DRAFT

The geometry of information retrieval
and
the dimensions of perception.

by

C.J. van Rijsbergen

Introduction

Geometry has played a critical role in the development of Information Retrieval. The very first paper on Information Retrieval to be published in *The Computer Journal*, almost 50 years ago, in 1958, was by Robert Fairthorne [1][1]. At the same time he also published a paper in *American Documentation* [2] in which he stated, "We will now consider the concept of 'distance' as applied to what texts are about. It certainly has some meaning or meanings in this application, because we can talk about topics being 'close to' or 'remote from' each other, and tell people something useful by such terms." This represents a very early use of the formal notion of distance in IR. In the same article [2], Fairthorne recommends that non-Boolean logic be used when reasoning in IR. He asserts that "The mathematics of systems, or geometries, using these distances has been worked out, and is available in mathematical journals". Furthermore, again in the same article, "…if a request should be written out at length by the enquirer, some of its statistics may be used as retrieval specification." Here he draws attention to the use of statistics (and hence probability) in IR. Thus Fairthorne foreshadowed much of the research in modern information retrieval.

It is clear from the above that geometry, logic, and probability always played a significant role in the early work of IR researchers. Salton[3], undoubtedly one of the leading pioneers of the subject, continued this tradition in his early and later work. In his 1968 book he devoted considerable space to

[1] Here I pay tribute to almost 50 years of publishing of IR papers in The Computer Journal.

Boolean and non-Boolean models of retrieval (see his Chapter 6). He is credited with introducing the vector space model for retrieval, largely based on the exploitation of distance and geometry. He also, together with Karen Sparck Jones[4] and others (eg Luhn[5]), produced convincing arguments for the use of statistical methods in IR. The development of probabilistic methods, as opposed to statistical methods, emerged slightly later, although Maron[6] was early in the field with his paper with Kuhns on probabilistic indexing. A detailed account of the emergence of probabilistic retrieval may be found in [7].

So, the groundwork for the use of geometry in IR was laid many years ago in the early research of a number of people. On it, the development of a number other retrieval methods were based.

Classification

The idea of classification is quite an intuitive one. In computer science simple classifications have been used for a long time although not always seen as such. For example, an inverted file can be viewed as a simple classification. Objects that are described by attributes can be grouped on the basis of the attributes they share. In the case of an inverted file only one shared attribute is needed to form a class. In this way a set of overlapping classes is constructed one for each attribute.

One can generalise this simple classification to one where the classes are determined by not just one attribute but by a number. Formally this can be defined as follows. Let V be the set of possible attributes, and Ω the universe of objects, Δ is a relation on Ω×V, for example this could mean an object 'bird' has 'feathers'.

$tr(A) = \{t \in V \mid a \Delta t \text{ for all } a \in A\}$ defines the at*tr*ibutes of a class A, and $in(T) = \{a \in \Omega \mid a \Delta t \text{ for all } t \in T\}$ defines the *in*dividuals instantiating a set of attributes.

We can now give a definition of a class A of objects,

$A \subseteq \Omega$
$T \subseteq V$
$tr(A) = T$, the attributes shared by the members of A is T, but,

in(T) = A, the individuals instantiating T make up A.

In the above definitions A determines T, and T determines A. The classes definition is very strict, that is, it is not easy for classes satisfying this definition to be found. Moreover, the logic associated with this class structure turns out to be *non-distributive*. Consider the following example, in which B, H, and L are classes satisfying the above definitions.

Let B: the class of birds,
H: the class of humans,
L: the class of lizards, and
V: the class of vertebrates.

The distribution law reads $B \wedge (H \vee L) = (B \vee H) \wedge (B \vee L)$. If B, H, and L were sets (subsets) then the distribution law would hold, to verify this simply translate conjunction into set intersection, and disjunction into set union. This is a well known result in Boolean algebra. But with the class structure defined as above, the distribution law in general fails. To see this, notice that $H \vee L$ is the class of vertebrates V, and

$B \wedge V = B$,
but, $B \wedge H = B \wedge L = \Phi$ (the empty set)

[insert diagram]

Hence the distribution law fails, $B \neq \Phi$.

