Salton Award Lecture[*]
On theoretical argument in information retrieval

Stephen Robertson

Microsoft Research
St George House
1 Guildhall Street
Cambridge CB2 3NH
email ser@microsoft.com

Preamble

First, let me say how pleased I am to receive this award. I am flattered indeed to be invited to join this – I was going to say "august" company, but *august* is a Latin word, more specifically Roman. I think in deference to our present location, I should perhaps say something different. I don't think the Greeks had an exact equivalent (not having gone in for emperors like the Romans), but perhaps "olympian" will serve the purpose. At any rate, I really appreciate the honour accorded to me by SIGIR and the committee of holders of the award who take the decision.

1 Theory versus pragma

I consider myself a theorist. That is, my inclination is to theoretical argument, to achieving theoretical understanding, in information retrieval as in other realms. To me, understanding is what theory is about; those other attributes of theory, prediction and application, are side-effects only, secondary to the main purpose.

However, I have to admit that the field of information retrieval in which I have chosen to be a theorist is not a very theoretical one. This is true in two senses: in a negative sense, there are few strong theories in IR, and certainly no overall theory *of* IR to which one might appeal to solve all difficulties. In a positive sense, the field is very strongly pragmatic: it is driven by practical problems and considerations and evaluated by practical criteria.

Actually, the pragma of IR comes in two distinct forms. On the one hand, we have commercial pragmatism: IR systems and services operate in the market-

[*] Talk given at SIGIR 2000, Athens, July 2000

place, and stand or fall by market forces – customer satisfaction, willingness to pay, competition *etc.* On the other hand, we have technological pragmatism: we design systems to perform certain tasks, and provided we have the ways and means to measure success or failure in the performance of these tasks, then we can try out mechanisms and techniques to our heart's content, selecting those that help in the pursuit of performance, and rejecting those that do not. *Why* they work or do not work is a secondary consideration.

Both these pragmatic views of IR are alive and well and living in all sorts of places—though somewhat curiously, they are almost completely disjoint. Commercial pragmatism lives where one might expect, in the twin commercial worlds of systems and services. It has given us the Dialogs and the BRS/Searches and the AltaVistas and the Autonomys, and no doubt has many more equally useful devices to come.

Technological pragmatism lives (slightly less obviously) in academia. In this tradition, we have a well-developed set of methods, rules, procedures and tools for testing IR systems (that set of methods that effectively began with Cranfield and continues with TREC). System designers may draw upon theoretical arguments, or techniques from other fields, or experience, or instinct, in choosing what to try; but whatever the original reason, any part of the system stands or falls by its effect on performance as measured by an appropriate combination of evaluation measures such as recall and precision. This tradition has given us ranking, relevance feedback, certain statistical methods, certain linguistic ditto, *etc.*

Between these two pragmatic traditions and its own shortcomings, theory tends to be the poor relation. Despite my opening claim to be a theorist, I sometimes find myself being apologetic about theoretical arguments: *e.g.* "I have this technique for relevance feedback. I can demonstrate that it is effective in retrieval performance terms on all the best test collections; I can also show you that at least under some circumstances users like it and use it; and I can also show you that it is a feasible practical device to incorporate into a working system. Oh, and by the way, it does have some theoretical basis."

2 Hidden theories

Given this apparent almost total domination of pragmatism, can I mount a credible defence for theory and theorising?

Actually, the apparent dominance of pragmatism is a little misleading. "Words are theories about objects"[1] – there are theories embedded in the language and in the words that we use to discuss retrieval (or anything else). An extreme example of this fact appears when we use the word 'relevance' in information retrieval. The word tends to carry with it a whole host of connotations, almost an entire view of the universe. Unfortunately in this particular case every user of the word appears to have his or her own view of the universe, different from everybody else's!

[1] Colin Blakemore, Reith Lectures, 1976

Let's take two more, preferably somewhat less contentious, examples: precision devices, and the relation of filtering to adhoc retrieval.

