On the Probabilistic Logical Modelling of Quantum and Geometrically-Inspired IR

Fabrizio Smeraldi[1], Miguel Martinez-Alvarez[1], Ingo Frommholz[2], and Thomas Roelleke[1]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2] School of Computing Science, University of Glasgow, UK

Abstract. Information Retrieval approaches can mostly be classed into probabilistic, geometric or logic-based. Recently, a new unifying framework for IR has emerged that integrates a probabilistic description within a geometric framework, namely vectors in Hilbert spaces. The geometric model leads naturally to a predicate logic over linear subspaces, also known as quantum logic. In this paper we show the relation between this model and classic concepts such as the Generalised Vector Space Model, highlighting similarities and differences. We also show how some fundamental components of quantum-based IR can be modelled in a descriptive way using a well-established tool, i.e. Probabilistic Datalog.

1 Introduction

Three main branches of IR are probabilistic, geometric or logic-based IR. [13] discusses the relationship between these branches, showing that geometric approaches can have a probabilistic or logic-based interpretation, as it is known from quantum probabilities and quantum logics. Subsequent work discusses the prospect of such an interpretation for context-based or interactive IR [2] and specific retrieval tasks like diversity and novelty [7, 5]. On the other hand, logic-based approaches combine concepts from databases and IR and offer advanced means to flexibly create sophisticated retrieval functions and for structured queries. Combining logic-based approaches with geometric ones is thus a straightforward step that has been started in [11].

In this paper, we contribute a first step in extending with concepts from quantum mechanics a well-established logic-based framework (probabilistic Datalog) that has been used for modelling several IR tasks such as Information Extraction, being reported that it produces programs than are easy to understand, debug and modify [12]. After introducing main geometrical concepts from quantum mechanics, we show how a well-known geometric retrieval approach, the generalised vector space model (GVSM), relates to quantum probabilities and the total probability. We then explain how geometric concepts (e.g. Euclidean normalisation) can be realised in probabilistic Datalog. In particular, we address how

traditional maximum-likelihood estimates (L1-normalisation) and Euclidean estimates (L2-normalisation) are expressed and related in PDatalog. The main technical contribution of this paper is the probabilistic logical modelling of the mathematical concept of GIR, and the theorems and proofs to show the correctness of the PDatalog programs to model L1 and L2 probabilities.

2 Geometric IR (GIR)

2.1 Background

Quantum logic, i.e. logic on Hilbert spaces, allows us to cast information retrieval in a geometric context. A Hilbert space is a vector space endowed with an inner product - in the finite-dimensional case, we can think of the Euclidean space. In quantum logic predicates are represented by linear subspaces. If V and W are two subspaces (predicates), conjunction is given by their intersection (also a subspace) and alternation by the span of their union [13]. Negation is represented by the orthogonal. With this in mind, projections and orthogonality become important notions, and a special notation (the bra-ket notation) is introduced to facilitate computation.

2.2 Notation and computation

Given a Hilbert space $\mathcal{H}$, vectors are denoted by greek letters in an angled bracket, eg $|\psi\rangle$. This is called a "ket". The corresponding element of the dual is denoted by $\langle\psi|$, a "bra". The inner product of two vectors $|\phi\rangle$ and $|\psi\rangle$ is given by $\langle\psi|\phi\rangle$, the aptly-named "bracket" of the two vectors.

In Euclidean spaces the scalar product establishes a natural correspondence between the bra and ket spaces as follows:

$$\langle\psi_K|\cdot\rangle := K(\psi, \cdot) \tag{1}$$

where $K(\cdot, \cdot)$ is the scalar product and ψ is a (ket) vector in $\mathcal{H}$. This correspondence is invariant up to a rotation of the basis. We can therefore use the scalar product in the space to compute brackets, essentially thinking of $|\phi\rangle$ as a column vector, $\langle\psi|$ as a row vector, and $\langle\psi|\phi\rangle$ as a scalar product $\psi^t\phi$ (observables H below would be symmetric matrices)

