Meta-reflections on TREC

Karen Sparck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, England
sparckjones@cl.cam.ac.uk

This paper in its final form appears in *TREC: Experiment and evaluation in information retrieval*, ed. E.M. Voorhees and D.K. Harman, Cambridge MA: MIT Press, 2005, 421-448.

Abstract

This paper considers the TREC programme in the larger context of IR systems and IR evaluation methodology. It summarises the main features of TREC, and TREC's messages for system design and testing. It then relates TREC to operational systems, first to conventional ones developed before the World Wide Web, and second to Web engines and the wider network-based world. Following a long-standing research tradition, TREC has concentrated on evaluation abstracted from the setup in which an IR system is actually used. This paper asks whether, to make future TREC research pertinent to the new opportunities provided by the Web and networking, it needs not only to do significantly more work on types of retrieval situation it has not hitherto addressed, but also radical change to bring more information management tasks, and user activities, within its scope.

1 Introduction

TREC has been the largest evaluation programme in the Natural Language and Information Processing (NLIP) field, as the collective proceedings, *TREC-1 - TREC-2002* (see *TREC* 1993-), make clear. In previous papers (Sparck Jones 1995, 2000) I reviewed TREC in detail, considering it primarily from the point of view of 'classical', or traditional, information retrieval (IR) concerns and methodologies. Thus my focus there was on what the TREC tests and their results had shown us about indexing approaches, strategies and devices, for example about the best form of indexing language, bearing in mind the generic evaluation framework within which these tests had been done. In general, the TREC evaluation has been within the paradigm established by the Cranfield experiments (Cleverdon 1967) and consolidated by the Cornell SMART research (Salton 1968, 1971, Salton and McGill 1983), of experiments designed to distinguish and determine the effects of specific system *performance factors* by controlled laboratory tests.

The TREC programme has added a very great deal to our understanding of retrieval system behaviour within this bounded framework. In my second review, after TREC-6, I raised the question of whether it was not merely desirable, but necessary, for TREC to extend its horizons. There have been five further completed TREC cycles since then, up to

2002, so the immediate question is whether these have introduced significant changes in the TREC style. There have clearly been new developments within TREC in the last five cycles, which have introduced 'more modern' types of data (i.e. net and Web documents, also video) and 'more current' tasks (i.e. question answering). But these have been handled in the same spirit as before. The aim of this paper is therefore to address the question: Does TREC need a radical shakeup? Should it adopt new goals and move in new directions? If so, what should these goals be, and how should it move towards them?

The reason for asking these questions is quite obvious. In what is, or is perceived to be, the very rapidly changing world of information provision, search and use associated with the astonishing growth of the World Wide Web, it is necessary to raise the issue loud and clear: how relevant is TREC? Thus on the one hand, how far have, should or can TREC findings be exploited for Web-based IR? On the other hand, what are the problems of Web-based IR that a presumably continuing TREC ought to address? The TREC programme has manifestly (and successfully) evolved both by changing task form in detail and by introducing new tasks: can it continue to do this, pertinently and usefully, i.e. tackle whatever IR issues the Web, and networked information systems generally, present?

In its most obvious form, this is a question about relations between TREC and Web engines. Here one possible view is that IR system development and provision on the Web is so fast moving, and so intensely competitive, that whatever is helpful will emerge from the market place itself, give or take a few brave spirits willing to invest in an idea or pick up and apply some accidentally-encountered suggestion, without any need for a prior classical evaluation programme: Google is the obvious example. A well-known economic theory suggests that the Web itself is the perfect environment for IR product evaluation, and certainly much better because more fast moving and global, than any of the older forms of IR service context like libraries or online bibliographic search services.

However there is more to the Web than as a vast and challenging arena for gladiatorial search engines. The Web, and network connections, are the enabling infrastructure for a much larger information world including, for example, bibliographic service sites, and database sites covering a very wide range of information types and supporting users with very varied needs. Thus where Web engines may focus on precision, other services may address recall, and where on Web pages anything goes, quality-wise, with other information bases quality control is paramount. Some of these areas fall under the the 'digital library' heading, but there are other information situations also of potential relevance, for example those associated with corporate intranets, where particular data types, for example financial reports, may loom large. This larger information scene is rapidly developing, and the question is whether and how TREC can connect with it, even whether TREC, with its paradigm of carefully-controlled and hence time-consuming scientific research, could keep up with such a changing world.

Thus in this paper I will concentrate on TREC lessons (or, perhaps, the lack of them) for the future. I will begin in Section 2 with a brief summary, for reference, of the TREC programme itself and its major findings, and note key points about its approach to evaluation. Section 3 considers the messages that TREC conveys for how to do IR and how to evaluate it in rather more detail, in relation to previous research. This leads in Sections 4 and 5 to a review of the strengths and weaknesses of the TREC approach first, and relatively briefly, in the context of conventional types of IR system and *setup*, i.e. computational system plus functional context involving the system's users (Sparck Jones and Galliers 1996); and second, in the new Web-based and -driven IR context. My main focus will be an attempt to characterise key IR properties of the Web world and the broader type of networked setup,

and to assess how far the traditional mode of evaluation to which TREC has subscribed can be applied in the new situation.

Because the Web's page data and search engines are the obvious initial point of entry to this world from TREC, and indeed TREC has already done some Web data evaluation, my initial assessment of TREC's lessons will be from this point of view. But this then leads to consideration of other modern information environments. There have already been comparative reviews of Web engine performance (e.g. Gordon and Pathak 1999), and I would not wish to claim that the Web, as an information management environment, is totally different from any previous one (though many assume it is). But it is clear that it is novel in important ways, and the question is therefore whether and how, traditional IR research concerns and evaluation methodologies, as so successfully applied on a larger and more varied scale than ever before in TREC, can be applied in the new environment. My analysis in Section 5 suggests that the traditional (and TREC) mode of evaluation that keeps users at arms length may have more legitimacy in the hands-on Web world than appears at first sight. This implies, therefore, that TREC's substantive findings should also be examined for their Web applicability. But the analysis also leads to the conclusion that the information seeking and management tasks that TREC addresses should change: it is time to move on, more quickly, from document retrieval as *the* IR research task. Looking further than the immediate comparison with the Web as the primary retrieval context leads to the same general conclusion: TREC has very wide potential pertinence, but much more work is needed, even for the retrieval task, on varied data types and ranges of needs as well as, beyond retrieval, on appropriate forms of other tasks and modes of connection between tasks.

To declare my interests: I have been a member of individual TREC participating teams; and I have been a member of the TREC standing Programme Committee for much of its life. This has been a very exciting and very valuable experience. But I trust that it has not biased my judgement so that the points I make about TREC for the future here are invalid.

I shall refer to individual TREC evaluation cycles thus: TREC-3, and to the corresponding proceeding thus: *TREC-3*; for full details see the bibliography under *TREC* 1993-.

2 Reference summary of the TREC programme

2.1 Design properties

For present purposes the key design features of the TREC programme have been that it has:

1. addressed a range of tasks, essentially of a mainstream IR kind, notably one-off (*adhoc*) searching and filtering but also, especially more recently, tasks that have not hitherto figured in mainstream IR systems, in particular, question answering;

2. concentrated on retrieval from full text of various sorts;

3. worked on a large (document) file scale, intended to be realistic in ensuring that undesirable volume effects are overcome by selective retrieval;

4. encouraged full automation in indexing and searching, though the manual construction of resources like thesauri has always been allowed, and explicit manual (and interactive) searching has normally been a possible option, for good reasons, and very instructively;

5. sought and applied well-founded evaluation protocols in terms of test collection design and data gathering, especially for the necessary *answer data*, i.e. relevance assessments;

6. used a range of formal performance measures chosen to present results from different points of view;

7. (once the programme became established) adopted a test cycle model with annual task specifications aimed at clear-cut and well-controlled experiments in any one cycle, and a cumulative attack on a task area over several cycles;

8. planned for reusable resources and results, to provide a platform for future experiments, both within and outside the programme.

2.2 Contingent features

Since the attractions of the TREC programme rapidly became apparent to the IR research community, and there were no formal barriers to entry and relatively low participation costs (given that data provision, assessment and programme management were centrally funded), TREC became a popular success with many participating teams, in many cases undertaking several tasks. This has had both first and second order advantages for TREC and IR research, in each case stemming from critical mass.

