Inf Retrieval (2007) 10:485–489
DOI 10.1007/s10791-007-9026-8

BOOK REVIEW

Keith van Rijsbergen, The Geometry of Information Retrieval

Cambridge. 2005, 119pp + Bibiliography

Paul B. Kantor

Published online: 1 May 2007

Perhaps with a nod to Galileo (1914) Keith van Rijsbergen, a noted authority, and the academic progenitor of an amazing number of important contributors to the field of Information Retrieval, begins this work with a delightful prologue. It is a ''discussion'' between himself (''K'') and two other individuals identified only as ''B'' and ''N'', who are both skeptical about his ideas. Since he wields the pen, he is able to move them some distance during the discussion, but is not able to persuade them completely. I compliment the author on capturing much of the speech and manner of one of the leaders of our field, who responds to the argument with brief, pointed and practical questions. I have scratched my head over the identity of the other participant. There are two individuals who come to mind, but neither of them has, in my own experience, been willing to make his responses as short as those presented here, nor would either of them end a discussion of this type by proposing to ''have some *coffee*''.

Briefly, the several chapters of the book are presentations at various levels of detail of some ideas from set theory, some ideas on vector and Hilbert spaces, a discussion of linear transformations, a discussion of conditional logic, and the proposed translation of the Hilbert space concepts into a discussion of a number of key issues in information retrieval.

The title ''The Geometry of Information Retrieval'' is itself an intriguing and attractive choice. It suggests a number of ideas. One is the idea that documents themselves can be accurately represented as points in some geometric space. Another is the idea that queries or information needs can be represented as points in some geometric space. And it might be the case that there is a natural mapping between the two (this is the point of view that follows from the use of Hilbert spaces). A third possibility, not explored in this book but as intriguing as either of these, is the notion that information retrieval systems themselves can be regarded as points in a suitable geometric space.

We focus on the questions examined in this book: the space of documents and the space of *operators representing queries*. Van Rijsbergen makes the correct and very important distinction between those properties which inhere in a document itself (which he calls the

P. B. Kantor (✉)
Rutgers University, New Brunswick, NJ, USA
e-mail: kantor@rutgers.edu

''aboutness'' of the document) and those properties which involve something more than the document (which he refers to as ''relevance''). He also states (page 17) that his key assumption is that the geometry of information space is significant and can be exploited to enhance retrieval. I think what this means, although it is not spelled out in detail, is that the important property of relevance is ''relatively smooth'' in some space in which the documents lie. This idea could be made quite concrete.

Specifically, in linear vector spaces with an inner product, one has a definite concept of ''between-ness''. The vector c is strictly (on a line) between points a and b if and only if there exists a real number x between 0 and 1 such that $c = xa + (1-x)b$. More generally, c is ''somewhere between'' certain parallel hyperplanes through a and b if and only if $0 \leq (x-a, b-a) \leq (b-a, b-a)$ *where* (x, y) *is the inner product of* x *and* y. But does betweenness have anything to do with ''relevance'' to a particular user with a particular query? The question has not been systematically investigated. Should documents ''between a and b'' have relevance that is ''between that of a and that of b''? What if it is strongly between them? And so on. This is not testable when the dimensionality of the space is larger than the number of documents whose relevance is known. But if one accepts a notion of reduced dimensionality (such as Latent Semantic Indexing) it could be tested. To my knowledge no one has yet done this.

The book also draws on another theme. Van Risjsbergen has proposed for some time that the path to a rigorous understanding of the relationship between something like a ''query'' and something like a ''document'' is to be found in some framework of *logic*, in which it is hoped that a document is relevant to a query precisely when the document implies some statement which is one of the possible useful answers to the query. I think this is a quite profound insight, as is the insight with regard to relevance and aboutness. It also is not yet a fully worked out idea.

The title, and, indeed the book, invites us to reflect on whether mathematics must prove as effective in Information Retrieval as it has in, for example, Physics. Starting with Luhn's insight, at once profound and trivial, that when we write about something we are likely to mention its name, and the names of its parts or its processes, ingenious engineers have built powerful structures, which pervade our lives. They lay open the impenetrable files of Government; they provide access to all kinds of documents whoever may post them on the web. But, is there an underlying structure to the mathematics behind all of this? Van Rijsbergen posits that the structure will be a linear vector space over the reals or the complex numbers. One hears often of the ''vector model for information retrieval''. One also often says ''a text may be represented by the vector of its salient attributes''. Let's examine these two interpretations.

The former is shorthand for Salton's model, which indeed treats documents (and perhaps queries) as having, in effect, a ''direction and magnitude''. The key point here is that when documents have a direction, one may speak of the ''angle between them'' and then, having dressed it in the cosinus function, write papers at will. In fact one can do more. The maverick researcher, Damashek, in a controversial paper (1995) proposed that we take the vector concept more seriously, and apply the important ''center-of-mass'' transformation to the vectors representing both documents and queries. This is a natural thing to think of when dealing with physical vectors. It seems, however, not as effective as the complex change-of-scale procedures which we know as ''inverse document frequency and its offspring (such as 'pivoting')''. All of these are simply changes in the metric of the underlying vector space. It remains a linear space over the reals.

