole machinery of complex vector spaces. We do not have a practical justification for the usefulness of vector spaces defined over the complex fields (see [29] for a discussion on these issues). However, we speculate that these could bring improved representational power and thus remains an interesting direction to explore.

7 Conclusion

In this work, we showed how VSM and LM can be considered dual in how they allocate the representation space of density matrices and in the nature of their scoring functions. In our interpretation, VSM adopt a symmetric scoring function which measures the concept similarity. LM fixes the standard conceptual basis and scores documents against queries based on how they spread the probability mass on such basis. We argued that leveraging both degrees of freedom could lend a more precise representations of documents and queries and could be especially effective in modelling compound concepts arising from phrasal structures. The soundness and usefulness of this general setting has been confirmed experimentally in [19].

References

1. Garrett Birkhoff and John Von Neumann. The logic of quantum mechanics. *The Annals of Mathematics*, 37(4):823, October 1936.
2. Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. In *ACM Computing Surveys*, 44(1):1:1–1:50, January 2012.
3. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, 41:391–407, 1990.
4. J. Gao, J. Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *Proc. of SIGIR*, pages 170–177, 2004.
5. Andrew Gleason. Measures on the closed subspaces of a hilbert space. In *Journal of Mathematics and Mechanics*, 6:885–893, 1957.
6. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. In *Machine Learning*, 42(1):177–196, 2001.
7. A. I. Lvovsky. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. In *Journal of Optics B: Quantum and Semiclassical Optics*, S556–S559, 2004.
8. Massimo Melucci. A basis for information retrieval in context. In *ACM Transactions Information Systems*, 26:14:1–14:41, 2008.
9. Massimo Melucci. An investigation of quantum interference in information retrieval. In *Advances in Multidisciplinary Retrieval*, pages 136–151, 2010.
10. Massimo Melucci. Deriving a quantum information retrieval basis. *The Computer Journal*, 2012.
11. Massimo Melucci and Keith Rijsbergen. Quantum mechanics and information retrieval. In *Advanced Topics in Information Retrieval*, pages 125–155. Springer Berlin Heidelberg, 2011.
12. Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, January 2004.
13. Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen. What can quantum theory bring to information retrieval. In *Proc. of CIKM*, pages 59–68, 2010.
14. B. Piwowarski, and M. R. Amini and M. Lalmas. On using a quantum physics formalism for multidocument summarization In *JASIST*, 2012.
15. K. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.
16. J Rocchio. Relevance feedback in information retrieval. *The SMART Retrieval System*, pages 313–323. 1971.
17. Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513523, August 1988.
18. D. Song, M. Lalmas, K. van Rijsbergen, I. Frommholz, B. Piwowarski, J. Wang, P. Zhang, G. Zuccon, P. Bruza, S. Arafat, L. Azzopardi, A. Huertas-Rosero, Y. Hou, M. Melucci and S. Rüger. How Quantum Theory is Developing the Field of Information Retrieval. In *Proc. of QI*, pages 105–108, 2010.
19. A. Sordoni, J.-Y. Nie and Y. Bengio. Modeling term dependencies with quantum language models for IR. In *Proc. of SIGIR*, pages 653–662, 2013.
20. A. Sordoni, J. He and J.-Y. Nie. Modeling latent topic interactions with quantum interference for IR. To appear in *Proc. of CIKM*, 2013.
21. K. Tsuda, G. Ratsch, and M. K. Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6(1):995, 2006.
22. M. K. Warmuth and D. Kuzmin. Bayesian generalized probability calculus for density matrices. *Machine Learning*, 78(1-2):63–101, 2009.
23. D. Widdows and S. Peters. Quantum logic of word meanings: Concept lattices in vector space models. 2003.
24. S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst.*, 13(1):38–68, January 1995.
25. ChengXiang Zhai. Statistical language models for information retrieval a critical review. In *Found. Trends Inf. Retr.*, 2(3):137–213, 2007.
26. Xiaozhao Zhao, Peng Zhang, Dawei Song, and Yuexian Hou. A novel re-ranking approach inspired by quantum measurement. In *Proc. of ECIR*, pages 721–724, 2011.
27. Justin Zobel and Alistair Moffat. Exploring the similarity space. In *SIGIR Forum*, 32(1):18–34, 1998.
28. G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Proc. of ECIR*, page 357–369, 2010.
29. Guido Zuccon, Benjamin Piwowarski, and Leif Azzopardi. On the use of complex numbers in quantum models for information retrieval. In *Advances in Information Retrieval Theory*, pages 346–350, 2011.