The first order benefits were:

1. more sets of results, so more informative performance comparisons;

2. more contributions to the assessment *pool* so more reliable test data;

3. more strategies and devices explored;

4. more confirmatory consolidation of results, promoting technology convergence (an effect noted in other government-sponsored evaluations e.g. of information extraction: see Cardie 1997).

The second-order benefits were:

1. more thorough discussion of task specifications and evaluation designs by more interested and informed participants;

2. more capacity to sustain evaluation over several cycles, with data variation or fine-grained task modification encouraging a better understanding of task requirements and of appropriate strategies for meeting these;

3. more ability to address more issues, contributing to the evolution of the programme as a whole, initially on the hub-and-spokes model with the specified Adhoc task as hub and others branching out from it, but latterly more as a set of *tracks*, each with their distinctive characteristics and variously related to one another.

Altogether, to summarise these aspects of TREC, the programme has involved increasingly large (and real) data sets. The Adhoc task involved at least 1.5 M regular documents altogether (so even subsets, sometimes used within TREC, or chosen to meet specific needs

outside it, can be substantial). The Web track has used 18.5 M Web pages in some experiments, the Filtering track has included tests with some 870 K documents, the 2002 Cross Language track used 380 K Arabic documents, the 2002 Question Answering track over 1.3 M news stories. The Adhoc task totalled 550 substantial topic requests (though in subsets of of 50), and the Web track, spectacularly, has done tests with 10 K queries drawn from Web engine logs (though only 150 of these had output assessed). There have also been input sets for other user need types including Web home page queries and about 2000 questions to be answered. There have been hundreds of thousands of relevance assessments, often on pools of more than 1000 for an individual request, as well as candidate answer assessments for the Question Answering track and sentence-level relevance and novelty assessments for the Novelty track.

Taking the Adhoc and track specifications as defining distinct *problems*, by 2002 TREC had addressed 18 generic problems, the main Adhoc one 8 times, others on average 3; with 2 official runs per team as the default, there were at least 80 sets of results for an Adhoc cycle, about 20 for other earlier tracks, allowing in the former case for 80 individual performance comparisons for one system against others, or 6400 distinct paired comparisons. The growth in participation in recent years has increased the number of runs, to about 40 per track on average in 2002. The programme has also - a far from trivial point - published all the participants' submitted run performance figures and their papers reporting their work, and has made the test data available for future research.

This is clearly a huge mountain of material effort: it has certainly brought forth more than a mouse. But has it brought forth more than an unknown quantity of prairie dogs?

2.3 Major findings

This paper is explicitly not a detailed review, as would be required to do justice to all the individual TREC tracks. But since there are close relationships between many tracks, primarily because they address the same generic adhoc retrieval task, and since TREC has been imbued with the some common aims, it is possible to list some general findings about IR that emerge from the work as a whole.

These still, however, refer only to the type of evaluation test that has characterised TREC, i.e. they are heavily constrained in relation to the realities of information management at large. The TREC tests are primarily, though not exclusively, about *core* IR system objectives and contextual functionality of an immediate and narrow kind (Sparck Jones 2001). They are also, in general, for the kind of user request that has been taken as the norm for IR systems, namely the *topic* search: finding documents about X (even if the actual formulation may suggest something more like a direct question). This has been routine for the main Adhoc task, and typical of the others, with a particular form in Routing and Filtering. The main exceptions have been home page searching in the Web track and, more importantly the recent topic distillation task in the Web track, the sentence set identification in the Novelty track, and the yet more distinctive requirements represented by the Question Answering track. The Video track introduced in TREC-2001 has also covered other forms of search specification.

Subject to these qualifications, the TREC evaluations have confirmed that:

1. fully automated systems can deliver reasonable retrieval performance;

2. they can do this for full text;

3. they can do it for languages, and documents, and requests, with quite different properties;

4. they can do it with robust, simple strategies;

5. they can do as well as minimal manual searching, though not as well as with heavy-duty manual query development.

In other words TREC appears to endorse, after exhaustive, large experimentation, the modern approach to retrieval, i.e. the approach that is motivated, explicitly or implicitly, by statistical models, that starts from simple natural language terms, that relies on weighting and feedback strategies, and that delivers ranked output. This is the approach that the research community has developed over several decades and, during that period, has consistently advocated in the face of the 'Boolean thesaurus/keyword' approach entrenched in conventional bibliographic search systems, and has now shown can scale up satisfactorily (though TREC has not made significant, direct Boolean/ranking comparisons).

2.4 Evaluation methods

As the foregoing suggests, the TREC evaluation methodology has followed a well-established protocol. It has continued to apply the laboratory experiment paradigm and performance criteria, focused on controlled system comparisons and hard output measures, that have been the mainstays of IR research. That is, in characterising IR systems for evaluation purposes, and in concomitant performance measures, TREC has engaged in:

1. heavy abstraction from system *environments*, so user properties (background, purpose etc etc) are represented only by the facts of their expressed information requests and independent relevance assessments;

2. aggressive reduction of information management and its varied elements to the 'search loop';

3. narrow concentration in performance assessment on precision and recall and their siblings and derivatives, especially Mean Average Precision.

Even with manual searching, the usual style has usually been that of a trained intermediary or deeply committed searcher. The Interactive track has naturally used 'ordinary', or at least pseudo-ordinary (i.e. library school student) users, but still within a relatively controlled laboratory setting and with test designs typically encouraging energetic and extended searching. The Routing and Filtering tracks have presupposed professional needs as starting points. The main exceptions to this model of users as both fairly dedicated and fairly skilled have been those where queries have been straightforwardly derived from Web engine logs, notably in the Web track. However within the Adhoc track increasing attention was also paid to very short queries deemed similar to 'ordinary user' Web engine ones.

3 TREC messages

3.1 TREC and how to do IR

It could be said that TREC has reprised the old research tune to the beat of a bigger drum. The main novelty is that the older claims hold when tremendously scaled up. Thus what

TREC says about how to do IR is what it says about how to build a respectable *core retrieval system*, particularly a general-purpose system where there can be little presumption about user experience or long-term commitment, in relation either to individual search sessions or to repeated or regular usage over a long period of time. However this focus on building robust general-purpose systems also has implications, because of the strategies these systems embody, for many particular specialised applications. Thus the types of term weighting, query formation, and iterative feedback procedures which have become established within TREC are ones that can deliver reasonable outputs even when supplied with rather poor inputs e.g. as requests. They can therefore be expected to work well when supplied with better quality inputs, as indeed tests with the TREC-1 and -2 topics showed.

TREC has, secondly, not merely confirmed previous research lines about how to do automatic indexing and searching, and in ways that can do without significant manual assistance either through the provision of support resources like thesauri or in the search process itself. It has also, primarily through being the first sustained series of IR experiments with full text, developed and honed the statistically-based techniques of earlier research. These are directly usable specific tools (even if they are not always appropriately used in practice).

As mentioned in the introduction, these techniques are the products of lines of work long familar from the Cornell SMART research (Salton 1968, 1971; Salton and McGill 1983; Salton and Buckley 1988), the City University probabilistic model and Okapi system research (Okapi 1997), and the INQUERY system (Croft 2000). General ideas suggested in the 1950s and 1960s (see Stevens' 1965 survey), but which could not be fully evaluated then, have now been significantly tried and tested within TREC not only by the teams just mentioned but also, for instance by Kwok (Kwok 1995; Kwok and Chan 1998). TREC trials of Latent Semantic Analysis (Dumais 1995) and, more recently, so-called "Language Modelling", i.e. Markov modelling, methods drawn from speech research (Ponte and Croft 1998, Miller et al. 1999), fall under the same broad heading. This generic, statistically-based class of methods 'fits' the IR problem and performs in an appropriate, resilient way.

At the same time, and somewhat surprisingly, the TREC experiments have not shown that more refined indexing with complex terms is especially advantageous, even though this might seem necessary for bulk files of full text. The automated natural language processing (NLP) systems required to test this on a large scale have only become available in the last decade. But the NYU/GE results (Perez-Carballo and Strzalkowski 2000) did not do better, compared with much cruder approaches, than their manual analogues in Cranfield (Cleverdon 1967) or automated predecessors in Fagan (1987). Thus insofar as phrases may be of modest use, 'statistical phrases' are good enough (Mitra et al. 1997).