In the above discussion no mention was made of geometry or distance. One can indeed generalise these classificatory structures even further by introducing a notion of similarity/dissimilarity/distance between objects. To do this one first needs to specify a set with a metric structure that reflects the distance between pairs of objects. In general such a metric satisfies the following properties, for any pair of objects

1. $d(x,y) \geq 0$
2. $d(y,x) = d(x,y)$ symmetry
3. $d(x,x) = 0$
4. $d(x,y) \leq d(x,z) + d(z,y)$ triangle inequality

Using our geometrical intuition derived from our knowledge of 3-dimensional Euclidean space these properties seem entirely reasonable. However, each one of these has been challenged by psychologists, for example by Tversky [8].

A set endowed with this kind of metric, as defined in 1. – 4., is usually called a metric space. A typical example of such a space is our Euclidean space, in 3 dimensions, which has the following metric,

$$D(x,y) = (\Sigma (x_i - y_i)^2])^{1/2}$$

where x, y are the 3-dimensional vectors (x_i) and (y_i). In IR, documents or terms, are most often represented by n-dimensional vectors, and the metric distances are variations on the Euclidean distance. The geometric aspects of the space are emphasised by the triangle inequality (4.), which states that any side is less than or equal to the sum of the other two sides. A well known result from high school geometry. When objects are represented by vectors, a plethora of other measures are possible, for example, the angle between any two vectors is often used as a measure of similarity. That is,

$$\langle x|y\rangle = ||x||\ ||y|| \cos\theta$$

$$\cos\theta = \langle x|y\rangle / (||x||\ ||y||),$$

which requires an inner product, $\langle .|.\rangle$, to be defined on the space. Retrieval now becomes a simple matter of calculating a distance function, or similarity measure, between a query and documents, and ranking them in order of their matching value.

In general these spaces are very high dimensional and suffer from the 'curse of dimensionality'. There are various ways in which the situation can be improved, most of these amount to performing some kind of dimensionality reduction, or data reduction.

Let us consider the dimensionality reduction process first. Here the objects in the space are projected down onto a subspace of lower dimension. For example, if the space were merely three dimensional, then any point could be mapped down onto a plane spanned by any two co-ordinate vectors. The simplest version of this would be to map any (u,v,w) onto ($u,v,0$), thus the 3-

dimensional space is mapped down to a specific 2-dimensional space. Clearly this can be generalised in a number of ways, firstly, the subspace into which one maps can be any subspace. And, secondly, the criterion that optimises the mapping could be different, for example one may want to preserve a property like the rank ordering of interpoint distances, so that if $d(x,y) > d(z,p)$ in R^3, this is preserved in R^2.

The data reduction process is similar, yet different. The most common data reduction process is a classification of the objects. The usual way is to use the distance function in the space and decide whether a pair of objects belong to the same cluster, that is, are they close enough with respect to some given threshold, k, for example $d(x,y) < k$? There are highly organised ways of doing this, often generating a stratified hierarchic classification as output.

There is an interesting geometric interpretation of this classification method. A distance function between pairs of objects can be represented as a symmetric matrix where each entry is the distance between a pair of objects. A classification can now be viewed as a transformation from this matrix to a stratified hierarchy, often called a dendrogram. But there is more, a dendrogram can be represented by another matrix. Whereas any three entries, such as d(i,j), d(i,k), d(j,k) in the input matrix satisfy the triangle inequality, any three entries in the output matrix satisfy the ultrametric inequality, $d(i,j) \leq \max \{d(i,k), d(j,k)\}$. Hence a classification method can seen as a mapping from metric to ultrametric matrices. In the original space any three objects x,y, z form an arbitrary triangle, that is $d(x,y) \neq d(y,z) \neq d(z,x)$. The effect of a stratified hierarchic classification is to transform all these triangles into isosceles triangles, ones with at least two sides equal, thus satisfying the ultrametric inequality. So classification can be seen as a data reduction process whereby certain information about the original geometry is lost. Naturally there may be other criteria that can be optimised, since there will be in general many cluster methods that lead to different dendrograms, the choice of cluster method may depend on the one that satisfies the chosen optimality criterion best.

