Towards a Geometrical Model for Polyrepresentation of Information Objects

Ingo Frommholz and C. J. van Rijsbergen
Department of Computing Science
University of Glasgow
{ingo|keith}@dcs.gla.ac.uk

Abstract

The principle of polyrepresentation is one of the fundamental recent developments in the field of interactive retrieval. An open problem is how to define a framework which unifies different aspects of polyrepresentation and allows for their application in several ways. Such a framework can be of geometrical nature and it may embrace concepts known from quantum theory. In this short paper, we discuss by giving examples how this framework can look like, with a focus on information objects. We further show how it can be exploited to find a cognitive overlap of different representations on the one hand, and to combine different representations by means of knowledge augmentation on the other hand. We discuss the potential that lies within a geometrical framework and motivate its further development.

1 Introduction

One of the promising recent developments in information retrieval (IR) is the idea of *polyrepresentation*, which came up as a consequence of cognitive theory for interactive IR [Ingwersen and Järvelin, 2005]. The basic idea is that entities may be interpreted or represented in different functional and cognitive ways. Finding relevant documents goes along with finding the *cognitive overlap* of functionally or cognitively different information structures.

We can regard polyrepresentation w.r.t. *information objects* [Skov *et al.*, 2006]. For instance, a Web document can be represented by its content (which reflects the authors view on the document). Nowadays, it is common that users annotate a document in several ways. Annotations may be, for instance, comments, opinions, tags or ratings. Such annotations provide a cognitively different representation of a document, in this case reflecting the users' view on it. Another form of polyrepresentation considers the user's *cognitive state* [Kelly *et al.*, 2005] and different search engines [Larsen *et al.*, 2009]. The former one includes the work task, the perceived information need, the experience and the domain knowledge, and others. The latter one sees different *search engines* as different reflections of the cognitive view of its designers on the retrieval problem. One of the conclusions from evaluating all these facets of polyrepresentation is that the more positive evidence is coming from different representations, the more likely is the object in the cognitive overlap relevant to a given information need.

The experiments on polyrepresentation suggest that search effectiveness can benefit from a retrieval model which explicitly supports the principle of polyrepresentation. What is missing so far is a unified view which incorporates the different facets of polyrepresentation which allows for determining cognitive overlaps, but can go even beyond. For instance, different representations may be combined, as it is possible with knowledge augmentation (see below), to create a new representation. A unified view for polyrepresentation should also consider the combination of the concept of polyrepresentation with the dynamics arising from interactive retrieval. Such a view can be based on a geometrical model, as it was discussed in [van Rijsbergen, 2004]. A growing number of geometrical models, inspired by quantum theory, were introduced recently. For example, Piwowarski and Lalmas propose a geometrical framework which takes into account the evolution of the user's information need (represented as a vector in a Hilbert space) [Piwowarski and Lalmas, 2009]. So far, the concept of polyrepresentation has not been discussed in this model.

The considerations presented here are a first attempt to describe the notion of polyrepresentation (in particular w.r.t. information objects) in a geometrical way. They are a starting point for further discussion into how we can combine the idea of polyrepresentation and geometrical models, possibly inspired by quantum theory.

2 Towards a Geometrical Model

Our discussion starts with an example of how document features in different representations can be expressed geometrically. A representation of a document can be based on a set of distinct features. Such features can be topical, like the appearance of a term in a document, or non-topical, for example the document genre or the page rank. Documents may have static features (like terms and their weights), but they can also be dynamic (e.g., a property which shows whether a document was presented to the user or not). In general, we assume that for a feature f we can estimate the probability $\Pr(f|d)$ that we observe the feature given that we observed d. Similarly, $\Pr(\overline{f}|d) = 1 - \Pr(f|d)$ denotes the probability that we do not observe the feature.

Before we continue our considerations by giving an example, we introduce the notation, which is used in quantum mechanics as well.

2.1 Notation

We give a short introduction to the *Dirac notation*, which we are going to use in the following. A vector x in a real[1] n-dimensional Hilbert space $\mathcal{H}$ can be written as a so-called

[1] Our considerations can be expanded to complex Hilbert spaces, but for the time being it is sufficient to assume that $\mathcal{H}$ is spanned over $\mathbb{R}$

ket in Dirac notation:

$$|\,x\,\rangle = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

with $x_1, \ldots x_n \in \mathbb{R}$. Transposed vectors are represented as a *bra*, that is $\langle\, x\,| = (x_1, \ldots, x_n)$. Based on this we can define an *inner product* between two vectors x and y as $\langle\, x \mid y\,\rangle = \sum_{i=1}^{n} y_i x_i$ if we assume a canonical basis.