These are all points about natural language directly as the indexing vehicle, with only light normalisation (notably by stemming). The other long-standing research issue has been the relative merits of controlled and natural languages for IR. Many TREC participants have made use of any thesauri or other lexical resources that they have been able to find (e.g. WordNet), though these have not usually been thesauri of the conventional bibliographic kind that supply controlled subject labels. But the heterogenous test collections used for TREC (as much as proprietary rights) have meant that there have been no controlled languages with adequate coverage to apply for any systematic comparative evaluations. In many cases also, thesauri have been exploited for the manual searching option, making it difficult to assess their independent contribution (see e.g. Adi et al. 2000; Mahesh et al. 2000). The TREC tests have shown that reasonable performance can be obtained without conventional types of vocabulary control and search aid, but have not been able to make comparisons between

index language types that have long been of interest.

The strategies most commonly used in TREC, and in particular term weighting, have had another important consequence for comparisons with earlier work. These strategies deliver a genuinely ranked output. Many studies in the past, and especially those related to conventional search services, delivered Boolean outputs. Comparing these different types of output is comparing chalk and cheese, and leads to the difficulties encountered in Salton (1972). Modern approaches to IR are based on good theoretical arguments for ranking, and with very large and full text systems, Boolean query constraints tend to be treated as filters before ranking rather than simple output determiners. TREC here is thus reflecting a more general shift, as most obviously seen in Web engines.

These remarks are for the mainstream adhoc case. The Adhoc test findings have in many cases been paralleled by track results, most obviously where the track task is the same but under other conditions, as for the Spanish and Chinese, Confusion, Database Merge, NLP, Spoken Document Retrieval, Cross-Language, and Web tracks. However there are few if any significant earlier tests for these other conditions with which to compare the TREC results. The main point of note is some (albeit gratified) surprise that methods tried and tested have carried over with little hiccup to the other conditions. Scaling up to the Large Web collection (18.5 M pages) was found less stressful than expected (Hawking 2001).With the High Precision and Very Large Collection tracks the task was the same though the detailed performance measures differed, and the same generalisations apply.

The non-adhoc tracks have therefore been only the Routing/Filtering, Novelty and Question Answering ones and also, in a complicated way, the Interactive one. There have been evaluations of the routing/filtering (i.e. selective dissemination of information) task in the past, e.g. by Barker et al (1972), but nothing on the TREC comparative scale. Getting novel information, rather than only relevant information, has long been recognised as a matter of interest, but is hard to investigate without large files offering multiple relevant documents and a heavy investment in assessment. The nearest connection is probably with the recent Topc Detection and Tracking evaluation programme (see Allan 2002). With question answering, apart from some initial investigations by e.g. O'Connor (1973), there are no precedents for the scale of the TREC comparative evaluations, not even for the cruder task of selective passage retrieval. These tracks have, however, been approached, like the others, with their focus on automation and hence on *system* requirements and behaviour.

The Interactive track has necessarily been different (see *IP&M* 2001). The task is again adhoc searching, but to meet particular requirements e.g. associated with particular forms of request, implying appropriately tailored performance measures. However though human searchers are involved, the whole is under laboratory conditions, not natural usage ones, for instance in having set tasks. This is more constraining than laboratory observation experiments of the kind reported in Beaulieu (1990) and Sullivan et al. (1990), though even these were far from unconstrained. However the new 2003 HARD track is intended to take richer user requirements into account, like the search purpose and desired output genre, and also to allow for sub-document retrieval. (Unfortunately the data obtained by monitoring real user behaviour through logging Web engine searches is a less rich source of information about users than one would wish, even with very large search samples.)

3.2 TREC and how to do IR evaluation

The TREC evaluation methodology, aimed at system performance comparisons under (relatively) careful control and with a high degree of abstraction from the contexts in which systems are actually used, has an obvious motivation in the interest in automation itself: can good quality retrieval performance be obtained with fully automated indexing and searching? The TREC form of evaluation can, moreover, be justified by the fact that, since it is impossible to evaluate IR systems without requests, and hardly helpful to evaluate them without relevance assessments, the key properties of real retrieval situations are preserved by the use of soundly designed and constructed test collections. The TREC collections have been formed with care, to obtain realistic document files and requests, and extensive relevance assessments. With several different collections moreover, and broadly-based relevance pools, the results obtained should be free from hidden biases and usefully general or generalisable.

These are good arguments for the TREC approach. The care about test collections helps to offset the abstraction, and to protect TREC from claims that the emphasis on generic *technology* development, which has been characteristic of DARPA-sponsored programmes in NLIP since the 1980s, has sometimes thrown the task baby out with the specific application waters (Sparck Jones 2001). NLIP is not an end in itself: NLIP systems are for tasks that are (directly or indirectly) of human interest. These tasks have their different distinctive, i.e. *core* characteristics and requirements, so systems on the one hand and the criteria and measures used to evaluate systems on the other have to properly address this core. In the TREC case (for documents, setting aside question answering), the core is clearly covered. Whatever else it should do, an IR system should deliver documents that are relevant rather than ones that are irrelevant, and performance is measured by ability to do this not, indeed, only in one but in several different ways. Thus the information supplied by Mean Average Precision (aka AvP), a single-number measure, can be enriched by, for example, Precision figures at different Document Level output ranks. Indeed the core technology is required not only by the important retrieval task in itself, but also by the need for it to underpin more selective tasks, for example by supplying material from which to draw a resource set, as in the Web topic distillation task, or to extract sentences, as in the Novelty task and full-blown question answering.

The TREC methodology continues that of earlier laboratory research from Cranfield onwards (see Sparck Jones 1981), with some gains from working with much larger test collections. Maintaining this laboratory paradigm can be justified by the nature both of the detailed run results and the broader TREC findings. Though we have learnt a great deal from TREC, we often do not really know what works or fails to work, and why. TREC has repeatedly shown on the one hand that plausible ideas do not work, and on the other that very different strategies and devices deliver similar performance. There is thus plenty of scope for further specific comparative analysis like those promoted by the Query track, and for new experiments in the same style, especially in environment conditions not so far represented by the TREC test collections, that can throw further light on the environment factors determining performance and hence leading to one choice of system strategy rather than another. There is also scope for more detailed analysis in relation to individual document and query properties, as pursued in the 2003 Robust track.

Moreover, though TREC has worked, for these good reasons, within a traditional research evaluation framework, it has not been completely static. It has worked with new types of document, particularly news material and web data, also spoken documents and, most

recently, video, (though the problem of adequate evaluation is well illustrated with the Novelty tests, where sentence recall and precision can only feasibly be computed if all participants started with the same ranked list of documents, rather than from their own varied lists). TREC has also developed specific new evaluation methods, notably for the routing/filtering task and the question answering task. It has carried the analysis of test data and methods itself further, as in Zobel (1998)'s and Voorhees (2000)'s relevance pool and reliability studies. Thus insofar as controlled experiments are intrinsically desirable, TREC has helped to ensure that these are properly done for large datasets, and for NLIP tasks of current concern as well as long-established ones.

It is therefore not difficult to make a case for TREC as a distinguished modern representative if an old and eminent family, adding new lustre to its name. The question is whether TREC is moving in the right circles: what does TREC have to say to operational systems and their real users? How does it relate to information retrieval on the Web? Who cares in the Web context, for instance, about performance measured by computing by Mean Average Precision over 1000 ranks? Again, what does it have to say to users of the rapidly growing specialist information sites, for example genome ones, which cover both databases in the orthodox sense, specialised 'catalogue' records, and conventional text? Further, interpreting "information management" broadly, how important is document or text retrieval compared with other information-seeking tasks like question answering, or in relation to other information processing tasks like summarising? Thus especially, but not only, in the Web context, even if topic-based retrieval is important both as a task in its own right and as a precursor to others like summarising, it may be that we have now learnt enough about retrieval from TREC and need to move on. But in that case, what is the best direction to move in, given that it clearly more sensible to try to apply what we have already learnt from TREC than to simply begin anew on some wholly independent task, however interesting and critical that may be?

4 TREC and operational systems

Conventional bibliographic systems, from DIALOG and ORBIT in the 1960s onwards, have subscribed largely to the controlled language paradigm and have consistently adhered to the Boolean paradigm. Operational systems adopted natural language searching, for titles and abstracts, primarily for practical reasons rather than because research suggested it would work as well as anything more elaborate; and they remained wedded to Boolean searching and hence eschewed weighting as a major element in indexing. Early systems for full text searching, particularly in the legal field, also endorsed the Boolean model, often at substantial performance cost (Blair and Maron 1985). Operational services only began to adopt the ranking model, derived from the previous decades of research, to any noticeable extent during the 1990s (Tenopir and Cahn 1994).