In a way a classification process is also a dimensionality reduction process. Once a classification has been constructed, individual objects are referred to the clusters they belong to, the clusters acting as the dimensions. Both clustering and dimensionality reduction provide a way of specifying a

context from which to perceive an object. By mapping into a particular context each object can be viewed from within that context. Similarly, by creating a classificatory structure, any object can be viewed from the point of view of any cluster. The perception of an object does not require that the object belong to the cluster in question. It is perfectly reasonable to perceive an object that lies outside the cluster.

Classifications of either documents or terms were used extensively from the very beginning of research in IR [9]. The motivation for their use was various. Documents were grouped to speed up searching, or in some cases to improve retrieval performance. Terms were grouped to provide substitutes, or additional terms, for searching [10].

Logic lost and regained.

The most intuitive way to introduce logic is via set theory. We are all familiar with how all subsets of a universal set X make up what is called the power set 2^X. The set theoretical operators $\cap$, $\cup$, and $\perp$, respectively intersection, union and set-theoretic complementation, correspond to the logical connectives, $\wedge$, $\vee$, and $\neg$ for the Boolean lattice made up of the propositions corresponding to the subsets. Subset containment in 2^X corresponds to logical entailment on the lattice, usually denoted as $\leq$.

Early retrieval systems exploited this interpretation of Boolean logic to the full. Unfortunately Boolean retrieval turned out to be rather ineffective, and of a hit and miss nature, either too much or too little was retrieved, and indeed, some like Fairthorne thought we were using the wrong logic anyway. To go beyond Boole became imperative [11]. The earliest successes were mostly based on the use of distance/similarity and statistics, where the representation space was a vector space. Regrettably in the move to vector space both logic and probability were lost, to be regained later in the development of the field.

Kinematics

A good example of how, with a little help from geometry, logic and probability were recovered in vector space, is the work on Probability Kinematics [12]. The basis for this work was again to assume that the

geometric structure of the space of objects was essential in modelling retrieval.

The starting point for this research was to try and cast the matching process in IR as a form of plausible reasoning. The inspiration came from Modal Logic where possible world semantics had been used to great effect. The truth of a statement in a world is determined by the truth of it at accessible worlds. In Modal Logic the accessible worlds are given by Kripke structures, for which a world is either accessible or not. Of particular concern were the evaluation of conditional statements, and motivated by the Ramsey Test[2], these were evaluated using a generalised Kripke structure where accessibility was not a simple yes/no matter but one of degree. From this it was only a small step to consider the probability of statements as a function of the distance that other relevant worlds would have from a given world. The high point of this analysis was to evaluate the probability of a conditional in terms of a probability revision process known as 'imaging'.

[Insert Example and diagram]

So once again the underlying geometric structure of the space plays a critical role in IR matching. Another aspect of this work was the realisation that the space with which one worked could be a Term-space, that is, the objects were terms, as opposed to a Document-space, in which the objects were documents. In fact it is now clear that the distinction need not be made. A Hilbert space can accommodate both points of view, as already noted by researchers who applied LSI to IR. [13]

We are now in the position where a space, with a generalised Kripke structure can support a logic and is also able to deal with probabilistic reasoning.

Enter context (dimensions of perception)

In the last few years it has become apparent that future improvements in retrieval effectiveness can be achieved by considering retrieval in context.

---

[2]'To evaluate a conditional, first hypothetically make the minimal revision of your stock of beliefs required to assume the antecedent. Then evaluate the acceptability of the consequent on the basis of this revised body of beliefs.'

Context can mean different things to different people, here I shall be considering context in terms of topical, thematic, conceptual context. Physical context, emotional context, etc, although important, I shall ignore.