2.3 Representing probabilities

As we have seen, a subspace identifies a predicate. We can represent probabilities by a *state vector* normalised to one, so that the square of the norm of its projection onto the subspace represents the probability of the predicate being true [13]. It is natural in the light of what we have seen above to use a Euclidean

normalisation $\langle\phi|\phi\rangle$=1. Remembering the analogy with the dot product above, this is expressed in components as follows: let $\{|e_i\rangle\}$ be the basis vectors, then

$$\langle\phi|\phi\rangle = \sum_i \langle\phi|e_i\rangle \langle e_i|\phi\rangle = \sum_i |\langle e_i|\phi\rangle|^2 = \sum_i \phi_i^2 \tag{2}$$

where by ϕ_i we indicate the component of ϕ along e_i. The projection of a state $|\phi\rangle$ onto $|e_i\rangle$ is obtained by applying the projection operator $|e_i\rangle\langle e_i|$ to $|\phi\rangle$, which following the mechanics of the notation gives

$$(|e_i\rangle \langle e_i|) |\phi\rangle = (\langle e_i|\phi\rangle) |e_i\rangle = \phi_i |e_i\rangle . \tag{3}$$

If $|\phi\rangle$ is normalised, it is immediate that ϕ_i^2 can be interpreted as a probability.

2.4 Representing retrieval

Documents are represented as vectors in a Hilbert space. In a two-dimensional space with basis vectors $|drive\rangle$ and $|school\rangle$, a document about driving schools might be seen as normalised coherent mixture of the basis states, taken for instance with equal weight:

$$|\psi\rangle = \frac{\sqrt{2}}{2} |drive\rangle + \frac{\sqrt{2}}{2} |school\rangle \tag{4}$$

$|\psi\rangle$ is a so-called *superposition* of the two states $|drive\rangle$ and $|school\rangle$. This differs from the case in which we do not know if the document is purely about driving or purely about schools. In quantum mechanics, such a condition is called a *mixed state* which is represented by a *density operator*

$$\rho = \frac{1}{2} |drive\rangle \langle drive| + \frac{1}{2} |school\rangle \langle school| . \tag{5}$$

This is the analogous of classical density matrices, see Equation 19.

These two alternative descriptions seem similar if we are interested for instance in the probability that the document is about driving:

$$|\langle drive|\psi\rangle|^2 = |\frac{\sqrt{2}}{2} \langle drive|drive\rangle + \frac{\sqrt{2}}{2} \langle drive|school\rangle|^2 = \frac{1}{2} \tag{6}$$

because $|drive\rangle$ and $|school\rangle$ are orthogonal. Similarly,

$$\begin{aligned} Tr(\rho |drv\rangle \langle drv|) = \\ \langle drv| \left(|sch\rangle \frac{1}{2} \langle sch| + |drv\rangle \frac{1}{2} \langle drv| \right) (|drv\rangle \langle drv|) |drv\rangle + \\ \langle sch| \left(|sch\rangle \frac{1}{2} \langle sch| + |drv\rangle \frac{1}{2} \langle drv| \right) (|drv\rangle \langle drv|) |sch\rangle = \\ \langle drv| \left(|sch\rangle \frac{1}{2} \langle sch| + |drv\rangle \frac{1}{2} \langle drv| \right) |drv\rangle = \frac{1}{2} \end{aligned} \tag{7}$$

where Tr denotes the trace function, the sum of the diagonal elements of a matrix. However the superposition case Equation 4 expresses the extent to which the document part-takes of both concepts $|drive\rangle$ and $|school\rangle$. This is made evident by the following example. Suppose we are interesting in finding documents about driving schools. We can then define the following *observable*:

$$H = |drive\rangle \langle school| + |school\rangle \langle drive| \tag{8}$$

In matrix notation it is represented by a symmetric matrix (one of Pauli)

$$H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \tag{9}$$

It is easy to see that the average value of the observable H on $|\psi\rangle$ is as follows:

$$\begin{aligned} &\langle\psi| H |\psi\rangle = \\ &\quad \frac{\sqrt{2}}{2} \left(\langle drv| + \langle sch|\right) \left(|drv\rangle \langle sch| + |sch\rangle \langle drv|\right) \frac{\sqrt{2}}{2} \left(|drv\rangle + |sch\rangle\right) = \\ &\qquad\qquad \frac{1}{2} \langle sch| \left(|drv\rangle + |sch\rangle\right) = \frac{1}{2} \end{aligned} \tag{10}$$