It is not surprising that may be called the 'Boolean thesaurus' model has remained a major force, and even more so the 'boolean (key)word' one. Past research, on small test collections, seemed irrelevant to large operational systems and was essentially ignored. Service organisations like Chemical Abstracts and Inspec deal in vast technical literatures calling for, and getting, skilled and informed searching that can deliver high quality output with these tools. When a thesaurus is lacking, the Boolean keyword natural-language model can be as effective in experienced hands, and also seems the natural strategy for newer services like

ScienceDirect to offer end users. Changing big, entrenched systems is extremely expensive, and systems based on the research paradigm have been new ones, for instance WAIS and now the Web engines. Even with new services, straying from conventional models for estabished literature types, like scientific journals, seems unnecessarily rash.

The scale of the TREC evaluations, compared with those of earlier research, should make the results more pertinent to the conventional services. As the early TREC Adhoc cycles demonstrated, with carefully formulated initial requests, automated system performance using only statistically-based natural language techniques, without invoking thesauri or subject classification schemes, can be very good indeed. It is true that TREC has not addressed retrieval in challenging technical subject areas, with complex specialist terminology, like chemistry, but this has been due to the difficulty of getting the necessary test data. Running TREC Adhoc and Filtering cycles for, say, a biochemistry text collection in a way which would allow some proper comparisons with conventional service operations is thus something potentially worthwhile for future TRECs. In the meantime, the TREC Genomics track introduced in 2003 is a welcome first effort in this area, with a restricted type of query. However the difficulty of obtaining 'mainstream' scientific/technical literature test sets means that it is still, though nearly half a century after modern methods promoting the use of natural language were first adumbrated by pioneers like Luhn (Schultz 1968), impossible to carry out the large-scale performance comparisons between natural language approaches and the use of controlled language subject indexing that continues to play a major role in modern bibliographic services.

The performance synopses over TREC cycles given in Sparck Jones (1999/2000) show very clearly that performance levels decline with request 'quality' (brevity, ill-definition, etc). But this applies as much to conventional services. The TREC results suggest that where users are willing to supply initial requests formulated with moderate care, the research model may be applied to conventional services. The natural route for this, already adopted by Web engines, is to use some Boolean constraints as initial filters and rank the selected documents using statistically-based weighting. This may not be formally optimal, but may be perfectly satisfactory in practice. However this depends on the setup within which a conventional service is used.

With very few exceptions (e.g. Saracevic et al. 1988), evaluation within the research paradigm that TREC adopted has ignored setup characteristics beyond those encapsulated by given documents, requests and assessments. For example there has been no concern with why the documents are wanted, whether particular forms of document are sought, whether other kinds of search key are available, whether users are occasional or habitual, whether there are also other retrieval resources available as well, not to mention a host of economic factors.

This has in part been through failures to document collection formation properly, or because users are out of the reach of researchers. But this abstraction has a more fundamental rationale. The emphasis on the system itself has been justified by the 'core assumption' mentioned earlier; by the associated assumption that the more the core can do the better, because it reduces the need to pay attention to the consequences of individual setups for system design; and by what has been an article of faith in IR research, namely that the less the user has to do, the better. Thus one of the attractions of blind relevance feedback is that the user is not even asked to make any actual relevance judgements to develop search queries. Again, one of the attractions of ranked output is that where the system is not pushed towards a Precision or Recall preference, users can make their own choice of rank cutoff in

their specific situation. Other points that may be of importance for users, for instance not delivering already-seen documents or avoiding content overlap in the output set, have been assumed to be either matters of mere mechanism (though identifying duplicate documents is far from trivial and is a concern for Web system engineers), or far too difficult and dangerous to tackle. Trying to make inferences from the semantic content of a request about performance preferences (e.g. for Precision at the expense of Recall), and about what search strategies to apply, is an extremely hard problem. Solving it is a motivation for the TREC Query track, but there are serious difficulties about obtaining appropriate test collections with large request and accompanying relevance assessment sets, not to mention the user context data that is needed to support performance analysis. The Interactive track has focused more on the role of user interaction as part of the search process than on user properties *per se*.

Overall, in the research context within which TREC has placed itself, it has been assumed that requests reflect typical user needs, so if the system can deliver some relevant documents, especially at top ranks, it will ipso facto meet those needs, without any further concern with setup detail or differences between setups. The system generality that IR research seeks should be either immediately hospitable to setup differences as embodied in requests, or at any rate automatically adaptive to them, so there is no need for researchers to pay any explicit attention to contextual factors, and notably human users. Thus all of the concern for the information seeking context, and the properties of users, that the information science literature at large exhibits, and to which conventional information services pay a good deal of attention, can be legitimately ignored. The precise point about the core assumption is that unless setup properties that might determine system design can be specified in such a (concrete) way that they can be taken into account in system design - and of course this may be the case - everything that matters about the setup happens either *before* the request is submitted, or *after* the documents have been delivered. The 2003 HARD track specifically makes this assumption: thus *if* some additional information about the user can be gathered beforehand, can the system exploit it effectively in addition to type of collection data (including past query data) that it already has? But, for example, though TREC evaluations have measured Recall, this has been in an extremely abstract way, and the TREC evaluations so far have not said much about meeting high recall needs in a humanly acceptable way (without, as in Web engines, relying e.g. on page hopping as one way of improving recall).

All this does not mean that there is no rationale to investigations, using systems as black boxes, that might lead to different system design specifications. On the view just taken, that is someone else's concern, not TREC's. But it can be argued that TREC has not addressed some rather obvious specifications and thus limited itself unduly. For example, though TREC evaluations have measured Recall, this has been in an extremely abstract way, and the TREC evaluations so far have not said much about meeting high recall needs in a humanly acceptable way (without, as in Web engines, relying e.g. on page hopping as one way of improving recall).

But whether or not this is a reasonable attitude to take to long-standing, conventional systems and services, it is at least as important to ask what TREC has to do with the Web, not just as an access route to conventional resources but as a resource in its own right. Then, further, what has TREC to do with the varied types of other information resource that are appearing, enabled by the Web or networking more generally. What are the banners that TREC should be marching into this new world with? This is a pressing question, and not just because there are a good many who think that the Web in particular is the only action in town. One possibility is that though IR may be fundamentally the same on the Web as off it, it may be less easy to relegate setup issues to outer darkness. The other possibility is that

IR on the Web is *not* fundamentally the same. So the questions to address now are, first:

1. Is IR on the Web the IR task the research community has long known and continued to love in TREC?

2. If not, what is the Web IR task that TREC, if it believes in trying to solve real world problems, ought to tackle?

and then, further,

3. Are there new types of information resource facilitated by system connectivity that require new approaches to retrieval, or other tasks, that TREC should address?

5 TREC and the Web

5.1 Current Web engines

The facilities a good many Web engines offer suggest that IR on the Web is in fact the task we know. Some engines, like Yahoo!, have indeed brought historical ideas about subject classification to market on the Web. But many engines have adopted research ideas. AltaVista, for example, was built right from the start to apply statistical weighting and ranking algorithms.

This might imply that TREC has no message for the Web. Perhaps, moreover, the boot is on the other foot. Searching for simple sentence or term list requests with engines that use the familiar statistical kind of weighting often does not work very well. Precision is low, and users are invited to apply all the conventional apparatus of compulsory terms (and hence some Boolean constraints), as well as e.g. quoted phrases, to improve it.

Why don't these Web engines based on long-standing IR research work better? Setting aside problems like the one that page header 'spamming' presents, there are good reasons for this unsatisfactory performance. One is that the files are shatteringly large. At least one engine indexed over 2 *billion* pages in June 2002, and since relevant documents will always be few, they have to compete with a lot of noise. The second is that the file is amazingly heterogeneous. Lecture slides, for instance, often have few words but a high proportion of good content words; they may therefore rank high compared with other short or much longer documents, but are often depressingly insubstantial, mere bullet lists. It is hard for uniform statistical methods to respond appropriately not only to variation in document length but to variation in discourse structure and genre. However the Web engines' user needs are too varied for systems to categorise document types as useless and ignore them. The third reason is indeed that, partly because there is so much 'information' available and partly because access is free, the Web engine user community is gigantic, the range of request and need types the engines have to serve is large and the range of individual requests and needs is enormous.