At the root of the modeling of context is the perception of properties. It is conventional wisdom in IR to perceive properties/attributes as belonging to objects. For example it is assumed that we can assign a term such as 'bank', or 'rabbit' to a document as if the document posessed that property. At first glance this seems intuitive. But now consider an image/picture, it may be that the image contains a line drawing which can be perceived by a user in one context as a rabbit, and in another context as a duck. In other words a property emerges from an interaction in a context. More explicitly one could argue that a property such as rabbit is only realised with a certain probability.

Thus a request for objects about rabbits needs to be augmented by information about the context. The proposal here is to represent a query by an observable which implicitly embodies the contextual information. Such an observable can be represented by a matrix, namely, an Hermitian matrix[3]. Motivation from QM suggests that these observables are non-commutative, that is, the result of measuring an observable followed by another depends on the *order* in which the measurements are made.

A good example of this non-commutativity in IR is the order in which 'relevance' and 'aboutness' are observed. A relevance assessment may change the perception of what a document is about, and of course the reverse applies too.

A formal way in which to capture these notions is to represent documents and terms (objects in general) in a Hilbert space, and to represent observables of any kind as Hermitian operators. This is of course how systems and observables are represented in Quantum Mechanics.

Many things follow from this way of representing the objects and processes in IR [14]. Most importantly it delivers a logic for reasoning about objects as subspaces. Birkhoff and von Neumann [15] in a seminal paper, showed how in general the lattice of subspaces is a non-distributive lattice, and hence

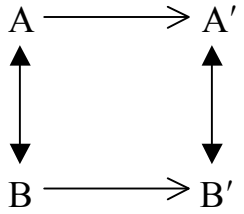
[3] A matrix is Hermitian if the conjugate transpose is equal to itself; when the entries are real numbers this reduces to it being symmetric.

non-Boolean. Each subspace corresponds to a proposition which is the answer to an elementary question, these questions are the projectors corresponding to the subspaces. An elementary question has a binary Y/N, or 1/0 answer, depending whether a proposition (subspace) is true for an object. Clearly if the object lies in the subspace the answer is true, and if it lies in its complement the answer is false. When an object is neither, the answer occurs with a certain probability. So we now have the situation where any object can be queried about any elementary property which corresponds to a subspace. In mathematical terms, if projector P corresponds to a question, then all the eigenvectors of $Px = x$ corresponding to eigenvalue 1, form a subspace of objects for which the answers to P is Yes.

Earlier we made the point that any general query could be represented by a Hermitian matrix[4]. So what happens when the query is applied. It turns out that any Hermitian operator, A, can be resolved into a number of elementary questions. This is result is known as the Spectral Theorem[16], and it goes as follows,

$$A = \lambda_1 P_1 + \ldots + \lambda_n P_n.$$

In the simplest case, for an n-dimensional Hilbert space, the λ_i are distinct the eigenvalues of A, and the P_i are 1-dimensional projectors. The eigenvectors corresponding to the P_i will make a basis for the space, so that any vector can be expressed as a linear combination of those basis vectors. That basis specifies a point-of-view, or context, from which to perceive all the vectors in the space. A context shift can now occur in two ways, either one transforms A to A′ to represent a context shift, or one specifies a change in the basis B to B′. Naturally these are equivalent, as illustrated by the following diagram,



The important point is that here is a mechanism for modelling context, and a way of observing it.

[4] Hermitian matrices and operators are now treated indistinguishably.

Enter Probability.

What we have thus far is a Hilbert space to represent our objects (documents, terms)[5], a set of Hermitian operators to represent queries, and a set of projectors to represent subspaces (or propositions). The final missing ingredient is probability.

There is a large literature on Quantum Probability[17] which distinguishes it from Kolmogorov Probability. The difference between the two is not important here, suffice it to say that it deals with the fact that in quantum theory the probability is defined on the elementary events which are subspaces, as opposed to subsets in the classical case. What is important is that there is famous theorem, Gleason's Theorem[18], that ties the algebraic structure (the lattice of subspaces, and operators) in an interesting, and useful, way to the probability structure on the subspaces.