The other concept of vector is more familiar to computer scientists. Taken as stated here, it is too broad. That is, we might say that one salient attribute of any document is ''how useful it will be to Paul Kantor on Feb 26th when he decides to do some more background research related to his review of Van Rijsbergen's book on Geometry of Information Retreival''. (This is representative, the full vector lists all the people who will ever search, and all the searches they will ever do). So, by slipping the apparently harmless word ''salient'' into the definition I have concealed all the hard science. Realistically, we should narrow the definition to include only those things that could be determined about a document, by a well-meaning computer program that can see all the characters (and formatting information, and images) in their naturally occurring order (the ''aboutness''). Some of these attributes (like the color of the pictures) do not very naturally lend themselves to representation as ''geometric vectors''. Others, such as the language of the text (that is to say, Urdu, or Tamil, or French, ...), seem not to have any ''geometric interpretation'' at all.

In the vector space of Quantum Mechanics, the vectors represent physical systems, their individual components (those complex numbers) do *not* represent physical quantities. Instead, QM requires a rich structure of Self-Adjoint Linear Operators, which can ''act upon'' these vectors. Some such operator corresponds to every physically interesting observable property (such as position, or 'spin', or electric charge). Every one of these operators defines a set of real numbers, and the results of any specific measurement of the corresponding observable must be a member of that set. Along the way, not surprisingly, the analysis of these operators shows that each of them breaks the space of all possible ''state vectors'' into orthogonal subspaces, corresponding to the possible results of the measurements. Each such subspace is equivalent to a ''projection operator'' which throws away all the parts of a vector, which do not lie in that subspace (as an example of a projection, think of the shadow of a building on the ground).

Can this powerful framework be put to use in the service of IR? Much of the framework, having to do with the special role of ''time'' and the time evolution operators, seems not relevant at all, so let us leave it off the table. What is left are the state vectors, and the operators corresponding to observables. And the challenge is to associate the fundamental concepts of IR with these objects, in a way that ''bakes some bread'' for those seeking more effective IR.

Well, what are the candidates? Documents are pretty fundamental. Are users fundamental (in this sense) or are they merely handles attached to ''queries''? Is relevance a concept, or a relation between entities? Clearly there is a lot of work to be done. All of the current work with the finite dimensional vector space of the vector model uses an important fact. *The set of all linear functions defined on a d-dimensional vector space is, itself, a d-dimensional vector space with a natural correspondence to the original vector space.* This is important because it enables us to jump (without even noticing it) from the idea that a query is a *linear function defining the relevance of documents* to the representation of that query as a vector. Now, in mathematical vector spaces, one may perform an operation called ''adding two vectors''. In some representations (such as the Bag of Words) this can be realized when documents are the objects. The concatenation of two documents has a vector corresponding to the sum of the vectors of the documents. (This is true even with IDF scaling, but not true with pivoting, which takes account of document length). Of course, so does the concatenation of the same documents with their words randomly rearranged.

Does (or should) relevance behave like a linear operator in this case? Probably not. And this leads naturally to the search for another alternative. Van Risjbergen has proposed that

the relevance (or, indeed, the probability of relevance) may instead be like a physical observable in Quantum Mechanics.

In this book, he argues that because relevance does not reside within the object itself, perhaps a good representation is to be found in the measurement theory developed for quantum mechanics. This approach has the further attractive property that while relevance may remain binary, a particular document can have the property of ''having a probability p of having relevance 1''. However, and this is a large ''however'', the key consequence of the quantum mechanical theory of measurement, as demonstrated in something called the Stern-Gerlach experiment, is an odd kind of interference between measurements.

If this interference property (which is called in the quantum mechanical theory ''non-commutativity'') is to translate into information retrieval we would have to find that the following situation holds:

1. A document is presented and its relevance to some question, query, quest or proposition is determined to have a specific value, let us say ''1''.
2. Immediately thereafter its relevance to some other specific issue is determined and, let us say, it is also found to have the value ''1''.
3. A third prompt re-evaluation of the document with respect to the first issue now reveals a *different* result (for example, ''0'').

This is the key odd property of quantum mechanical measurement. The author does not propose any example of how this would occur in information retrieval and I have been unable to conjure one up.

This leaves us in a situation where the proposed QM machinery seems to have as its crucial strength the ability to answer a question that simply does not arise in information retrieval.

From another perspective, perhaps even this framework is, in some ways, not rich enough. The Hilbert space formulation of quantum mechanics requires that queries, whatever they are, and documents, whatever they are, ''live in the same space''. The Hilbert spaces are in fact one among a continuously infinite variety of spaces called the L_p-spaces. Of these spaces, the L_2, or Hilbert space is the only one which is ''self-dual''. That is, it is the only one for which ''linear questions'' can be mapped naturally into the space itself. Indeed, in addition to the infinity of other possibilities corresponding to $p < 2$, one might seriously question whether the important property of relevance can be accurately captured by a linear-space model at all.

To sum up, reaching almost half a century back into the history of physics, I am reminded of the theoretical physicist Marvin (''Murph'') Goldberger, one of the founders of the analytic S-matrix theory and later head of the Institute for Advanced Studies and then California Institute of Technology. When the analytic S-matrix theory was seeking to displace quantum field theory, which was the reigning doctrine of its day, Murph was fond of saying that quantum theory was ''ornamental but useless''(M. L. Goldberger, 1962, personal communication, What he actually said was ''rather like the teats on a brass monkey''). Subsequent events and the astonishing progress of quantum field theory have shown that Murph's judgment, while entertaining, was wrong. I must conclude this review by saying that while I do find this book, at present, ornamental but not useful, I also hope that the next fifty years will prove me wrong.

Reference

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. Science 10 February 1995: 843–848

Dialogues Concerning Two New Sciences, translated by Henry Crew and Alfonso di Salvio, Prometheus Books, 1991. ISBN 0-87975-707-8. The classic source in English, published in 1914