Actually one can show that H has Eigenvalues +1 and -1 and that the corresponding Eigenvectors are respectively $|\psi\rangle$ as in Equation 4 and

$$|\phi\rangle = \frac{\sqrt{2}}{2} |drive\rangle - \frac{\sqrt{2}}{2} |school\rangle \tag{11}$$

Now if we interpret the components ψ_i of the vectors ψ as TF-IDF frequencies, then we obtain:

$$\psi_i = \mathrm{TF}(t_i, \psi) \cdot \mathrm{IDF}(t_i) \tag{12}$$

Hereby, t_i is the term corresponding to dimension i, $\mathrm{TF}(t_i, \psi)$ is a frequency component, and $\mathrm{IDF}(t_i)$ is a measure to reflect the inverse document frequency (e.g. $\mathrm{IDF}(t_i) = -\log(n_D(t_i)/(N_D - n_D(t_i)))$, where $n_D(t_i)$ is the number of documents in which t_i occurs, and N_D is the total number of documents). Note that the vector component is negative for the case $n_D(t_i) > \frac{N_D}{2}$.

Of course, any other frequency or probabilistically motivated measure can be chosen as a vector component.

We can see that $Span(|\psi\rangle)$ represents the subspace on which the frequencies of the terms are positively correlated while $Span(|\phi\rangle)$ is the subspace on which they are negatively correlated.

3 Relationships between GIR and Traditional Concepts

This section outlines the relationships between "Geometric IR" and traditional concepts, namely the GVSM (section 3.1) and the total probability (section 3.2).

3.1 Generalised Vector-Space Model (GVSM)

The scalar product of two vectors can be re-written using the identity matrix as the intermediate between the vectors:

$$\boldsymbol{d}^T \cdot \boldsymbol{q} := \boldsymbol{d}^T \cdot I \cdot \boldsymbol{q} \tag{13}$$

where by ^{T} we indicate transposition.

In more general, the VSM is based on the idea to use the matrix G (a term-times-term) matrix between document and query.

$$\boldsymbol{d}^T \cdot \boldsymbol{q} := \boldsymbol{d}^T \cdot G \cdot \boldsymbol{q} \tag{14}$$

The matrix G may be used to associate semantically related terms. For example, setting $g_{12} = 1$ leads to the following equation:

$$\boldsymbol{d}^T \cdot G \cdot \boldsymbol{q} = d_1 g_{11} q_1 + d_1 g_{12} q_2 + d_2 g_{22} q_2$$

Now, the first term (e.g. the term "dog") is related to the second term (e.g. the term "animal"), i.e. a query containing "animal" will retrieve documents that contain "dog". Here, g_{21} may be not equal to g_{12}, if we wish to model a generalisation of terms. For synonyms, e.g. "classification" and "categorisation", we have $g_{ij} = g_{ji}$, i.e. the matrix G is symmetric for synonyms.

In a real-valued Hilbert space, a symmetric matrix G is a Hermitian operator and corresponds exactly to the observable H we introduced in Equation 8.

3.2 Total Probability

The total probability theorem is as follows:

$$P(q|d) = \sum_{t \in T} P(q|t) \cdot P(t|d) \tag{15}$$

Here, q and d are events, and T is a set of disjoint events. In the context of IR, let q be a query, d be a document, and t be a term. Using Bayes' theorem for $P(t|d)$, the theorem can be rewritten:

$$P(q|d) = \frac{1}{P(d)} \cdot \sum_{t} P(q|t) \cdot P(d|t) \cdot P(t) \tag{16}$$

This form relates the total probability to the GVSM, as is demonstrated in the following:

$$\boldsymbol{d} = (P(d|t_1), \ldots, P(d|t_n)) \tag{17}$$

$$\boldsymbol{q} = (P(q|t_1), \ldots, P(q|t_n)) \tag{18}$$

$$G = \begin{bmatrix} P(t_1) & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & P(t_n) \end{bmatrix} \tag{19}$$

Matrix G thus defined is a representation of a document vector $\boldsymbol{d}$ in terms of probabilities of disjoint events. In the quantum framework, disjoint events are orthogonal subspaces; thus G corresponds to the density matrix ρ introduced in Equation 5. We note that there is no classical analogy for the coherent quantum superposition as introduced in Equation 4.

