IRFIS 3:
Proc of the
3rd International
Research Forum on
Information Science
T. Henriksen (ed)
Statens Bibliotekskole
Oslo, 1979

Relevance, Retrieval and Document Spaces

S.E. Robertson

Centre for Information Science
The City University
Northampton Square
London EC1V OHB, U.K.

It has long been accepted that, in documentary information retrieval, we can seldom expect fo find a document or documents that match the initial statement of a query precisely in every detail. Some documents may be found which match the query to a greater or lesser degree, and of these some may turn out to be useful and some not.

This inherent vagueness in the process of retrieval renders the idea of a space in which documents and queries are distributed an intuitively attractive one. If we have such a space, then we may look first at those documents which are nearest to the query and then if necessary spread the net wider; we may move the query around if it appears to be biased in one direction; we may identify clusters of closely related documents; and so on.

Indeed, some of the notions discussed in the context of retrieval appear to have some implicit base in the idea of a document space. As suggested above, document clustering is an obvious one, and the notion of a matching function (which measures how closely a document matches a query) seems to be closely related to the notion of distance in a space. Term clustering, too, suggests itself as being associated with the idea of a space - although, as we shall see below, the relation may not be obvious.

So even before we consider any specific definition of a space and distance measure, the idea of a space seems worth exploring. This paper is an attempt to do just that.

A relevance-theoretic argument for a space

If we want to proceed further in this exploration, then we must begin to construct some idea of what the space might look like. I referred above to documents and queries being distributed in space, but this obviously begs some questions, such as whether the documents and queries are located at or identified with points in the space. The following rather more technical argument suggests a specific connection between the idea of a space and that of relevance, and in the process sheds some light on the nature of the space.

I have argued elsewhere (1) that we must postulate the existence of a variable, which I called synthema, underlying relevance. Synthema is seen as a continuous variable relating all documents to a given query: that is, any document is more or less synthematic to the query. I regard relevance (as usually assessed) as a partitioning of this variable into a small number of classes (e.g. highly relevant, partially relevant, non-relevant).

One could argue for such a variable purely on the grounds of experiments on relevance itself, such as those surveyed by Saracevic (2). However, my argument went beyond that. I was concerned to explain some aspects of retrieval test results, particularly the apparent variations in the non-relevant set: one document, even though judged non-relevant to a query, may be quite frequently retrieved by different systems or search strategies, while another is never retrieved. One can explain this variation as a variation in the synthema variable: the two documents have different degrees of synthema to the query, albeit both below the upper threshold of the non-relevant set.

So there are some powerful arguments, both theoretical and empirical, for postulating the existence of a variable such as synthema. The problem with synthema is that it refers only to one particular query, and therefore apparently says nothing that can be used in a retrieval theory, except possibly in the context of relevance feedback. How can we generalise the idea so as to describe the relation between different queries and documents?

One obvious next step is to regard synthema itself (for a particular query) as deriving from an underlying variable which is common to all queries. But because the sets of documents that are judged relevant to different queries differ (but overlap in various complex ways), this underlying variable has to be multi-dimensional. Which of course makes it a space, which I will refer to as theme-space.

Description of theme-space

What can now be said about this hypothetical space that I have conjured up?

(a) The query: The point of the synthema argument was that it was trying to say something about relevance; hence the "query" must be taken as the underlying need, against which relevance is judged. The relation between this underlying need and any expression or formulation of it, in terms of the space, is discussed below.

(b) The document: Similarly, this is taken as the document itself, with a question mark over its relation with any representation.

(c) Points in the space: In the context of synthema, each document has a value of synthema or a point on a synthema line. The query can be associated with the point of maximum synthema. In generalizing to a theme-space, we must again associate documents and queries with points in the space (different queries now have different points).

(d) Synthema: This must now be represented as some form of distance in the space (zero distance being maximum synthema), so that for a particular query, the synthema of each document relates to the distance from the document-point to the query point.

Retrieval

Theme-space, as described above, is a hypothetical variable which we cannot measure directly. How might we proceed to organise our retrieval system on the basis of this idea?

One way (not the only way) would be to construct a space which in some sense approximates to theme-space, a pseudo-theme-space, and use that for retrieval. Points in pseudo-theme-space would now be identified with the representations of queries and documents. Obviously, one would be looking to use an appropriate distance-measure in this pseudo-theme-space as the retrieval criterion: that is, documents would be retrieved (or ranked for retrieval) according to their distances in pseudo-theme-space from the query.