The theorem goes as follows.

Theorem[6]

Let L be the lattice of subspaces on Hilbert space H. Given a probability function on a space of dimension ≥ 3 there is an Hermitian, non negative operator W on H, whose trace is unity, such that $P(\mathbf{x}) = \langle x \mid Wx \rangle = tr(WP_x)$ for all atoms $\mathbf{x} \in L$, where $\langle . | . \rangle$ is the inner product, and x is a unit vector along **x**. In particular, if some $\mathbf{x}_0$ satisfies $P(\mathbf{x}_0) = 1$ then $P(\mathbf{x}) = |\langle x_0 | x \rangle|^2$ for all $\mathbf{x} \in L$.

This theorem is a slightly modified version taken from Pitowsky[19] which suits our purpose here. A more abstract version of the Theorem can be found in a number of places [20][21]. The meaning of this theorem is rather simple. Firstly it shows that given an operator such as W, which is known as a *density operator*, will generate a probability function on the lattice of subspaces. Conversely given any probability function then it defines uniquely a W which through an inner product is equal to that function. In

---

[5] or, images and features, if one is working in image processing, or pattern recognition.
[6] Atoms are nonzero elements in a lattice such that no other element lies between them and the lattice minimum.

other words, the theorem tells you that the probability function defined in that way is all there is, any function violating the definition will not be a probability function. This makes it a very powerful theorem. The special case at the end of the theorem shows how simple it really is. If you are certain about a 1-dimensional subspace then the probability of any other 1-dimensional subspace (in a real Hilbert space) is the square of the cosine of the angle between the two subspaces: $\cos^2 \theta$, where θ is the angle between $\mathbf{x}_0$ and $\mathbf{x}$. Many important mathematical consequences follow from this theorem in quantum mechanics, for example, no hidden variable theories[22]. For us in IR the theorem lays down an algorithm for consistently computing a probability function on the lattice of subspaces which we use in retrieval.

Applications

I will now give three brief applications of the theorem, or perhaps more accurately three illustrations of how the theorem might be used.

a. *Clustering*. In many applications where it is required to search a clustering it often becomes necessary to match a document or a query against an individual cluster. This can only be done if there is some sensible description or representative of the cluster. Traditionally these representatives are derived from the members of the cluster. The proposal here is to construct this representative with the aid of a density operator W. If the cluster has k elements, each one a vector, and P_i is the projector onto the *i*th vector. Any linear combination of projectors, for example, $\lambda_1 P_1 + ... + \lambda_k P_k$, where $\sum \lambda_i = 1$, will construct a density operator W. The λ_i are to be chosen appropriately. This operator is now a representative for the cluster and using $tr(WP_x)$ will give us algorithm for calculating the probability of any **x** in the context of that cluster.

b. *Conditionals*. In Kinematics the Stalnaker conditional plays an important role. For example in using the following rule of inference, $A, A \rightarrow B \mid\!= B$, the proposition A is often uncertain, and $A \rightarrow B$ too, but when the latter conditional is non-classical its status as a proposition needs to be justified. It turns out that in the logic of subspaces one can define a conditional which is a Stalnaker conditional[23], and hence is evaluated using a possible world

semantics. More importantly in the lattice of subspaces that conditional corresponds to a subspace itself. This means that one can calculate the probability of $A \rightarrow B$ via Gleason's Theorem with respect to any density operator. One uses here a slightly more general version of the Theorem: $\mu(A \rightarrow B) = tr(WP_{A \rightarrow B})$. The reverse calculation, when $\mu(A \rightarrow B)$ is known, will give us a W to be used in further computations.

c. *Relevance feedback*. In IR relevance feedback is one of the most important techniques for improving retrieval performance. It involves a user giving relevance assessments of documents that have been retrieved thus far in a session; the system then learns from these assessments to improve its retrieval. There is a version of this called 'pseudo relevance feedback', in which the user is not directly involved; it is assumed that the top k documents are assumed relevant. It is now easy to see how this feedback process could be formulated with the use of Gleason's Theorem. One takes these k documents and associates with each one weight λ_i proportional to the value of the matching function that was used to retrieve it. One then constructs a density operator with the λ_i's scaled to sum to one. One can now calculate the probability of any document with respect to this density operator, and repeat the process indefinitely.