4 Probabilistic Datalog: A Language for Probabilistic Logical Modelling

Probablistic Logical Modelling (PLM) is a modelling approach that is based on probability theory and logic. In principle, PLM is a theoretical framework composed of possible world semantics (logic and probability theory). Probabilistic extensions of standard languages (e.g. Probabilistic Datalog, [3], probabilistic SQL, [10]) are instances of probabilistic modelling languages. We present a brief overview over Probabilistic Datalog and show its application for the probabilistic logical modelling of the GVSM, the total probability, and TF-IDF.

Probabilistic Datalog (PDatalog) combines Datalog (query language used in deductive databases) and probability theory [3, 9]. It was extended to improve its expressiveness and scalability for modelling ranking models [10]. In addition, it is a flexible platform that has been used as an intermediate processing layer for semantic/terminological logics in different IR tasks such as *ad-hoc* retrieval [4, 6] or annotated document retrieval [1].It allows Bayesian goals and subgoals (for the modelling of probability estimation and conditional probabilities) and assumptions like SUM, PROD, SQRT or ROOT for combining probabilities of tuples. For example, given the information about grades for a given degree. $P(grade|person)$ can be inferred with the model shown in Figure 1.

4.1 Probabilistic Logical Modelling of the GVSM

Figure 2 shows the modelling of an scalar product and the GVSM model in PDatalog using the relation "g" as the matrix representing relationship between different terms.

4.2 Probabilistic Logical Modelling of the Total Probability

The modelling of the Total Probability is illustrated in figure 3. $P(term|doc)$ and $P(query|term)$ are obtained using the Bayesian assumption of PDatalog.

4.3 Probabilistic Logical Modelling of TF-IDF

Figure 4 shows a PD program illustrating the main ingredients for the probabilistic logical modelling of TF-IDF.

```
1  p_grade_degree SUM(Grade, Degree) :- grade(Student, Grade, Degree) | (Degree);

3  # Extensional evidence:
4  grade(john, "B", art); grade(anna, "A", art);
5  grade(mary, "B", maths); grade(peter, "B", maths); grade(paul, "C", maths);

7  ?- p_grade_degree(Grade, Degree);
8  # 0.5 ("B", art);  0.5 ("A", art); 0.667 ("B", maths); 0.333 ("C", maths)

10 # For a person Mr. X that has joined both arts and maths, what is the probability
       of "A", i.e. P(grade|person)?
11 0.5 register(mr_x, maths); 0.5 register(mr_x, arts);

13 # P(degree|person)
14 p_degree_person(Degree, Person) :- register(Person, Degree) | (Person)

16 # P(grade|person): Using Total Probability
17 p_grade_person SUM(Grade, Person) :-
18         p_grade_degree(Grade, Degree) & p_degree_person(Degree, Person);
```

Fig. 1. PDatalog example: Estimation of $P(\text{grade}|\text{degree})$

```
1  # vec(d) * vec(q) and vec(d) * G * vec(q)
2  scalar_product SUM(D, Q) :- vec_q(T, Q) & vec_d(T, D);
3  g_scalar_product SUM(D, Q) :- vec_q(T1, Q) & g(T1, T2) & vec_d(T2, D);

5  # Example: Main diagonal of G:
6  g(sailing, sailing); g(boats,boats) ;...

8  # Lower and upper triangles:
9  g(sailing ,boats); # For a query with "sailing" retrieve docs containing "boats"
10 g(boats, sailing); # For q query with "boats" ...
```

Fig. 2. GVSM in PDatalog

```
1  #P(term|doc)
2  p_t_d(Term, DocId) :- term(Term, DocId) | (DocId);

4  #P(query|term)
5  p_q_t(Term, QueryId) :- qterm(Term, QueryId) & pidf(Term);

7  #P(query|doc)
8  p_q_d SUM(DocId, QueryId) :- p_q_t(Term, QueryId) & p_t_d(term, DocId);
```