But because pseudo-theme-space is an approximation to theme-space, distance in pseudo-theme-space reflects not only distance in theme-space (i.e. synthema) but also the approximation process. Thus such a ranking must combine the two ranking principles identified by Robertson and Belkin (3), degree of relevance and probability of relevance.

Because of this introduction of probabilistic ideas, it might not be appropriate to use the same form of distance measure in pseudo-theme-space as is postulated for theme-space.

We have now to ask the question: should we identify any document-spaces that are constructed for retrieval systems with the pseudo-theme-space here described, or are there other forms of space which might be used in retrieval? Since the notion of a space is so tied up with the notion of distance in the space, we can ask the same question about distance as a ranking device in retrieval. Robertson and Belkin identify no ranking criterion proposed in the literature other than degree or probability of relevance. So it seems that if a space is to be used at all, then distance in that space must be interpreted either as degree of relevance, or as probability of relevance, or as some combination of the two.

In the argument above, I started with the idea of degree of relevance, and was forced to add that of probability of relevance at the moment of making the space concrete. There remains the possibility that a space could be constructed relating only to probability. As far as I know, no space has been specifically constructed in an IR context with this idea in mind. Indeed, one of the reasons for constructing a space is to make use of the formal structure imposed by the space itself and by the particular distance measure used; but since probability theory has its own formal structure, of a quite rigid kind, it seems that it may well be counter-productive to try to introduce a spacial structure as well.

Probability and relevance judgements

We have introduced the idea of probability between the hypothetical theme-space and the pseudo-theme-space we construct for retrieval. This suggests that, given sufficient feedback, we might be able eventually to improve our pseudo-theme-space until it gets very close to true theme-space. In turn, this suggests that retrieval is perfectable.

One facet of the spatial model already described militates against that: although we may be able to improve the positions of points representing the documents, each new query must start in an approximate position. But then if we performed a retrospective experiment of the sort described by Robertson and Sparck Jones (4) we should be able to find a perfect position for each query, which retrieves all and only relevant documents.

Experience so far suggests that this will not be possible. This may simply be a function of the methods at present available, but I suspect that it is a more fundamental property of retrieval. Is there any form of spatial model which is consistent with this prejudice of mine?

Yes, there is. If we introduce a probabilistic process between synthema or theme-space and the actual relevance judgements by an individual, as described elsewhere (5), then there remains a residual probabilistic property of retrieval which cannot be eliminated by any amount of feedback.

In practice such a change may make little difference to our models: whatever happens, we have some probabilistic problems to contend with between any retrieval act and the relevance judgements which the retrieval system is trying to predict. But the possibility exists, within the spatial framework, of devising a formal scientific test of the following hypothesis against its converse: Retrieval is perfectable.

Note on matching functions

I said above that the notion of a matching function as used in retrieval seems to be related to the notion of distance in a space. We should ask in what sense this relation exists, and whether it applies to any matching function.

Matching functions obviously measure closeness (similarity) rather than distance (dissimilarity), but we can certainly find a transformation that could be used to turn any particular matching function into a measure of dissimilarity or vice versa. But one must then ask whether such a transformation can be found which will make the measure of dissimilarity into a metric in the technical sense. (A metric is the mathematical notion of a distance measure, and has to satisfy certain criteria (6); the space together with its metric is a "metric space".)

The answer is that it depends on the matching function. For example, it can be shown that any measure of dissimilarity deriving from level of coordination cannot be a metric, in that it would not satisfy the triangle inequality (this result is demonstrated in the appendix).

This suggests that, if we wish to proceed with a spatial view of information retrieval, we must either abandon some traditional matching functions, or adopt a notion of a 'space' (and'distance measure') which does not satisfy the technical conditions for a metric space. The latter course may be dangerous, because our intuitive ideas about spaces normally include such conditions as the triangle inequality; we would therefore have to beware of our intuition as well. But one further possible way of looking at traditional matching functions in a spatial framework is suggested below.

SMART space

Perhaps the best-known space in IR that is explicitly considered as such, and certainly the prototype for many ideas in IR that have a spatial base, is the document space used in the SMART system (7). This space is defined by the index terms used to describe the documents; each of the T terms is associated with one dimension of a T-dimensional space. Each document (and each query) occupies a point in space, determined by the presence or absence (or weights) of each index term in the document or query.

How does the SMART space relate to the ideas discussed above? It has strong similarities to theme-space, in that documents and queries are represented as points, and the assumption is that the documents most relevant to a query will be those nearest to it (indeed, this is remarked

on as an observed effect (8)). But since it is an actualised version which is clearly not precisely correct, and in which the positions of queries and/or documents may be improved (9), it must be regarded as a pseudo-theme-space in the sense described above.