Having a vast *reference library* at one's fingertips is of course what the Web is all about, and what the Web engines are intended to provide. The issue here is how the model of the IR task that much IR research - and to a considerable extent conventional search services - have been bound fits that underlying the Web. Is the *literature access* model the same as the *reference data* model?

Of course the two overlap. One may use a reference library to find about, say, elephants. But many Web engine searches are not for documents about something at all. They are what may be called 'location' searches, i.e. means to the end of finding, e.g. "Where's there a

university that runs a course on elephant training?" There are 'definition' searches: "[What does] hermeneutic [mean]?" Again, many are not intended to be selective e.g. "[Find me some] elephant pictures [i.e. any ones will do]". Many are naked direction seeking, e.g. "What's the way to the Elephant House?" None of these fit the 'classical' IR research paradigm model of a request, as in: "Give me documents about the manufacture of wastepaper bins from elephant feet".

The Web engines have nevertheless in general sought to apply the classical topic request model to all these (and other) varied types of need, or at least have retained it as a substrate, taking some combination of user-friendly boolean structure and data-reflective statistical weighting as their basic retrieval strategy. They have then added all kinds of elaborations and modifications designed to help users gain precision in topic searching or to adjust query expression to need type, for example by using compulsory terms, applying category constraints etc., but also to refine backend matching in ways that can both aid topic precision and satisfy some other types of query, for example by emphasising term matching on page titles or preferring pages that say "home page". At the same time, the engine builders have recognised that document searching is only one element in a user's complex information management activities, and so have imported other capabilities like translation or summarising on demand. They have also sought to enhance retrieval itself for the user by, for example, routinely providing minimum, query-oriented 'snippet' summaries. The Web engines have of course been able to take advantage of general interface developments e.g. having multiple windows, which facilitates browsing and can compensate for lack of retrieval accuracy.

But there have also been more radical departures. One, developed by Google, has been to wholeheartedly exploit the information supplied by page links, not just individually but collectively: i.e. to write the old idea of citation indexing new and large. When compared with 'ordinary' citation indexing, Google takes advantage on the one hand of the fact that there may naturally be a much denser supply of connectivity data in Web page links than in conventional bibliographic citation, and on other of the fact that these links are better grounded and hence more likely to be useful for retrieval than simple lexical overlap between pages. This strategy also has the substantial advantage of characterising documents which may well not make or get the usual kind of bibliographic citation.

The other, rather more substantial departure from the document retrieval model has been that on which Ask Jeeves is based, i.e. to start from the presumption that when users seek information they already have a specific question they want answered. Users are not in a rather general anomalous state of knowledge about some topic, so more specific questions only arise, and are answered, when they read the texts on the topic they have retrieved. With any luck moreover, given a very large user population, many user questions will in fact be Frequently Asked Questions, or at any rate will instantiate familiar generic question templates, so the user can be served by preprocessing the incoming data to extract potential question answers.

The engine builders' holy grail is to be able to learn enough about individual users, or rather individual needs, to be able to target material to them. But it is very hard to get direct feedback (as in relevance feedback for the individual request), and even with a good deal of log data it is hard to make reliable inferences about users (certainly in legally or ethically acceptable ways). The problem is compounded by the way users fall over the edge, out of the engine and into individual sites, taking their further information management operations, and the user data these might provide, with them. The only real weapon the engine builders have is that they can get such an enormous amount of log data that some reliable patterns

may be observed, even if any individual search does not provide much information at all.

Indeed the main challenges, both for Web engine developers and for researchers hoping to offer good ideas and tools for practical use, is that user queries are typically extremely short, averaging 2.6 terms per query in a large Excite sample, for instance. It is very hard to get any leverage at all from such a minimal starting point, when it may also not be possible to draw on the user's 'off-engine' working on actual pages, and when it is not certain that a sequence of submitted queries is actually part of a user's search to meet a single need.

Thus though Web engines may seem, by being used online with nice friendly interfaces, to involve the very close interaction between system and user that IR research ought, on some views, to cover, the relation is in reality much less close than it appears. Interaction with the Web engine is only part of the user's whole online activity, just as the user's interaction with the card catalogue in a regular library was only one part of the getting and using of information. Thus if one seeks to apply TREC findings to the Web, there appears to be less of a requirement than might have been expected to modify the core orientation, with its narrow view of the system environment, that has been characteristic of IR research and has been maintained through the TREC programme. The fact that it is difficult for anyone building a public Web engine to really *integrate* their systems with the user's own information management environment as a whole, as opposed to simply supplying one or more tools among many for the user to choose from, seems only to imply that there is scope for developing more tools for a bigger and better toolbox: i.e. for following the traditional IR research path.

Important generic ideas developed in earlier IR research, notably statistically-based weighting, have long been deployed in Web engines: as mentioned earlier, the first Web engine, AltaVista, explicitly applied these, (indirectly) encouraged by TREC results. Exploiting hyperlinks, as in Google, is another form of statistical processing, making use of a type of information, citations, long recognised as of value. Anchortext is just a piece of text, albeit perhaps an especially useful one, with words in it open to statistical weighting like other text. At the same time, while links and their anchortext may seem to supply especially well-focused and hence rich forms of index information, experiments with these in TREC-9 and TREC-2001 found that they were of no special value for topic searches, though they were helpful for the rather particular home page finding task (Hawking 2001; Hawking and Craswell 2002).

Some research-derived strategies, like statistically-based relevance feedback, have proved difficult to deploy on the Web, because they do not fit the modus operandi of Web engine users, even though they have been clearly shown to be effective offline (Sparck Jones et al. 2000) and also in filtering. Thus though they may not always work online (see Koenemann and Belkin 1996), there is no reason to suppose they *cannot* work on the Web. But this is not a central issue for TREC Web pertinence. For the primary task, adhoc search, the issue is different: given that there has already been a Very Large Collection/Web track in TREC that has had large Web document sets to work on (Hawking et al. 1999; Bailey et al. 2003; Craswell and Hawking, 2003), along with a very large log-derived query set, and given that the Web track results have been those just mentioned, what more is there for TREC to do to get closer to the Web world and its needs? The Web track results might rather imply a 'been there, done that' status for TREC in relation to the Web.

But this is a very dubious conclusion, as consideration of the current lead engine, Google, suggests. Google is a complex beast, with indexing emphasising links and, it seems, anchor text but also using term frequency information and document structure. Queries are handled on a Boolean filter with output ranking basis. In addition matching is apparently phrasally 'oriented' or proximity 'biased'. The overall thrust is towards Precision, with little or no

reference to Recall. All of this is a response, given the file data realities, to short queries that are often phrasal, especially with names, and users wanting a few good hits early on.

It can be argued that TREC has essentially failed to get to grips with the realities that an actual engine like Google has to deal with. Thus while the Web track has sought to address at least the most significant issues for Web retrieval, it has been subverted by the need to sample the Web: this has reduced linkage and, probably, the impact of varied data types. In TREC in general both topics and files have not been Web-like, as well as puny by comparison with the Web. With all its collections, and the forms of performance measure that have been applied, TREC has perhaps underestimated what a Precision focus requires, as well as what can be done with Boolean queries, or proximity constraints, or document structure, and the like.

So how could, or should TREC develop in relation to the Web?

5.2 Web directions for TREC

The obvious, perhaps most obvious, direction for TREC is to engage, much more fully than hitherto, with the heterogeneity of Web documents and requests, and with their all too common inadequacy as levers to move the information world. The miscellaneous additional (indexing and) searching devices that the engines have adopted very forcefully suggests that the statistically-based strategies familiar from TREC research are too weak in the face of the Web's characteristic messiness, its string bag mix of really useful lengths of stout cord, odd pieces of string, and little bits too short or frayed for anything much. Again, there is more to investigate in links and, especially, anchor text, as e.g. Westerveld et al. 2002 implies. At the same time, it would be instructive as well as useful for retrieval research to engage more fully with capturing and using document structure: this is a practical and theoretical challenge, for instance for weighting formulae. It might also be useful, though this has not been neglected in TREC (see e.g. Rose and Stevens 1997), to reconsider Boolean and proximity-based matching.

But really tackling these issues, and particularly exploring 'words vs. links', presents enormous challenges, most obviously in establishing adequate test collections or environments. 'Lifting' test collections from the Web presents all sorts of practical and formal difficulties; staying inside the Web presents complementary ones. But without larger-scale experiments and a fuller engagement with Web data properties at the very least but also, ideally, Web users, it may not be possible to demonstrate that what look like counterintuitive results about the value of links really hold, or that current retrieval research can contribute to improving Web search performance.