Fig. 3. Total Probability in PDatalog

```
1  # Within−document term probability:
2  # P(t|d):
3  p_t_d SUM(T, D) :− term(T, D) | DISJOINT(D);

5  # Collection−wide IDF−based term probability:
6  # Probabilistic IDF:
7  pidf(T) | MAX_IDF() :− term(T, D);

9  # Query term weighting:
10 w_qterm(T, Q) :− qterm(T, Q) & pidf(T);

12 # Normalisation of query term probabilities:
13 norm_w_qterm(T, Q) :− w_qterm(T, Q) | DISJOINT(Q);

15 # Retrieval:
16 retrieve SUM(D, Q) :− norm_w_qterm(T, Q) & p_t_d(T, D);
```

Fig. 4. Probabilistic Logical Modelling of TF-IDF

For "pidf(T)", the probability estimation is based on $\text{idf}(t)/\,\text{maxidf}$, and this value between 0 and 1 has a probabilistic semantics (see [8]), namely the probability to occur ($P(t$ occurs)) is equal to being not informative in maxidf trials, where the probability of being informative is $P(t \text{ informs}) := \text{idf}(t)/\,\text{maxidf}$.

The details of the meaning of TF and IDF-based probabilities is beyond the focus of this paper; however, important is that the TF and IDF-based probabilities described in the PDatalog program have a proper probabilistic semantics, and this leads to a probabilistic interpretation of the TF-IDF score.

The rule for "w_qterm(T,Q)" models IDF-based query term weighting. This is followed by a normalisation. The normalised tuple probabilities are then used for obtaining a probabilistic TF-IDF-based score in "retrieve(D,Q)".

In summary, this example illustrating the probabilistic logical modelling of TF-IDF highlighted that TF-based and IDF-based probabilities are combined to obtain a probabilistic TF-IDF-based score.

5 Probabilistic Logical Modelling of Geometric IR

Geometric IR can be viewed as a perspective for IR where the modelling of documents and queries is based on vectors. Quantum-inspired IR may be viewed as a modelling approach that combines geometric IR and probability theory. Essentially, the vector components are probabilities, and the combination of vectors (and/or matrices) yields probabilities.

The following sections present the modelling of GIR. Each section is related to the respective GIR section in which the mathematical foundations were reviewed.

5.1 Modelling probabilities (GIR 2.3)

As pointed out above, a central property of Quantum-inspired IR is related to the Euclidean norm, also referred to as the L_2 norm.

$$L_2(\boldsymbol{x}) := \sqrt{(\sum_i x_i^2)} \tag{20}$$

The L_1 norm is simply the sum over the vector components:

$$L_1(\boldsymbol{x}) := \sum_i x_i \tag{21}$$

With respect to probabilistic logical modelling, the L_1 norm corresponds to the assumption 'DISJOINT' (corresponds to maximum-likelihood (ML) estimate), and the L_2 norm is covered by the assumption 'EUCLIDEAN'.

Figure 5 shows the modelling of ML-based and Euclidean-based probabilities.

```
1  # Maximum−Likelihood P(t|d): based on the L1−norm.
2  # P_L1(t|d) = (sum_{t,d in term} P_term(t,d)) / (sum_{t' in d} P_term(t',d))
3  p_L1_t_d SUM(T, D) :− term(T, D) | DISJOINT(D);

5  # Example for doc2[sailing sailing boats]: docLen(doc2)=3
6  # L1(doc2) = 3 = 2 + 1
7  # 0.667 (sailing,doc2) # 2/3: Since 2 occurrences of sailing in doc2
8  # 0.333 (boats,doc2)   # 1/3: Since 1 occurrence of boats in doc2

10 # Euclidean P(t|d); based on the L2−norm.
11 # P_L2(t|d) = P_L1(t|d) / sqrt ( sum_{t' in d} square(P_L1(t'|d)) )
12 p_L2_t_d(T, D) :− p_L1_t_d(T, D) | EUCLIDEAN(D);

14 # Example:
15 # L2(doc2) = sqrt(5) = sqrt(2^2 + 1^2)
16 # 0.894 (sailing,doc2) # 2 / sqrt(5)
17 # 0.447 (boats,doc2)   # 1 / sqrt(5)
```