We have seen above that we must think about the distance measure as well as the space itself. The obvious distance measure for a T-dimensional space is the Euclidean metric (which is ordinary length if T is 3 or less). But this is not the measure that is used for retrieval in the SMART system - instead, cosine correlation is used. Cosine correlation is a matching function, and is equivalent to a metric (or technically a pseudo-metric); but not to the Euclidean metric.

In fact, the cosine correlation between two points is determined by the angle between the two vectors creating the points in relation to the origin. As a corollary, any two points lying in the same direction (but different distances) from the origin are regarded as equivalent - all vectors can be regarded as being normlized to length 1, so that all points are constrained to lie on a unit sphere, centred at the origin. Thus the metric space is not the ordinary Euclidean space, but the boundary of a sphere, the metric being the angle subtended at the centre. This point demonstrates the importance of specifying the distance measure as well as the space itself.

When one looks at the SMART space in formal mathematical terms, several curious aspects emerge. For example, extensive use is made (for example in document clustering) of the centroid of a group of points, as representing some kind of average point. But the centroid is only an average in terms of the Euclidean metric; since cosine correlation is being used for retrieval, it would seem more appropriate to use an average point based on cosine correlation.

Index terms in SMART

A second aspect of the SMART space that is worth looking at in this way is the relation of index terms to the space.

The index terms are not identified with points in the same sense that documents or queries are identified as points, but instead are used to define the dimensions of the space. We can, somewhat artificially, associate each term with the point representing a hypothetical document or query which contains that term and no other. But now we have the situation that the different terms are regarded as strictly unrelated to each other: that is, the point (vector) associated with term A and that associated with term B have zero closeness if A and B are different. This is so whether or not there is any semantic or statistical association between A and B. Thus the notion of term clustering does not fit within the spatial framework provided by the SMART space.

This problem is evident in some of the ideas investigated in the SMART system. Consider for example Salton's Theory of Indexing (10). Here each term is examined for the contribution it makes to the separation of points in the space. This is achieved by removing the term as a dimension of the space and comparing the separation of the points before and after the operation.

This operation (of removing the term as a dimension of the space) is a slightly curious operation in spatial terms. If we were dealing with a Euclidean space (that is, with the Euclidean metric), then one would describe it simply as a projection; but because we are dealing in effect wit the surface of a T-dimensional sphere, the operation is rather more complex.

The theory of indexing is used to decide whether certain terms should be dropped from the indexing vocabulary. But terms are found to be of varying usefulness; an obvious possibility would be to reduce the effect of the less useful terms (e.g. by weighting), rather than eliminating them altogether. This is, indeed, one of the possibilities suggested (10). But such modification, again, does not fit within the spatial framework provided by the SMART system; it must be grafted on in a somewhat artificial way. More recent work on term importance in individual queries suffers from the same problems (11); indeed, in many instances, the traditional (for SMART) cosine correlation is abandoned, and the simpler level of coordination is used instead.

How to get round these problems? One might imagine a pseudo-theme-space in which the terms, as well as queries and documents, are identified with points which may be moved around given certain kinds of feedback. What is not obvious then is what the basic dimensions of the space should be; but for some simple kinds of 'space', this problem may not arise.

Graphs

One kind of structure which has been used in retrieval (e.g. by Oddy (12)) is a graph structure: that is, nodes connected by links. Although this is not a space in the usual sense, it has similarities which are worth exploring. Each document and each term is a node in this structure; index terms assigned to a document are represented by links between the corresponding nodes. The system is designed for interactive retrieval: so the query is not tied to a particular node, but is associated with a region of the graph which may change with feedback.

Thus the graph structure does, as suggested above, treat terms as "points" (i.e. nodes) in the same way that it treats documents. Of course this structure, while in some ways freer and less constrained than a space, is also more restricted in that nodes are either connected or not (at least at the lowest level), and degrees of connection are not allowed for (though they could perhaps be incorporated without too much trouble). But there are obviously strong associations between the idea of a graph and that of a space, which may be worth exploring in the search for a suitable structure.

The way in which the query is associated with a region of the graph rather than with a single point is of interest, and at first sight looks very different from the point-queries of (say) the SMART system. However, in the latter case the point serves as a focus for a region (defined in terms of the distance function) in any act of retrieval. In general, one might say that a focus-point is the simplest way of defining a region in a space with a suitable distance measure, but that the region it defines is of a rather restricted type (e.g. a sphere in Euclidean space).