Besides inner products, we can also define an *outer product* as $|\,x\,\rangle\langle\, y\,| = xy^T$, which yields a square $n \times n$ matrix in our case. Each such matrix can be regarded as a linear transformation or *operator*. *Projectors* are idempotent, self-adjoint linear operators; they can be used to project vectors onto subspaces. For example, let $|\,e_0\,\rangle = (1,0)^T$ and $|\,e_1\,\rangle = (0,1)^T$ be the base vectors of a two-dimensional vector space $\mathcal{H}$, and $|\,x\,\rangle = (x_1, x_2)^T$ a vector in $\mathcal{H}$. Then $\mathbf{P} = |\,e_0\,\rangle\langle\, e_0\,|$ is a projector onto the one-dimensional subspace spanned by $|\,e_0\,\rangle$; $\mathbf{P}\,|\,x\,\rangle = (x_1, 0)^T$ is the projection of $|\,x\,\rangle$ onto that subspace. $||x|| = \sqrt{\sum_i^n x_i^2}$ denotes the *norm* of a vector, and $||x|| = 1$ means the vector is a *unit vector*. If $|\,e_0\,\rangle, \ldots, |\,e_n\,\rangle$ form an orthonormal basis of a Hilbert space, then $\mathrm{tr}(\mathbf{T}) = \sum_{i=1}^{n} \langle\, e_i\,|\,\mathbf{T}\,|\,e_i\,\rangle$ is called the *trace* of the matrix $\mathbf{T}$. It is the sum of the diagonal elements of $\mathbf{T}$.

2.2 Polyrepresentation of Information Objects

Representing Document Features

Our basic idea is to encode every feature in a *qubit* (quantum bit). A qubit is a two-dimensional subspace whose base represents two possible disjoint states $|\,0\,\rangle = (1,0)^T$ and $|\,1\,\rangle = (0,1)^T$. We give some examples of how a document feature, in particular a term, can be expressed as a qubit.

Let us assume we have a probabilistic indexer which assigns two probabilities to each term w.r.t. its corresponding document: $\Pr(t|d)$ is the probability that document d could be indexed with t, and $\Pr(\bar{t}|d) = 1 - \Pr(t|d)$ is the probability that it could not. Let $|\,0_t\,\rangle$ and $|\,1_t\,\rangle$ be the base

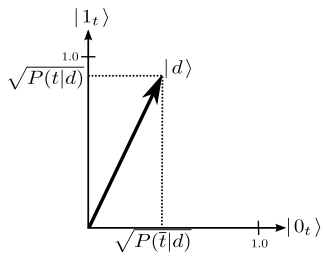

Figure 1: A term feature of d in a qubit

vectors of the qubit for term t. If we set $\alpha = \sqrt{\Pr(t|d)}$ and $\beta = \sqrt{\Pr(\bar{t}|d)}$, then $|\,d\,\rangle = \alpha \cdot |\,1_t\,\rangle + \beta \cdot |\,0_t\,\rangle$ is a unit vector (length 1). The situation is depicted in Figure 1 with $\Pr(t|d) = 0.8$ and $\Pr(\bar{t}|d) = 0.2$ resulting in $|\,d\,\rangle = (d_1, d_2)^T = (\sqrt{0.8}, \sqrt{0.2})^T$ in this qubit.

Retrieval with Polyrepresentation Example

Let us assume that we have a collection consisting of two terms, t_1 and t_2, and a document d with a user comment (annotation) a attached to it, so we have two cognitively different representations of the same document. We denote these two representations by two vectors, $|\,d_c\,\rangle$ for the content view and $|\,d_a\,\rangle$ for the annotation view on d. We give an example of how we can derive a simple well-known retrieval function from our representation, namely the traditional vector space model (VSM) which measures the similarity between a document and query vector in a term space. In order to support this, we need to transform our representation based on qubits into the classical vector space representation where the terms are the base vectors. One way to achieve this is to create a new vector $|\,d'_c\,\rangle = (d'_1, d'_2)^T$ with $d'_1 = |\,1_{t_1}\,\rangle\langle\, 1_{t_1}\,|\,|\,d_c\,\rangle$ (the projection of $|\,d_c\,\rangle$ onto $|\,1_{t_1}\,\rangle$) and $d'_2 = |\,1_{t_2}\,\rangle\langle\, 1_{t_2}\,|\,|\,d_c\,\rangle$. $|\,d'_c\,\rangle$ is then a vector in the classical term space known from the VSM. We can create $|\,d'_a\,\rangle$ out of $|\,d_a\,\rangle$ analogously. The new situation is depicted in Fig. 2. We can see that in contrast to the classical VSM, where the document is represented by only one vector, we now have two vectors for d, namely $|\,d'_c\,\rangle$ and $|\,d'_a\,\rangle$. We further assume that $\sum_{i=1}^{2} \Pr(t_i|d) = 1 = \sum_{i=1}^{2} \Pr(t_i|a)$, which means that $|\,d'_c\,\rangle$ and $|\,d'_a\,\rangle$ are unit vectors. We can