Fig. 5. Maximum-Likelihood and Euclidean Probabilities

Theorem 1. *The rule for "p_L1_t_d" is correct, i.e. the tuple probabilities in "p_L1_t_d" correspond to ML-probabilities of the form n/N where n is the sum of tuple probabilities in "term(t,d)", and N is the sum of tuple probabilities "term(·,d)", i.e. the sum of document tuples.*

Proof. The L1-based probability $P_{\mathrm{L1}}(t|d)$ is modelled in the rule for relation "p_L1_t_d". The rule body generates a probabilistic relation in which each rule

probability (from relation "term") is divided by the evidence probability, i.e. the sum of the tuple probabilities of the tuples that share the same evidence key. Here, "(D)" is the evidence key, i.e. the document id constitutes the evidence key. Therefore, the tuple probabilities generated by the rule body have the semantics $P_{\text{term}}((t,d))/\sum_{t' \in d} P_{\text{term}}((t',d))$.

The aggregation assumption in the rule head, i.e. SUM in p_t_d SUM(T,D), aggregates the tuple probabilities of non-distinct tuples.

For a non-probabilistic relation "term", "p_t_d" is the normalised within-document term frequency, i.e. $n_L(t,d)/\sum_{t' \in d} n_L(t',d)$, where $n_L(t,d)$ is the total occurrence of t (also denoted as $\text{tf}_d := n_L(t,d)$).

Theorem 2. *The rule for "p_L2_t_d" is correct, i.e. the tuple probabilities in "p_L2_t_d" correspond to the probabilities as required for GIR.*

Proof. Let x_i be the vector component for the i-th dimension. The Euclidean normalisation is:

$$\frac{x_i}{\sqrt{\sum_j x_j^2}}$$

Let $P_{L1}(t) := x_t/\sum_{t'} x_{t'}$ be the L1-based probability, and according to theorem 1, we find this probability in relation "p_L1_t_d".

The norm EUCLIDEAN in the subgoal of "p_L2_t_d" forms the sum of the squares of the probabilities that share the same evidence key. Then, for each tuple, the tuple probability is computed as follows:

$$P_{L2}(t|d) = \frac{P_{L1}(t|d)}{\sqrt{\sum_{t' \in d} (P_{L1}(t'|d))^2}}$$

Given the computation of $P_{L1}(t|d)$, the following equation holds:

$$\frac{P_{L1}(t|d)}{\sqrt{\sum_{t' \in d} (P_{L1}(t'|d))^2}} = \frac{x_t}{\sqrt{\sum_{t'} x_{t'}}} \tag{22}$$

Thus, the tuple probabilities in name p_L2_t_d are correct in the sense that they are based on the Euclidean normalisation.

5.2 Modelling retrieval (GIR 2.4)

The following PDatalog program illustrates the modelling of the GIR-based approach where L2-norm-based probabilities are combined in the rule body.

The PD program shows some rules to illustrate the modelling of various retrieval models. The rules shown underline that the models share the same pattern: a query representation is joined (matched) with a document representation, and the evidence from this match is aggregated.

```
1  # Euclidean P(t|d) and P(t|q) as defined previously:
2  p_L2_t_d(T, D) :- p_L1_t_d(T, D) | EUCLIDEAN(D);
3  p_L2_t_q(T, Q) :- p_L1_t_q(T, Q) | EUCLIDEAN(Q);

5  # Geometric IR:
6  gir_retrieve  SUM(D, Q) :- p_L2_t_q_idf(T, Q) & p_L2_t_d_idf(T, D);

8  # TF-IDF:
9  tf_idf_retrieve  SUM(D, Q) :- tf_q(T, Q) & pidf(T) & tf_d(T, D);

11 # VSM:
12 vec_q(T, Q) :- tf_q(T, Q) & pidf(T);
13 vec_d(T, D) :- tf_d(T, D) & pidf(T);
14 vsm_retrieve SUM(D, Q) :- vec_q(T, Q) & vec_d(T, D);

16 # GVSM:
17 gvsm_retrieve SUM(D, Q) :- vec_q(T_q, Q) & g(T_q, T_d) & vec_d(T_d, D);
```

Fig. 6. GIR Retrieval in PDatalog

6 Conclusions

This paper reviewed the basic concepts of a geometric and probabilistic approach to IR. In essence, vector and matrix components correspond to probabilities, and multiplications in the underlying vector spaces generate probabilities.