3 Precision devices

You sometimes hear the use of phrases (as opposed to single words) in IR described as a precision device, or something intended to enhance precision. [I believe the phrase 'precision device' is due to a previous Salton Award winner, Cyril Cleverdon.] At some level, it's an obvious idea, which maybe doesn't need further analysis (aside from the pragmatic task of discovering whether it actually does do so). However, I would claim that the idea would benefit from a theoretical analysis. So I propose to spend a little time on this idea and some of its components, trying to tease out some of the theoretical constructs and presuppositions inherent in the description. I will be approaching this analysis rather obliquely; those among you who are impatient with theorising may find yourselves out of sympathy with this paper. I can only apologise and suggest you go and have a swim for half an hour!

3.1 M-factors

Suppose that we have a system and a collection and a query and a retrieval rule – one that gives us a retrieved set of documents, not a ranked list: it could be a scoring or best-match system coupled with an explicit threshold, or (say) a Boolean system. Suppose further that we are considering making a single change, either to the system, or to the retrieval rule (in general or in this instance, or to the instance of the query). How could we analyse the effects of such a change?

Some candidate changes are specifically intended to expand the retrieved set. Thus for example if we move from a system which looks for exact matches on specific words to one which allows a match if the stems of the two words are the same, but ignores suffixes, then we can expect that we will get more documents matching. This is a logical expectation, not a statistical one, in the sense that the direction of change is determinate, though the amount of change is not. Such a change can be labelled a recall device, for the following obvious reason: that if some documents in the collection are relevant and some are not, and if there is a finite (non- zero) probability that some of the documents newly brought into the retrieved set are relevant, then recall will increase. And despite my use of the word 'probability' there, the *direction* of change is again determinate – recall cannot be reduced by such a change.

(In parenthesis here I will just observe that there are of course many assumptions, deserving of theoretical analysis, embedded in my use of the words 'relevant' and 'recall' and perhaps even 'document', and in the sentences around them in what I have just said. I will however skip over these problem areas for the purposes of this paper.)

What about the other direction: a change designed to restrict the retrieved set? (For example, requiring a match on a phrase rather than on the constituent words.) In a paper published in JASIS in 1975[2], I discussed both cases at some length. I found it easier to frame the discussion and to draw conclusions in terms of recall and fallout rather than recall and precision: then the opposite of a recall device is a fallout device, *i.e.* a device intended to restrict the retrieved set and thus to reduce fallout, or the number of non-relevant documents retrieved. However, the argument can be converted back into precision terms: there are good common-sense reasons why a fallout device might be expected to increase precision, just as there are good common-sense reasons why a recall device might be expected to reduce precision. This would then justify the use of the term 'precision device' rather than fallout device, although the relationship here is much more clearly statistical: there is no *logical* reason for such a device to change precision in one direction only.

For such single changes (either way: either to expand or to restrict the retrieved set), I coined the phrase 'M-factors'. The problems arise when it comes to the use of two or more M-factors simultaneously. If we allow two such variables to change at once, one in one direction and the second in the other direction, the effects on either recall or precision are much less predictable: such a combination of changes could quite plausibly increase or reduce both recall and precision simultaneously.

3.2 Ranked output

In scoring or best-match systems, the threshold itself is an M-factor. Furthermore, it is generally an implicit variable (because such systems tend to ignore absolute scores once they have determined document ranking). A threshold expressed explicitly as a number of documents (rather than an absolute score) suffers from the same problem. In these circumstances the effect of any other M-factor is bound to be confounded with the effect of the hidden threshold changes. To make matters worse, the scoring itself may induce some degree of confounding, if for example it includes a document or query length normalisation which is affected by the device in question.

How would we identify a precision-enhancing change in a ranked-output system? In the TREC tradition, this is taken to mean a device which enhances precision at low document cut-off, say 5 or 10 (or even zero, since the analysis method normally used extrapolates to a 'precision at zero' figure). Taking a 5-document cut-off as example, we can pose the question as "does the use of this device increase precision at 5 documents?" But since the effective threshold score which gives five documents is different before and after using the device, the question already mixes at least two M-factors.