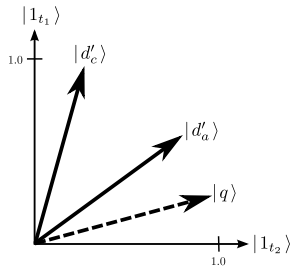

Figure 2: Document representation and query

represent a content query as a normalised vector $|\,q\,\rangle$ in the term space. We measure the similarity between the query and the two document representations by applying the trace function: $\mathrm{tr}(|\,q\,\rangle\langle\, q\,|\,|\,d_c\,\rangle\langle\, d_c\,|) = |\,\langle\, q \mid d_c\,\rangle\,|^2$. This equals $\cos^2\alpha$ (where α is the angle between $|\,d_c\,\rangle$ and $|\,q\,\rangle$) because we assume all vectors to be normalised. The beauty of this is that the resulting similarity value can be interpreted as a probability; see [van Rijsbergen, 2004, p. 83] for further details. By calculating the similarity between $|\,q\,\rangle$ and $|\,d_a\,\rangle$ analogously, we get two probabilities, one for the content representation and one for the representation of the document by its comment. These probabilities can now be used to determine the cognitive overlap of these representations and they can be combined to calculate the final score of d w.r.t. q.

We have seen that we can derive a well-known retrieval function from our representation. But what is the benefit from expressing features as qubits, when at least for term features we could have created a term space without introducing them? To answer this, we will now give an example of the possible combination of different representations, which relies on the proposed description of features as qubits.

Combination of Representations and Knowledge Augmentation

One may wonder whether the representation of features as a qubit is too redundant, since at least for the term features we also store $\Pr(\bar{t}|d)$, the probability that a document cannot be indexed with a certain term. While in general for other features it might be useful to store this probability as well, it can be incorporated in a senseful way when we want to combine different representations to create a

new one. This happens when for example we apply the concept of *knowledge augmentation*. Here, we augment our knowledge about an object with other objects which are connected to it, according to the probability that we actually consider them (see, e.g., [Frommholz and Fuhr, 2006] for a discussion of knowledge augmentation with annotations). Knowledge augmentation basically means to propagate features and their weights from connected objects to the one under consideration. A simple example shall illustrate knowledge augmentation in a geometrical framework. Again, we have a document d and an associated annotation a. We want to augment d with a, which means to propagate all term probabilities in a and also d to an augmented representation of d, denoted d^*. In d^*, the terms (features) and probabilities of d and a are aggregated. Along with a goes $\Pr(c_d^a)$, the probability that we consider a when processing d. One can think of this as a propagation factor[2]. We can store this probability in a qubit as discussed above; the corresponding vector is $|\,c\,\rangle = (c_1, c_2)^T = \sqrt{\Pr(c_d^a)} \cdot |\,1\,\rangle + \sqrt{1 - \Pr(c_d^a)} \cdot |\,0\,\rangle$. Based on this, we can now propagate a term t from d and a to d^* as follows. Qubits can be combined by means of tensor products, and we perform knowledge augmentation by calculating the tensor product of $|\,d\,\rangle$, $|\,c\,\rangle$ and $|\,a\,\rangle$:

$$|\,d^*\,\rangle = |\,d\,\rangle \otimes |\,c\,\rangle \otimes |\,a\,\rangle = \begin{pmatrix} d_1 \cdot c_1 \cdot a_1 \\ d_1 \cdot c_1 \cdot a_2 \\ d_1 \cdot c_2 \cdot a_1 \\ d_1 \cdot c_2 \cdot a_2 \\ d_2 \cdot c_1 \cdot a_1 \\ d_2 \cdot c_1 \cdot a_2 \\ d_2 \cdot c_2 \cdot a_1 \\ d_2 \cdot c_2 \cdot a_2 \end{pmatrix}$$

$|\,d^*\,\rangle$, which represents d^*, is a vector in an 8-dimensional space. The first element of $|\,d^*\,\rangle$ expresses the event that we index d with t (d_1) *and* consider a (c_1) *and* a is indexed with t (a_1). The fifth element denotes the case that we do *not* index d with t (d_2) *and* consider a (c_1) *and* a is indexed with t (a_1). Similarly for the other 6 elements. Each base vector thus represents a possible event, and all these events are disjoint. In fact, the resulting vector represents a probability distribution over these events and is thus a unit vector.