This paper has two main contributions. Firstly, section 3 relates the terminology of "Geometric IR" to traditional concepts such as the generalised vector space model (GVSM) and the total probability theorem, whereby we also underline the relationship between the GVSM and the total probability.

Secondly, section 4 reviews Probabilistic Datalog, and shows the probabilistic logical modelling of some standard models (TF-IDF, GVSM). Thirdly, section 5 added the modelling of selected concepts of "Geometric IR" (Euclidean-based probability estimation, modelling of retrieval models to outline the dualities between IR models).

This paper advocates probabilistic logic (PDatalog and descriptive modelling in general) as a potential platform to model IR. The overall finding of this paper is that the expressiveness of PDatalog is sufficient to model traditional IR models and concepts of "Geometric IR".

Future work will include to investigate quality and scalability of the shown approach. Given that PDatalog scales for TF-IDF, VSM, PIN, and language modelling, the hypothesis is that Euclidean-based normalisations and other concepts of GIR can be transferred to large-scale. Having discussed Euclidean probability estimations in this work, we will include other components of quantum mechanics, like projection, state vector ensembles (mixed states) and compound

systems expressed as tensor spaces, together with dynamics like state changes due to observation, to probabilistic Datalog.

7 Acknowledgements

We acknowledge the support of the UK EPSRC Project EP/F015984/1 Renaissance.

References

1. I. Frommholz and N. Fuhr. Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In *Proceedings of Joint Conference on Digital Libraries (JCDL'06)*, pages 55–64, 2006.
2. I. Frommholz, B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. van Rijsbergen. Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework. In *Proceedings of IIiX 2010*, pages 115–124, New Brunswick, Aug. 2010. ACM.
3. N. Fuhr. Probabilistic Datalog - a logic for powerful retrieval methods. In *Proceedings of the 18th ACM SIGIR Conference on Research and development in information retrieval (SIGIR'95)*, pages 282–290, 1995.
4. C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *ACM SIGIR conference on Research and development in information retrieval*, pages 298–307, 1993.
5. M. Melucci. A basis for information retrieval in context. *ACM Transactions on Information Systems (TOIS)*, 26(3), 2008.
6. H. Nottelmann. PIRE: An Extensible IR Engine Based on Probabilistic Datalog. In *Proceedings of the European Conference on Information Retrieval (ECIR'05)*, pages 260–274, 2005.
7. B. Piwowarski, I. Frommholz, M. Lalmas, and K. Van Rijsbergen. What can Quantum Theory Bring to Information Retrieval? In *Proc. 19th International Conference on Information and Knowledge Management*, pages 59–68, Oct. 2010.
8. T. Roelleke. A frequency-based and a Poisson-based probability of being informative. In *ACM SIGIR*, pages 227–234, Toronto, Canada, 2003.
9. T. Roelleke and N. Fuhr. Information retrieval with probabilistic Datalog. In F. Crestani, M. Lalmas, and C. J. Rijsbergen, editors, *Uncertainty and Logics - Advanced models for the representation and retrieval of information*. Kluwer Academic Publishers, 1998.
10. T. Roelleke, H. Wu, J. Wang, and H. Azzam. Modelling retrieval models in a probabilistic relational algebra with a new operator: The relational Bayes. *VLDB Journal*, 17(1):5–37, January 2008.
11. I. Schmitt. QQL : A DB&IR Query Language. *The VLDB journal*, 17(1):39–56, 2008.
12. W. Shen, A. Doan, J. F. Naughton, and R. Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. In *VLDB '07: International conference on Very large data bases*, pages 1033–1044, 2007.
13. C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004.