Conclusion.

I have shown how geometry has influenced the construction of several IR models, and how it can be exploited. The culmination of this story is the illustration of how Gleason's Theorem naturally arises in a vector space in which subspaces play a significant logical role. Gleason's theorem captures nicely how to construct probabilities in a vector space consistent with its geometry and logic. The three applications at the end show how elegant the application of the theorem can be in the analysis of particular IR techniques.

References

[1] Fairthorne, R. A. (1958) Automatic retrieval of recorded information. *The Computer Journal*, **1**, 36–41.

[2] Fairthorne, R.A. (1958) Delegation of classification. *American Documentation*, **9,** 159-64.

[3] Salton, G. (1968) *Automatic Information Organization and Retrieval*, Mc Graw-Hill, New York.

[4]Sparck Jones, K. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**, 11-21.

[5]Luhn, H. P. (1957) A Statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development*, **1**, 309-317.

[6]Maron, M. E. and Kuhns, J. L. (1960) On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, **7**, 216-244.

[7]van Rijsbegren, C. J. (2005) The emergence of probabilistic accounts of information retrieval, In Tait, J. (ed) *Charting a New Course: Natural Processing and Information Retrieval. Essays in honour of Karen Sparck Jones*, 23-38, Springer, Dordrecht.

[8]Tversky, A. (1977) Features of similarity. *Psychological Review*, **84**, 327-52.

[9]Needham, R. M. (1961) The application of digital computers to classification and grouping, PhD Thesis, University of Cambridge.

[10] Sparck Jones, K. (1971) *Automatic Keyword Classification for Information Retrieval*, Butterworths, London.

[11] Cooper, W. S. (1988) Getting beyond Boole. *Information Processing and Management*, 24, 243-48

[12]Crestani, F. and van Rijsbergen, C.J. (1998) A study of probability kinematics in information retrieval. *ACM transactions on Information Systems*, **16**, 225-55.

[13] Deerwester, S., Dumais S. T., et al. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, **4**, 391–407.

[14]van Rijsbergen, C.J. (2004) *The Geometry of Information Retrieval*, CUP, Cambridge.

[15] Birkhoff, G. and J. von Neumann (1936). "The logic of Quantum Mechanics." *Annals of Mathematics*. **37**, 823–843.

[16] Halmos, P. R. (1958). *Finite-Dimensional Vector Spaces*, D. van Nostrand Company, Inc., Princeton.

[17] Pitowsky, I. (1989). *Quantum Probability – Quantum Logic*, Springer-Verlag, Heidelberg.

[18] Gleason, A. M. (1957). "Measures on the closed subspaces of a Hilbert Space." *Journal of Mathematics and Mechanics*. **6**: 885–893.

[19] Pitowsky, I. (2006) Quantum mechanics as a theory of probability. In Demopoulos, W. and Pitowsky, I. (eds) *Physical Theory and it Interpretation: Essays in honour of Jeffrey Bub*. 213-39, Springer, Dordecht.

[20] Hughes, R. I. G. (1989). *The Structure and Interpretation of Quantum Mechanics*, Harvard University Press, Cambridge.

[21] Parthasarathy, K. R. (1992). *An Introduction to Quantum Stochastic Calculus*, Birkhauser Verlag, Basel.

[22] Bub, J. (1997). *Interpreting the Quantum World*, Cambridge University Press, Cambridge.

[23] Hardegree, G. M. (1976). The Conditional in Quantum Logic. In Suppes, P. (ed) *Logic and Probability in Quantum Mechanics*, 55-72, D. Reidel Publishing Company, Dordrecht.