Since research done in the past, though in less taxing system environments, has provided some of the foundations for many Web engines, so new research, undertaken within more taxing environments characteristic of the Web, might in turn supply better-grounded strategies and devices to replace the adhoc assemblies of gadgets that current engines have put together on top of their system foundations. The fact that the engines continue, indeed are obliged to continue, to seek general techniques that will work across a range of cases representing many actually different types of request, implies that there is further research in the traditional style, on better retrieval methods, to do. In particular, while it is not clear how to replicate Web query sequencing within a single search activity, the number of past queries that Web engines accumulate offers the opportunity to explore the value of query clustering. This is an old idea (see e.g. Worona 1972) which can now be much more thoroughly investigated;

and Scholer and Williams' (2002) study suggests that with sufficient queries, the weakness of the inferences that can be drawn about relevance from single query-document matches can be overcome. On similar lines, there appears to be much more to investigate in relating the TREC filtering, as well as adhoc, technology to the Web data and its clientele. While so-called intelligent agents may already figure on the Web, there is much more scope for principled approaches to filtering related to Web engines.

Simply comparing the output, for the same request, obtained by searching AltaVista, Google, and other engines (also including Ask Jeeves), is extremely instructive: the results may overlap in some cases, be quite distinct in others. But the output from any one engine is not consistently superior to that of any other: the most striking point in many cases is that when different engines deliver good output it is also complementary output; indeed there are distinctive glittering nuggets in each system's dross. So the crucial issue for TREC is how to scale its work up and out, to get the test data and task specifications that are required to satisfy both operational pertinence and scientific propriety needs

6 Going beyond the Web

6.1 Other webs

The 'ordinary' public Web, significant though it is, is not the only arena for retrieval research. There is the world of corporate intranets, of Usenet (the original stimulus for AltaVista). Research-based systems may be more helpful in such dedicated enviromments than in the general Web as, for example, Autonomy claims (Autonomy 2004). But this of course raises the problem of getting public test collections. The same problem arises with other manifestations of the hidden Web' for instance the very large conventional document databases being made available by journal publishers.

6.2 New data types

Both the public Web, as a directly accessible information resource, and these other information worlds, raise the issue of data types other than text in more or less conventional and familiar forms.

The Spoken Document Retrieval track in TREC has already explored, albeit on a small scale, retrieval from speech data; and it has shown that good performance can be obtained, though transcription is far from perfect, using standard text retrieval techniques. The recently introduced Video track, addressing the image retrieval task that is increasingly important both on and off the Web, draws attention not only to the need to rethink the notion of index key but also that of what a query looks like. What sort of thing can a video query be: "Find me a sequence showing a horse race", "Find me a striking closeup of a horse?", "Show me some cool panning"? There is much more challenging work on image retrieval to do, not just on its own but in applying and using language keys related to images, and for many image types including, for example, those in scientific image databases.

In the same way, the range of resource types available on the Web emphasises the need to develop hybrid or multi-function search techniques, able to search different types of file - including conventional structured databases, semi-structured data, and text, from a single starting point in, say, a natural language request, and to integrate the results. Effective combinations of search methods, from simple statistical ones to those properly requiring

natural language processing, are a necessary precursor for Web data mining (see *AI* 2002) and are also a good line forward for TREC work. The main immediate challenge is devising suitable test environments and obtaining appropriate test data for meaningful evaluations; but the new Genomics track is a natural starting point for this.

The Genomics track draws attention to the opportunities more generally, with networks, for multi-facetted information bases. This is not a new notion, or even actuality, but the ease with which connections between one base and another can be made, both by the system and the user, emphasises the challenge for future TRECs of dealing with information requests that are single entry points to a range of resource types, for example textual and numerical, or in unstructured and structured databases, and require a retrieval mechanism able to develop a set of query types from this one starting point. This is an area which goes far beyond the visible Web.

6.3 New tasks

Pursuing a more orthogonal line, the question is what new tasks rather than data types TREC ought to address. TREC has already taken a major step forward here in introducing the Question Answering track. While many Web engine queries are topic ones or, like home page searches, approximate to 'known item recovery', it is quite clear that many users would like specific questions answered. The TREC Question Answering tests so far have shown that when longer answer passages, i.e. passages hopefully embedding answers, are allowed, established text retrieval methods can be quite competitive with only modest elaboration. However when only brief answer snippets are permitted, and even more when exact answers are required as in TREC-2002 (Voorhees, 2003), a significant ramp-up in analysis and search techniques exploiting natural language processing, to at least a non-trivial extent, is needed.

The Question Answering track, challenging though it is, has been a success, with an animating effect on TREC as a whole. It has served to emphasise the fact that while document retrieval is a valid task in its own right, it is also part of a spread of information seeking, modification, and presentation tasks. From this point of view TREC has been a standard-bearer outside the retrieval world. The level of performance that has been obtained with a difficult task has been impressive (e.g. Moldovan et al, 2003). But it is also significant that while explicit natural language processing appears to be essential, it can be materially enhanced with statistical learning, as illustrated by Yang and Chua (2003). There is manifestly much more, very hard research to do to support online, user-specific, question answering. This means not only being able to handle a range of question types, but also being able to accommodate the many intrinsic uncertainties about what the user's question is, and to degrade gracefully when direct answering cannot be done. This is clearly a line of work that TREC can grow further with, though it clearly also raises the issue of interactive question answering and a system's ability to extract pertinent contextual information.

Even here, however, traditional retrieval has a contributory role in selecting long passages, potentially containing answers, for more detailed analysis. Text retrieval has similar natural roles in relation to other tasks, e.g. in supplying likely text for detailed information extraction, in selecting documents pertinent to specific topic tracking, as in following news stories, and in delivering key documents to be summarised. It may not be appropriate for TREC to extend itself to one or more of these other tasks simply as a matter of course, - some already have their own evaluation programmes, for example DUC (the so-called Document Understanding Conference) addressing summarisation (DUC 2003). TREC needs to

move into task areas either where there is a natural issue as to whether current IR methods are applicable or can be extended to be applicable, or whether the connection between the retrieval sub-task and the other task(s) is particularly close. Question Answering illustrates the former, and topic tracking and summarising, especially extractive summarising, could illustrate the latter. TREC thus needs to monitor other evaluation initiatives, taking advantage of the Road Mapping exercises that these may involve, and the changing opportunities for research partitioning or collaboration that they offer, as already illustrated with the two question answering programmes running respectively under TREC and ARDA's AQUAINT initiative (AQUAINT 2003).

6.4 New resources

The relation between TREC and the Web is not a one-way street. The Web is also a source of resources for those engaged with retrieval and similar tasks. This is not only in the obvious sense, as in supplying e.g. parallel texts as a source of translation equivalents for Cross Language retrieval, or handy dictionaries. The Web is also, as a huge text base, a source of information about word usage and discourse forms. It was thus used, for instance, to supply additional forms of question answer-patterns for searching the target file for the Question Answering task (Brill et al. 2002). Opportunities like this imply that, just as modern information retrieval began by recognising the value of direct text clues, its indexing and searching tools can be further refined by exploiting the vast quantities of text the Web makes available. The same of course applies to off-Web resources, as Autonomy's corporate applications imply.

7 The future: multi-tasking

TREC can naturally, and valuably, continue along its existing lines, primarily

1. pushing adhoc retrieval and its variations for new types of need or material, and
2. tackling other individual information-seeking tasks.

There is plenty here to occupy the research community, in collection building and evaluation design as well as system development and testing. However there are problems about just continuing with more of the same.

As noted earlier, TREC has implications for retrieval in general, not only for retrieval on the Web. At the same time, Web engines illustrate major constraints on the core retrieval system that has been the main focus of TREC so far. First, Web engines deal very successfully with the 'quasi' known-item searches that figure so largely in Web usage but to which TREC's topic-based search model can contribute very little. Second, Web users can remain, as they seem to be, lacking in search enterprise because there are so many pages out there on virtually any topic that even the most minimal search specification can usually retrieve something useful, but where the richer search strategies that IR research offers cannot get much leverage. Third, Web engines have been driven, by the quantity and heterogeneity of the material they are dealing with, to adopt 'everything including the kitchen sink' approaches to indexing and searching that fall far outside IR research practice so far, as illustrated by TREC, even if more all-embracing but still principled approaches could be developed through research study.