Considering our general notion of a pseudo-theme-space in this light, we might therefore wish to replace the point-query with a region (and thus avoid the restriction). But then the question arises: should documents and/or terms be treated in the same way? There is something

to be said for this, particularly in the case of documents. Unfortunately, the more one throws away restrictions of this kind which are suggested by the structure of the space, the less this structure actually helps us formulate the problem (which is of course the object of introducing the structure in the first place). So I will continue to assume point-queries and point-documents, accepting that we may have to modify or abandon this notion at a later date.

Knowledge spaces

Meincke and Atherton (13) propose a notion of a "knowledge space" which has more in common with my theme-space than with pseudo-theme-space, in that it is an idealised way of looking at the organisation of information. The dimensions of the space are defined by a set of basic concepts, which have to be independent (in some semantic sense) of one another. Then (a) other concepts, (b) documents, (c) states of knowledge of individuals are all vectors in this space. Each query, on the other hand, is identified with a 'search volume' which is a region of the space, not defined by a focus-point but by a choice of interval on each dimension; thus the region takes the form, in Euclidean space, of a rectangular solid rather than a sphere.

The length of vectors in the Meincke and Atherton model (unlike SMART space) are of some importance; more sophisticated concepts are envisaged as being further away from the origin than simple ideas. As for the state-of- knowledge vector, this is longer insofar as the individual's understanding of the basic concepts is "deeper".

The Meincke and Atherton space clearly overcomes the problem of the relationships between terms, discussed above in connection with the SMART space, in that two different concepts are allowed to be related to each other. But the importance of the lengths of vectors and the representation of states of knowledge by vectors seem to present anomalies. For example, the model apparently implies that if an individual deepens his/her understanding of two basic concepts A and B, then s/he automatically deepens his/her understanding of any concepts that derive from A and B - which surely cannot be the case.

It seems to me that the objective of representing any individual's state of knowledge by a single point (as well as documents and single concepts) is over-ambitious. Indeed, it does not seem compatible with some of their initial comments which give rise to the idea. They quote from Miller (14):

> "There is a user who has a system of concepts, and there is an information store that also has a system of concepts. When the user discovers a gap that he needs to fill in his own system, he formulates a question about it..."

This would imply that the user's state of knowledge includes some image of the structure of the space, rather than being represented by a point in it. This idea is reinforced if we replace Miller's "gap" with the more general notion of "anomaly" (15) - except that we then have to allow that the user's image of the structure may not only be incomplete, it may also be inaccurate.

So if we are to include the user's state of knowledge in these spatial ideas that we are trying to develop, it seems that it should be regarded as an (incomplete and inaccurate) image of the structure of the space, rather than as a point in it. There are obvious connections here with

the work of Belkin, Brooks and Oddy, reported earlier at this forum.

Retrieval from knowledge spaces

Meincke and Atherton's model is oriented towards information retrieval; McGill (16,17) conducts some retrieval experiments based on knowledge-space ideas, which are suggested by Meincke and Atherton's work.

McGill's work is directed at the question of the dimensionality of the space in which searching takes place. The idea is that the query is normally specified by rather few terms, which suggests that the dimensionality of the document space has to be reduced by projection before searching. (There is a question not considered by McGill: whether terms absent from the query should be regarded as having unspecified weights, or zero weights, in the query. The former implies that the dimensionality of the query is reduced; the latter does not.) McGill considers various projection methods, and tests the most obvious one (orthographic projection) using SMART-type methods.

The tests indicate that orthographic projections are not useful in retrieval. Unfortunately, some aspects of McGill's method make this result difficult to interpret. For example, we have already seen how one must consider the distance measure together with the space when discussing operations upon the space. But both Meincke and Atherton and McGill implicitly use ideas which are appropriate to Euclidean space (with the usual Euclidean metric), but not necessarily to SMART space with its cosine correlation.

However, the basic idea of manipulating the space in some way (e.g. by projection) in relation to a given query, before searching, is an important one. It is in effect a means of defining a region for the query which still uses a focus-point, but is more flexible than the simple spherical region. It should be pointed out that some existing matching functions (e.g. level of coordination) effectively do the same thing, by simply ignoring the presence or absence of non-query terms. Indeed, this may be a way of reconciling some traditional matching functions with the idea of a space. We saw above that some matching functions are not equivalent to proper distance measures; but we may be able to interpret any given matching function as equivalent to a projection or other transformation of the space followed by the use of a distance measure.