How can we now calculate the probability $\Pr(t|d^*)$ that we observe t in the augmented representation d^*? We observe t in the augmented representation in the following five cases: when we observe it in d, and when we do not observe it in d, but consider a and observe t there. These are exactly the events described by the first 5 elements of $|\,d^*\,\rangle$. These elements contribute to $\Pr(t|d^*)$, whereas the last 3 elements of $|\,d^*\,\rangle$ determine $\Pr(\bar{t}|d^*)$. To get $\Pr(t|d^*)$, we project $|\,d^*\,\rangle$ to the subspace spanned by the first 5 base vectors, and calculate the trace the projection. If $\mathbf{P_t}$ is such a projector, then $\Pr(t|d^*) = \mathrm{tr}(|\,d^*\,\rangle\,\langle\,d^*\,|\,\mathbf{P_t})$. Similarly for $\Pr(\bar{t}|d^*)$. Having achieved both probabilities, we can store them in a qubit as discussed above, and repeat the procedure for the other terms. Note that in this example, we combined a term-based representation with another term-based representation, but we are not bound to this. We can also combine topical and non-topical representations of a document in a similar way.

[2] A discussion of this probability is beyond the focus of this paper. It might be system-oriented, e.g. determined by the number of comments, or user-oriented, for instance by rating comments as important or less important.

3 Discussion

We have seen examples for polyrepresentation of information objects in a unified geometrical framework. Document features, be it content features expressed as terms, or non-topical ones, can be represented with the help of qubits which encode the probabilities that a certain feature can be observed or not. In this way, we can integrate different representations of documents in one model, calculate their relevance and use this information to compute the cognitive overlap. Different representations of documents may also be combined, as we have seen for knowledge augmentation. This way, we can exploit the polyrepresentation of information objects to obtain a higher-level representation. This simple example can of course not properly define a whole geometrical framework. This paper is not meant to deliver such a definition, but to undertake a first step towards it and to further motivate it. The following discussion shall reveal what we potentially gain when we further specify a geometrical framework which also includes inspirations coming from quantum mechanics.

We showed an example with different representations of information objects. In fact, also a polyrepresentation of search engines is potentially possible within our framework. How different retrieval models (like the generalised vector space model, the binary independent retrieval model or a language modelling approach) can be described geometrically is reported in [Rölleke *et al.*, 2006]. It is in principal possible to transfer these ideas into our framework, although it has yet to be specified which further knowledge (like relevance judgements) needs to be incorporated. Another extension of the model might also introduce polyrepresentation w.r.t the user's cognitive state, which may be represented as a vector similar to information objects.

The framework discussed in this paper may be used to support other models which indirectly apply polyrepresentation. An example is the Lacostir model introduced in [Fuhr *et al.*, 2008]. This model aims at the integration and utilisation of layout, content and structure (and thus polyrepresentation) of documents for interactive retrieval. The core part of the model consists of certain operations and their resulting system states. For instance, a selection operator (basically a query) lets the user choose relevant documents. Once relevant documents are determined, the user can select suitable representations thereof with a projection operator. An organisation operator can be applied by the user to organise the projected representations, for instance in a linear list or graph. With the visualisation operator, the user can choose between possible visualisations of the organised results. During a session, the user can at any time modify these operators. To support this model, an underlying framework must be capable of handling the different states the user and the system can be in as well as the transitions between them. It also needs to deal with the polyrepresentation of information objects. A geometrical framework can potentially handle the different representations and the dynamics in such a system. At least the selection and projection operators might be mapped to geometrical counterparts, whereas the organisation and visualisation operators may benefit from a geometrical representation of system states as vectors.

While we used some notations borrowed from quantum mechanics, the examples so far are purely classical, but with a geometrical interpretation. They give us a clue of the tight relation between geometry and probability theory and show the potential to embrace existing models in one uni-

fied framework. However, we did not touch any concepts used in quantum mechanics yet, like entanglement or complex numbers. For instance, different representations of an information object can be related, a property which we apply with knowledge augmentation. This relationship may also be expressed by using one state vector per feature and document, but with a different basis for each representation. Different representations may be entangled, and such property could easily be included in our model. An open question therefore is how the relationship between different representations should be modelled.