Fourth, as noted earlier, most users' information-seeking behaviour on the Web takes place beyond the scope of its engines, even further from the core system focus that TREC has had.

There are therefore good reasons for taking a more radical approach to TREC in the future. Continuing with the 'one-task-at a time' approach, even if over a wider range of separate tasks than hitherto, would not be pursuing the capability that information management under Web conditions especially, but also under modern IT conditions in general, should really offer: namely of providing a properly *integrated* information management service subsuming different tasks that can be executed, as occasion demands, in any particular user situation. This is, of course, the "integrated solutions" mantra that business system vendors invoke, though in practice to rather limited effect: much more real power is needed. The Web services have already begun to move in this direction. They already take in one another's washing as document retrieval engines, and in some cases point to one another within the broader framework of response to enquiry, by referring to question answering as well as document retrieval. They have also, more importantly, begun to offer other task capabilities, like translation and summarisation on request.

But a vision of what a Web (or other modern) information management should be like has a larger task range - for example including information extraction, new text derivation as well as translation and summarisation - and looks for more integration than the superficial one represented by a collection of buttons in a menu. Being able to invoke different task facilities at the press of a button in a single menu is much better than nothing, as a convenience that information technology has brought us. But proper integration, allowing the user to move effectively from one task to another at will, implies a truly common information environment where pertinent, user-specialised detail can be moved between tasks and exploited as required. The current multiple task options that the engines offer are only superficially related, and hence not as productive as they should be. They do not maintain and use current context properly, and so do not take proper advantage of the information about the user that is in principle available to make the execution of any particular task more helpfully personalised.

There are beginning to be operational systems that offer integrated multi-tasking, for example MiTAP (Damianos et al. 2002). MiTAP draws on the knowhow developed in TREC and its companion evaluation programmes. But there is much more to explore here, especially in how far general, statistically-based methods can provide a common platform across tasks.

8 Rethinking TREC from the bottom up

TREC has been hugely successful in three different ways:

1. it has been a major IR research programme that has delivered many important results;
2. it has built a large community (around 80 teams took part in TREC-2001, for example) and has fostered links with other, hitherto separate, communities, importing participants and ideas e.g. from speech and natural language processing;
3. it has stimulated, and will continue to stimulate, retrieval research outside TREC, by reporting findings for comparison and supplying test data, and by encouraging other programmes, as in the cross language retrieval CLEF and NTCIR evaluations (see CLEF 2003; NTCIR 2003).

4. it has encouraged the application of statistical methods of information processing in task areas outside retrieval, e.g. by exporting *tf*idf*-type word weighting.

It would indeed also have been nice to point to clear evidence that TREC findings have been taken up by commercial systems, especially Web ones. But though individual researchers link TREC with the operational world, so one hopes there has been some carryover, those responsible for Web engines and the like do not publish details of how the engines work.

There are areas within IR that TREC has not significantly addressed, for example retrieval from large files of full-text scientific material, primarily because it is difficult to get suitable test collections. Since the proprietors of large journal data files continue to maintain conventional approaches, not being able to challenge this conventional wisdom in sound evaluations is unfortunate.

But this is not the critical future direction for TREC. After a decade's solid work on document retrieval, it is time for TREC to enter a radically new phase. This can be expressed by saying it is necessary to relate TREC more fully to the Web than hitherto. But it means more, however, than focusing on the Web because it is there, or taking advantage of thinking about IR and the Web to review what we suppose information *retrieval* is all about. Relating TREC more fully to the Web, and beyond the Web to information environments generally, implies we have to think again about what TREC's foundation, on some principled view of what information *management* is, should be, and that

The natural development is still to start from the notion of text, as with the first decade of TREC, albeit viewing this notion very broadly; but it is also, now, to start from *interpretation*, not just retrieval. This means moving upwards and outwards from texts to cover a range of tasks, some crude some complex, that are all related in doing something, in some way, with some information from some text(s), and are also in operation *dependent on one another* because they are invoked in common contexts. Studying the way common methods of processing text can be applied to different question answering requirements, notably for long or short extracts, in the TREC Question Answering track has been a modest move in the new direction. Other work, for example on summarising (Mani and Maybury 1999; DUC 2003) has long explored extractive techniques with much in common with those used for retrieval. The fact that tasks currently only rather contingently related to the TREC Programme, like information extraction, topic detection and tracking, and summarisation, share technologies with retrieval is thus one good reason to think about developing TREC to make connections with them. But the much more important reason for TREC to make these connections is that we want future information management systems to be able to carry out their tasks as subtasks supporting the users' information management activities. This implies a common, multipurpose evaluation framework so that, for example, if we take summaries as surrogates for full texts as inputs to question answering, we can relate question answering effectiveness to summarising effectiveness. The evaluation experience that TREC has gained in the last decade makes TREC well placed to tackle more evaluation scenarios for more ambitious information management situations, and also justifies the argument that it should advance in this direction, tough though this will be.

Acknowledgement

I am grateful to my referees for comments.

References

Adi, T., Ewell, O.K. and Adi, P. 'High selectivity and accuracy with READWARE's automated system of knowledge organisation', *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-246, Gaithersburg, MD: National Institute for Standards and Technology, 2000, 493-498.

AI. Special Issue on Intelligent Internet Systems. *Artificial Intelligence*, 118 (1-2), 2000, 1-275.

Allan, J. (Ed.). *Topic detection and tracking: event-based information organisation*, Boston: Kluwer, 2002.

AQUAINT. See http://www.ic-arda.org/InfoExploit/aquaint/. Visited August 2003.

Autonomy. See http://www.autonomy.com/. Visited April 2004.

Bailey, P., Craswell, N. and Hawking, D. 'Engineering a multi-purpose test collection for Web retrieval experiments', *Information Processing and Management*, 39, 2003, 853-871.

Barker, F.H., Wyatt, B.K. and Veal, D.C. 'Report on the evaluation of an experimental computer-based current-awareness service for chemists', *Journal of the American Society for Information Scientists*, 23, 1972, 85-99.

Beaulieu, M. 'Experiments on interfaces to support query expansion', *Journal of Documentation*, 53, 1997, 8-19.

Blair, D.C. and Maron, M.E. 'An evaluation of retrieval effectiveness for a full-text document retrieval system', *Communications of the ACM*, 28, 1985, 289-299.

Brill, E. et al. 'Data-intensive question answering', *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-250, Gaithersburg, MD: National Institute for Standards and Technology, 393-400.

Cardie, C. 'Empirical methods in information extraction', *The AI Magazine*, 18 (4), 1997, 65-79.

Carmel, E., Crawford, S. and Chen, H. 'Browsing in hypertext: a cognitive study', *IEEE Transactions on Systems, Man, and Cybernetics*, 22, 1992, 865-884.

CLEF. See http://clef.iei.pi.cnr.it/. Visited August 2003.

Cleverdon, C.W. 'The Cranfield tests on index language devices', *Aslib Proceedings*, 19, 1967, 173-194.

Craswell, N. and Hawking, D. 'Overview of the TREC-2002 Web track', *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-251, Gaithersburg, MD: National Institute for Standards and Technology, 2003, 86-95.

Croft, W.B. (Ed.) *Advances in information retrieval*, Dordrecht: Kluwer, 2000.

Croft, W.B. and Lafferty, J. (Eds.) *Language modelling for information retrieval*, Dordrecht: Kluwer, 2003.

Damianos, L. et al. 'MITAP for biosecurity: a case study' *The AI Magazine*, 23 (4), Winter 2002, 13-29.

Derr, R.L. 'Information seeking expressions of users', *Journal of the American Society for Information Science*, 35, 1984, 124-128.

DUC. See http://www-nlpir.nist.gov/projects/duc/. Visited August 2003.

Dumais, S.T. 'Latent Semantic Indexing (LSI): TREC-3 report', *Proceedings of the Third Text REtrieval Conference (TREC-3)*, Ed. D.K. Harman. Special Publication 500-225, Gaithersburg, MD: National Institute for Standards and Technology, 1995, 219-230.

Fagan, J.L. *Experiments in automatic phrase indexing for document retrieval: a comparison on syntactic and non-syntactic methods*, PhD Thesis, TR 87-868, Department of Computer Science, Cornell University, 1987.