To make a more direct statement: the concept of a precision-enhancing device has one meaning in the context of set-based retrieval, and another and

[2] S.E. Robertson, Explicit and implicit variables in information retrieval systems. Journal of the American Society for Information Science 26: 214–22 (1975).

quite different meaning in the context of ranked-output retrieval. The 'same' device (such as 'using phrases') might very well be a precision device in one context and not in the other. The term *precision device* itself was coined in the former context; whether it is a valid concept for the latter is not obvious.

To put it even more directly, in accordance with the status of old IR hand which I suppose the Salton award means I must have acquired: Precision devices aren't what they used to be!

I will not pretend that this is an earth-shattering revelation, or that other researchers have not reached the same insight, perhaps by other means (possibly including experiment!). Nor will I claim that this theoretical argument which I have just used would justify being called a capital-T Theory. I will, however, claim that it shows that insights *can* be obtained from theoretical argument alone.

Just to pursue the same line of argument a little further, recall devices are also problematic, despite their logical (rather than statistical) status in the set-retrieval context. Again, in the TREC tradition, we tend to measure recall at some arbitrary large cutoff (say 1000 documents). This immediately destroys any claim to logical status for a recall- enhancing device. Even if we do something which (logically speaking) can only increase the size of the retrieved set, such as expanding the query with a lot of synonyms, it might still reduce recall at 1000 documents.

3.3 Phrases

Given that we are now generally working in the ranked-output context, it has become much more difficult to see whether the use of phrases might be regarded as a precision device. We certainly now lack an *a priori* reason to consider it as such, or even as a candidate. It might, indeed, turn out to be a recall-enhancing device and not a precision ditto, or both, or neither; and this may depend on the scoring method used. Furthermore, the theoretical argument forces us to take one of the two pragmatic views of this: it is a precision device if and only if it enhances precision, either in formal experiments such as TREC, or in the informal experience of search-engine providers and their users. (In parenthesis, being an academic myself, despite my present affiliation, I would tend to rate the former above the latter.)

Either pragmatic view might look like a slightly depressing conclusion for a theorist. However, all it really means is that we have to carry the theoretical argument to a somewhat higher level of abstraction, and not rely on such simple but unjustifiable notions as that of 'precision device'.

4 The logic of filtering

I tend to espouse, and to be associated with, the probabilistic approach to IR and specific probabilistic models for IR. Actually, my theoretical bent is to a combination of logic and probability. (A hasty disclaimer at this point: this does *not* imply that I adopt Keith van Rijsbergen's logical model of IR, in which relevance is identified with the probability that the document implies the query. My own logic is at the same time much more basic and much more elementary than Keith's!)

I would claim that there is still substantial scope for insight from theoretical arguments based on elementary logic, perhaps combined with moderately elementary probabilistic or statistical ideas but perhaps on its own. The argument I have just made about precision and recall devices was one such; here is another example. There is a frequently-cited paper by Bruce Croft and Nick Belkin[3], on retrieval and filtering as "two sides of the same coin". There is indeed a useful logical symmetry between the two, which may be exploited to good effect – but the symmetry is by no means absolute. (I have discussed the general concept of symmetry and duality in a *Journal of Documentation paper*[4] in 1994, and some aspects relating to filtering are reflected in another more recent one[5]). What follows is an abbreviated analysis of both the symmetry argument and the counter-arguments, and some indication of their implications.

4.1 The duality between adhoc retrieval and filtering

The basic symmetry/duality argument arises if we assume that in some sense, documents and queries are similar kinds of objects, or that they are at some level interchangeable. Given such an assumption, then any statement we make about documents and queries has a dual statement in which the roles of documents and queries are interchanged. Some statements are self-dual: that is, the interchange will leave the *sense* of the statement unchanged. Others have duals which mean quite different things – indeed the dual may be incompatible with or contradict the original.