4 Related Work

The idea of using geometry for information retrieval, going far beyond the VSM, was formulated in [van Rijsbergen, 2004]. In this book, the strong connection between geometry, probability theory and logics is expressed. The examples in this paper are further inspired by Melucci's work reported in [Melucci, 2008]. Here, contextual factors (which may be different representations of information objects, but also reflect the user's point of view) are expressed as subspaces. Information objects and also queries are represented as vectors within these subspaces. Given this representation, a probability of context can be computed. This resembles the idea sketched in this paper, but the approach is not focused on polyrepresentation of objects. In the model presented by Piwowarski and Lalmas, a user's information need is represented as a state vector in a vector space which may for instance be set up by (possibly structured) documents [Piwowarski and Lalmas, 2009]. Initially, less is known about the actual information needs. Each user interaction gains more knowledge about her information need, which lets the state vector collapse until the information need is expressed unambiguously. Schmitt proposes QQL, a query language which integrates databases and IR [Schmitt, 2008]. In his work, he makes use of qubits as the atomic unit of retrieval values and interrelates quantum logic and quantum mechanics with database query processing. Further approaches about the relation of quantum theory and IR are reported in the proceedings of the Quantum Interaction symposium (see, e.g., [Bruza *et al.*, 2009]).

5 Conclusion and Future Work

In this short paper, we showed by an example how polyrepresentation of information objects can be realised geometrically. The goal is to undertake a first step towards a unified framework for polyrepresentation, which is missing so far. The example also shows how we can geometrically combine different representations to a new one. A subsequent discussion reveals some of the further possibilities coming from a geometrical approach.

We will also investigate the integration of existing quantum-inspired models into the framework, like the ones reported in [Piwowarski and Lalmas, 2009] or [Melucci, 2008], which do not deal with polyrepresentation yet. These models may thus be extended with the ideas that came up in the discussion so far, like knowledge augmentation and the possible entanglement of representations.

6 Acknowledgements

Our work was supported by the EPSRC project *Renaissance*[3] (grant number EP/F014384/1).

[3] `http://renaissance.dcs.gla.ac.uk/`

References

[Bruza *et al.*, 2009] Peter Bruza, Donald Sofge, William Lawless, Keith van Rijsbergen, and Matthias Klusch, editors. *Proceedings of the Third International Symposium on Quantum Interaction (QI 2009)*, LNCS, Heidelberg et al., March 2009. Springer.

[Frommholz and Fuhr, 2006] Ingo Frommholz and Norbert Fuhr. Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In M. Nelson, C. Marshall, and G. Marchionini, editors, *Proc. of the JCDL 2006*, pages 55–64, New York, 2006. ACM.

[Fuhr *et al.*, 2008] Norbert Fuhr, Matthias Jordan, and Ingo Frommholz. Combining cognitive and system-oriented approaches for designing IR user interfaces. In *Proceedings of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)*, October 2008.

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

[Kelly *et al.*, 2005] Diane Kelly, Vijay Deepak Dollu, and Xin Fu. The loquacious user: a document-independent source of terms for query expansion. In *Proceedings of the SIGIR 2005*, pages 457–464, New York, NY, USA, 2005. ACM.

[Larsen *et al.*, 2009] 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(4):646–654, 2009.

[Melucci, 2008] Massimo Melucci. A basis for information retrieval in context. *Information Processing & Management*, 26(3), June 2008.

[Piwowarski and Lalmas, 2009] Benjamin Piwowarski and Mounia Lalmas. Structured information retrieval and quantum theory. In Bruza et al. [2009], pages 289–298.

[Rölleke *et al.*, 2006] Thomas Rölleke, Theodora Tsikrika, and Gabriella Kazai. A general matrix framework for modelling information retrieval. *Information Processing and Management*, 42(1):4–30, 2006.

[Schmitt, 2008] Ingo Schmitt. QQL: A DB&IR query language. *The International Journal on Very Large Data Bases*, 17(1):39–56, 2008.

[Skov *et al.*, 2006] Mette Skov, Birger Larsen, and Peter Ingwersen. Inter and intra-document contexts applied in polyrepresentation. In *Proceedings of IIiX 2006*, pages 97–101, New York, NY, USA, 2006. ACM.

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