Conclusions

It seems that spatial ideas in IR are pervasive but ill-formulated. We may soon be in a position to formulate a reasonably complete and coherent spatial model of IR, but we certainly have not done so yet.

As an alternative to upgrading the idea of a space to a formal model, we might downgrade it to the status of an analogy. But analogies are notoriously dangerous, particularly when one seeks to derive mathematical techniques from them.

As a middle course, I suggest the careful exploration of the implications of spatial ideas in IR. One way of doing this is to construct IR systems based on these ideas and test them, as has been done in the past; but this is somewhat clumsy and roundabout. An alternative is to simulate spatial models and test the statistical predictions of the simulation against observed data. I know of only one such simulation so far: Jose Griffiths (18) fairly effectively demolished some rather

simple-minded spatial ideas that I had suggested. Testing by simulation seems to me to be useful and worth pursuing further.

References

1. ROBERTSON, S.E. The role of theory in the testing of IR systems. In: JONES, K.P. and HORSNELL, V. (Eds), Informatics 3. London Aslib, 1978 (p114-123).

2. SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science. J.A.S.I.S., 1975, 26 321-343.

3. ROBERTSON, S.E. and BELKIN, N.J. Ranking in principle. J.Doc., 1978, 34, 93-100.

4. ROBERTSON, S.E. and SPARCK JONES, K. Relevance weighting of search terms. J.A.S.I.S., 1976, 27, 129-146.

5. ROBERTSON, S.E. The probabilistic character of relevance. Inf.Proc. Man. 1977, 13, 247-251.

6. Many textbooks on topology or mathematical analysis contain a discussion of metric spaces. See e.g. KELLEY, J.L. General Topology. Princeton, N.J., Van Nostrand, 1955.

7. SALTON, G. The SMART retrieval system: experiments in automatic document processing. Englewood Cliffs, N.J., Prentice Hall, 1971.

8. LESK, M.E. and SALTON, G. Relevance assessments and retrieval system evaluation. In reference 7 (p.506-527).

9. See e.g. Papers on relevance feedbacks and document space modifications in reference 7.

10. SALTON, G. A theory of indexing. Regional Conference Series in Applied Mathematics. Philadelphia, Society for Industrial and Applied Mathematics, 1975.

11. YU, C.T. and SALTON, G. Precision weighting: an effective automatic indexing method. J.A.C.M. 1976 23, 76-88.

12. ODDY, R.N. Information retrieval through man-machine dialogue. J.Doc., 1977, 33, 1-14.

13. MEINCKE, P.P.M. and ATHERTON, P. Knowledge space: a conceptual basis for the organisation of knowledge. J.A.S.I.S., 1976, 27, 18-24.

14. MILLER, G.A. Psychology and information. Am.Doc. 1968, 19, 286-289.

15. BELKIN, N.J. and ROBERTSON, S.E. Information science and the phenomenon of information. J.A.S.I.S., 1976, 27, 197-204.

16. McGILL, M.J. Knowledge and information spaces: implications for retrieval systems. J.A.S.I.S., 1976, 27, 205-210.

17. McGILL, M.J. Projections withing knowledge spaces: the implications

for information storage and retrieval systems. Proceedings of the A.S.I.S. 1975, 12, 138-140.

18. GRIFFITHS, J.-M. The computer simulation of information retrieval systems. Ph.D. Thesis, University of London, 1977.

Appendix: Level of coordination is not equivalent to a metric

Suppose that we have a space with points q representing queries, d representing documents, and that λ (q, d) is the level of coordination of d with q (that is, the number of terms from q that occur as index terms in d). Suppose further that there does exist a metric or pseudo-metric m on the space (6), and a transformation f relating λ and m:

$$\lambda (q, d) = f (m(q, d))$$

$$m (q, d) = f^{-1} (\lambda(q, d))$$

Then perfect match must correspond to zero distance, i.e.

$$0 = f^{-1}(\lambda(q, q))$$

But λ also implies perfect match where the document contains all the terms of the query and others: i.e.

$$q \subseteq d \Rightarrow m(q, d) = f^{-1} (\lambda(q, d)) = 0$$

Suppose now that we have queries q_1 and q_2 and document d_1 such that

$$q_1 \subset q_2$$

$$q_1 \subset d_1$$

but q_2 contains a term not contained in d_1 Then:

$$m (q_1, q_2) = 0$$

$$m (q_1, d_1) = 0$$

but $$m (q_2, d_1) \neq 0$$

This contradicts the triangle inequality for metrics; hence level of coordination is not equivalent to a metric.