By 'statement' here I mean to include theories or models, empirical observations, system or function descriptions, *etc.*.. In the present context, the following pair of dual statements indicate the relationship between adhoc retrieval and filtering:

- We maintain a collection of documents. When a new query comes along, we search the collection, and identify appropriate documents for this query.
- We maintain a collection of queries. When a new document comes along, we search the collection, and identify appropriate queries for this document.

I was rather careful to formulate this in such a way that the two dual statements both make reasonable sense. The fact that I had to take care in this way has

[3] N. J. Belkin, W. B. Croft, Information Filtering and Information Retrieval: Two Sides of the Same Coin? CACM 35: 29–38 (1992)

[4] S.E. Robertson, Query- document symmetry and dual models. Journal of Documentation 50: 233–238 (1994)

[5] S. Robertson and S. Walker, Threshold setting in adaptive filtering. Journal of Documentation 56: 312–331 (2000)

to do in part with accidents of language, so that some things which are really dual are typically expressed in non-dual fashion – an example is that I would probably use the word 'profile' rather than query in the filtering context (and both are really shorthand for users with anomalous states of knowledge). But the difficulty also has to do with real asymmetries. Examples will follow.

If we were to take the second statement as a complete description of the filtering function, and the first as a complete description of adhoc retrieval, then it would follow that we could use exactly the same system for both, by simply interchanging the roles of documents and queries. It is probably already obvious to you that this is not going to work – that is, the situation is not as symmetric as these statements suggest. Let us explore some of these asymmetries.

4.2 Counter-arguments to duality

Here is a short list of some of the ways in which the situation may not be symmetrical – that is, reasons against the interchangeability of documents and queries.

- Asymmetries of representation

Some systems do not represent queries and documents in the same way. An obvious example is a Boolean system: a document is a 'bag-of-words' and a query is a Boolean statement. (In parenthesis, it is interesting to consider the dual possibility: a query (profile) as a bag-of-words and a document as a Boolean statement expressing the logical combination of characteristics which might define a query to which it would be relevant; but I've never come across such a system.)

- Asymmetries of ranking

A system which ranks documents in relation to each query must behave differently in the dual case for filtering. Ranking the queries in relation to each document is *not* a lot of use; one would need at the very minimum a decision threshold on the ranking, but in fact a single decision threshold *per document* is hardly likely to satisfy the different requirements of different users as regards the balance of retrieved non-relevant and missed relevant documents.

- Asymmetries of evaluation

The previous point relates to another: the source of evaluation. In the adhoc case, it is always assumed that the user (owner of the query) is the one who must decide if an act of retrieval was good or not. The strict dual would have the author of the *document* evaluating a filtering system. However, although one might imagine some special situations where that would be appropriate, the norm must surely be that it is still the user (owner of the profile) who must evaluate.

- Asymmetries of interaction

Documents tend to be static, while users may at any time provide additional information or interact with the system; perhaps more importantly, they or their interests may change over time. (However, we can envisage situations where documents change too.)

- Asymmetries of history

We would expect, in the filtering case, to maintain (at some level of detail) a history of each profile, including documents judged relevant and other user activity. It's not usual, in the adhoc case, for histories to be associated with documents – although it might indeed be interesting to maintain a history of the queries for which a document has been judged relevant.

- Asymmetries of statistics

There are many quantitative or statistical aspects of documents and queries which do not match (the most obvious one is the length). There are also such aspects which do not look as though they should be treated in a dual fashion. Thus for example many document scoring algorithms for text retrieval use idf, which relates to the document collection; it seems unlikely that in a filtering context it would be appropriate to replace idf by iqf, the inverse frequency in the query collection.

Just to point out an example of how these asymmetries may cause problems, I will look at the first, representation, in the context of the usual mechanisms of Boolean retrieval. If the document is taken as a bag-of- words which we index for adhoc retrieval in the usual inverted-file fashion, then the Boolean structure of the query is accommodated by the usual merge operations on inverted lists. If instead for filtering we index the terms in the (Boolean) queries, there is no obvious equivalent mechanism for the Boolean operations – they certainly have to be done in a different way.