Gordon, M. and Pathak, P. 'Finding information on the World Wide Web: the retrieval effectiveness of search engines', *Information Processing and Management*, 35, 1999, 141-180.

Hawking, D. 'Overview of the TREC-9 Web track', *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-249, Gaithersburg, MD: National Institute for Standards and Technology, 2001, 87-102.

Hawking, D. and Craswell, N. 'Overview of the TREC-2001 Web track', *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-250, Gaithersburg, MD: National Institute for Standards and Technology, 2002, 61-67.

Hawking, D., Craswell, N. and Thistlethwaite, P. 'Overview of TREC-7 very large collection track', *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-242, Gaithersburg, MD: National Institute for Standards and Technology, 1999, 91-103.

IP&M. Special Issue on Interactivity at the Text Retrieval Conferences (TREC). *Information Processing and Management*, 37, 2001, 365-541.

Koenemann, J. and Belkin, N.J. 'A case for interaction: a study of interactive information retrieval behaviour and effectiveness', *Proceedings of CHI 1996*, 1996, 205-212.

Kwok, K.L. 'A network approach to probabilistic information retrieval', *ACM Transactions on Office Information Systems*, 13, 1995, 325-353.

Kwok, K.L. and Chan, M. 'Improving two-stage ad-hoc retrieval for short queries', *Proceedings of the 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1998, 250-256.

Lu, X.A., Holt, J. D. and Miller, D.J. 'Boolean system revisited: its performance and its behaviour' *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, Ed. D.K. Harman, Special Publication 500-236, Gaithersburg, MD: National Institute for Standards and Technology, 1996, 459-473.

Mahesh, K., Kud, J. and Dixon, P. 'Oracle at Trec8: a lexical approach', *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-246, Gaithersburg, MD: National Institute for Standards and Technology, 2000, 207-216.

Mani, I. and Maybury, M.T. (Eds.) *Advances in automatic text summarisation*, Cambridge, MA: MIT Press, 1999.

Miller, D.R.H., Leek, T. and Schwartz, R.M. 'A hidden Markov model information retrieval system', *Proceedings of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1999, 214-221.

Mitra, M. et al., 'An analysis of statistical and syntactic phrases', *Proceedings, RIAO-97, Computer-Assisted Information Searching on Internet*, (Montreal), Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaires, 1997, 200-214.

Moldovan, D. et al. 'LCC tools for question answering', *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-251, Gaithersburg, MD: National Institute for Standards and Technology, 2003, 388-397.

NTCIR. See http://research.nii.ac.jp/ntcir/. Visited August 2003.

O'Connor, J. 'Text searching retrival of answer-sentences and other answer passages', *Journal of the American Society for Information Scientists*, 24, 1973, 445-460.

Okapi. Papers on Okapi, Special issue, *Journal of Documentation*, 33, 1997, 3-87.

Over, P. 'The TREC interactive track: an annotated bibliography', *Information Processing and Management*, 37, 2001, 369-381.

Perez-Carballo, J. and Strzalkowski, T. 'Natural language information retrieval: progress report', *Information Processing and Management*, 36, 2000, 155-178.

Ponte, J.M. and Croft, W.B. 'A language modelling approach to information retrieval', *Proceedings of the 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1998, 275-281.

Rose, D.E. and Stevens, C. 'V-Twin: a lightweight engine for interactive use', *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-238, Gaithersburg, MD: National Institute for Standards and Technology, 1997, 279-290.

Salton, G. *Automatic information organisation and retrieval*, New York: McGraw-Hill, 1968.

Salton, G. (Ed.) *The SMART retrieval system*, Englewood Cliffs NJ: Prentice-Hall, 1971.

Salton, G. 'A new comparison between conventional indexing MEDLARS and automatic text processing (SMART)', *Journal of the American Society for Information Science*, 23, 1972, 75-84.

Salton, G. and Buckley, C. 'Term weighting approaches to automatic text retrieval', *Information Processing and Management*, 24, 1988, 513-523.

Salton, G. and McGill, M.J. *Introduction to modern information retrieval*, New York: McGraw-Hill, 1983.

Saracevic, T. et al. 'A study of information seeking and retrieving. I. Background and methodology; II. Users, questions and effectiveness. III. Searchers, searches, and overlap', *Journal of the American Society for Information Science*, 39, 1988, 161-176; 177-196; 197-216.

Scholer, F. and Williams, H.E. 'Query association for effective retrieval', *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2002, 324-331.

Schultz, C.K. (ed),. *H.P. Luhn : Pioneer of information science*, New York: Spartan, 1968.

Sparck Jones, K. (Ed.) *Information retrieval experiment*, London: Butterworths, 1981. OCR version available in IRLIB at http://www.nist.gov/itl/div894/894.02/projects/irlib (visited in 2000).

Sparck Jones, K. 'Reflections on TREC', *Information Processing and Management*, 31, 1995, 291-314.

Sparck Jones, K. 'Summary performance comparisons, TREC-2 through TREC-7', in *TREC-7*, 1999, B1-B6; and 'Summary performance comparisons, TREC-2 through TREC-8', in *TREC-8*, 2000, B1-B5.

Sparck Jones, K. 'Further reflections on TREC', *Information Processing and Management*, 36, 2000, 37-85.

Sparck Jones, K. 'Automatic language and evaluation processing: rethinking evaluation', *Natural Language Engineering*, 7, 2001, 1-18.

Sparck Jones, K. and Galliers, J.R. *Evaluating natural language processing systems*, Lecture Notes in Artificial Intelligence 1083, Berlin: Springer, 1996.

Sparck Jones, K., Walker, S. and Robertson, S.E. 'A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2', *Information Processing and Management*, 36 (6), 2000, 779-808 and 809-840.

Stevens, M.E. *Automatic indexing: a state of the art report*, Monograph 91, National Bureau of Standards, Washington D.C., 1965.

Sullivan, M.V., Borgman, C.L. and Wippern, D. 'End-users, mediated searches, and front-end assistance programs on Dialog: a comparison of learning, performance, and satisfaction', *Journal of the American Society for Information Science*, 41, 1990, 27-42.

Tenopir, C. and Cahn, P. 'TARGET and FREESTYLE: DIALOG and MEAD join the relevance ranks', *Online*, 18 (3), 1994, 31-47.

TREC. The First Text REtrieval Conference (TREC-1), Ed. D.K. Harman, Special Publication 500-207, Gaithersburg, MD: National Institute of Standards and Technology, 1993; *The Second .. (TREC-2)*, Ed. Harman, 500-215, 1994; *The Third ... (TREC-3)*, Ed. Harman, 500-225, 1995; *The Fourth ... (TREC-4)*, Ed. Harman, 500-236, 1996; *The Fifth ... (TREC-5)*, Ed. E.M. Voorhees and D.K. Harman, 500-238, 1997; *The Sixth ... (TREC-6)*, Ed. Voorhees and Harman, 500-240, 1998; *The Seventh ... (TREC-7)*, Ed. Voorhees and Harman, 500-242, 1999; *The Eighth ... (TREC-8)*, Ed. Voorhees and Harman, 500-246, 2000; *The Ninth ... (TREC-9)*, Ed. Voorhees and Harman, 500-249, 2001; *The Tenth ... (TREC-2001)*, Ed. Voorhees and Harman, 500-250, 2002; *The Eleventh ... (TREC-2002)*, Ed. Voorhees and Harman, 500-251, 2003.

Voorhees, E.M. 'Variations in relevance judgements and the measurement of retrieval effectiveness', *Information Processing and Management*, 36, 2000, 697-716.

Voorhees, E.M. 'Overview of the TREC 2002 question answering track', *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-251, Gaithersburg, MD: National Institute for Standards and Technology, 2003, 57-68.

Westerveld, T, Hiemstra, D. and Kraaij, W., 'Retrieving Web pages using content, links, URL's and anchors', *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-250, Gaithersburg, MD: National Institute for Standards and Technology, 663-672.

Worona, S. 'Query clustering in a large document space', in Salton, G. (Ed.). *The SMART retrieval system*, Englewood Cliffs NJ: Prentice-Hall, 1971, 298-310.

Yang, H. and Chua, T.-S. 'The integration of lexical knowledge and external resources', *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, Ed. E.M. Voorhees and D.K. Harman. Special Publication 500-251, Gaithersburg, MD: National Institute for Standards and Technology, 2003, 486-491.

Zobel, J. 'How reliable are the results of large scale information retrieval experiments?', *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, 307-314.