The duality idea is interesting because it prompts these questions. In some cases, it might suggest doing something which no-one has yet tried, and which therefore might be worth trying. In other cases, particularly the evaluation one, it enables us to identify a genuine and fundamental asymmetry, which is interesting in its own right.

5 The nature of theoretical argument

I have been arguing in this paper, not for a Grand capital-T Theory of or about information retrieval, but for a very much lower level in the scheme of things. The structures of the objects or entities that we deal with in IR, or of collections of these objects or entities, the structures of the relationships between them, and the structures of the situations we observe or postulate, all provide us with a level of logical argument which has to be basic to our field. We appeal to these arguments all the time, albeit sometimes unknowingly because the ideas are implicit in the language, and sometimes inconsistently because the language hides that too. 'We' includes the theorists and both kinds of pragmatists.

Despite our lack of a capital-T Theory in IR, we are not lacking in small-t theory. But we do need to be more careful with the low-level theoretical arguments. They are sometimes treated as matters of common sense, and certainly common sense is important; but like all theories they need looking after.

But what about a Grand Theory?

5.1 Grand Theories

Do I believe in some future nirvana in which we *do* have a Grand Theory of IR? Not really. The problem of principle seems to me to lie in the range of different domains such a theory would have to encompass. I would like to illustrate this problem with an analogy. We build bridges – real bridges, over rivers or bays or creeks or chasms or roads or railways. We have had a lot of experience (hundreds, nay thousands of years' worth), and we have various kinds of theories or models which complement that experience. There are models of mechanics which tell us how loads are distributed; we have various ways of understanding the behaviour of different kinds of materials – mechanical behaviour under different stresses, chemical behaviour under the onslaught of the environment, etc. We have models of soil mechanics and hydrodynamics which tell us something about the supports. And so on. Occasionally in our history, even after thousands of years of experience, we have discovered great gaping holes in our understanding – for instance when we discovered (catastrophically) that under some conditions an apparently stable structure can simply shake itself to pieces.

But we don't have (I believe, though someone might like to contradict me here) a Grand Theory (Capital G capital T) of bridge building. A designer or builder of bridges has to juggle these various facets, and decide when a particular aspect needs worrying about or can be safely ignored – on the basis of a combination of experience, understanding and low-level logic of exactly the sort I have been arguing for in IR.

[I should perhaps point out to those who have been in London recently, that I actually wrote these words *before* the Millennium Bridge fiasco. For the rest of you, the Millennium Bridge is a new footbridge over the Thames in London – a slim steel-and-aluminium suspension construction, opened to great fanfare last month. Like all such bridges, it is designed to flex – to move with the forces upon it, rather than resisting them rigidly. And indeed it does – particularly when there are many people on it – to such an extent as to make them seasick! The bridge has now been closed temporarily for alterations.]

So, what are the analogous facets in the case of information retrieval? Well, we have cognitive science, and linguistics, and epistemology or ontology, and probability and statistics, and probably other things. There is some tendency to regard these as alternative ways of looking at IR, but of course they are really complementary. And their complementarity resides precisely in the combination of low-level logic and experience. It seems unlikely that we can find a Grand Theory that will tell us exactly when we should be worrying about the linguistics and when, by contrast, we should take the linguistic entities we have identified at their face-value and treat them as statistical clues. I'm not claiming that such a theory is impossible – just that it's a tall order.

This is not at all to say that the *search* for theory is futile – far from it. I believe that the models we have at present can indeed be extended, by theoretical argument as well as by both kinds of pragmatism, to cover more ground than they do at present and to be more useful as tools. But when I read a paper (as one does, occasionally) which seems to make a claim to represent a Grand Theory, then I shall continue to take it with a pinch of salt. And if I myself should ever seem to make such a claim, you have my full permission, nay encouragement, to do the same to me!