Pergamon

Information Processing & Management, Vol. 32, No. 3, pp. 273–286, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0306-4573/96 $15.00+0.00

0306-4573(95)00061-5

RELEVANCE JUDGMENTS FOR ASSESSING RECALL

PETER WALLIS and JAMES A. THOM
Department of Computer Science, RMIT, G.P.O. Box 2476V, Melbourne, Victoria 3001, Australia

(Received August 1994; accepted September 1995)

Abstract—*Recall* and *Precision* have become the principle measures of the effectiveness of information retrieval systems. Inherent in these measures of performance is the idea of a *relevant* document. Although recall and precision are easily and unambiguously defined, selecting the documents relevant to a query has long been recognized as problematic. To compare performance of different systems, standard collections of documents, queries, and relevance judgments have been used. Unfortunately the standard collections, such as SMART and TREC, have locked in a particular approach to relevance that is suitable for assessing precision but not recall. The problem is demonstrated by comparing two information retrieval methods over several queries, and showing how a new method of forming relevance judgments that is suitable for assessing recall gives different results. Recall is an interesting and practical issue, but current test procedures are inadequate for measuring it. Copyright © 1996 Elsevier Science Ltd

INTRODUCTION

Four decades of testing information retrieval systems suggests that automated keyword search is a very effective means of finding relevant material. Nevertheless, for some queries there are documents which keyword search will not find. Recent attempts to improve the effectiveness of information retrieval systems include adding natural language processing techniques (Fagan, 1987; Smeaton, 1987; Gallant *et al.*, 1992; Wallis, 1993), broadening the query using automatic and general purpose thesauri (Crouch, 1990; Thom & Wallis, 1992), document clustering techniques (El-hamdouch & Willett, 1989), and other statistical methods (Deerwester *et al.*, 1990). The measured performance of these techniques has not been encouraging. In this paper we argue that the reason for this result may lie in biases associated with the assessment method rather than with the information retrieval techniques being proposed.

The emphasis in information retrieval research has been, it seems, on users who want to find just a few relevant documents quickly. Once these users have found the required information, they stop searching. At the extreme, what is wanted is a system that finds a single relevant document and no non-relevant documents. In other words, these users need a system that emphasizes precision. In many situations, however, a user must be fairly confident that a literature search has found all relevant material; such a user requires *high recall.* Su (1994) has found that users are more concerned with absolute recall than with precision. Her results were based on information requests by users in an academic environment, but her findings are applicable to others. For instance when lodging a new patent at the patent office it is necessary to retrieve all relevant, or partially relevant, material. Finding precedence cases in legal work and intelligence gathering are other situations in which high recall is desirable.

Some might argue that there is no need for high recall retrieval systems because existing tools are good enough. Cleverdon (1974) has suggested that, because there is a significant redundancy in the content of documents, all the relevant information on a topic will be found in only one-quarter of the relevant material. But the converse does not hold: one-quarter of the relevant material does not necessarily contain all the relevant information. So a user interested in high recall will need to find much more than one-quarter of the relevant documents. Users who need high recall information retrieval get by with existing tools, but there is a need for systems that emphasize the recall side of the problem.

Apart from practical benefits, improving the recall of information retrieval systems is an interesting research problem. Swanson (1988) has said there are conceptual problems in information retrieval that have been largely ignored. One of the "postulates of impotence" he proposes is that human judgment can bring something to information retrieval that computers cannot because computers cannot understand text. Our work on semantic signatures goes part way toward computer understanding and shows where understanding can contribute to the information retrieval process. In this paper we show that the high recall part of the information retrieval problem may have been ignored, not because people have found the problem esoteric or uninteresting, but because they have not had the tools to effectively test ideas.

This paper is structured as follows. In the section on *relevance judgments and information retrieval*, we examine what it means for a document to be relevant, how relevance has been assessed in some existing test collections, and how the processes have incorporated a bias toward systems that emphasize precision. In the section on *information retrieval and assessment of recall*, we propose a modification to the assessment process that shifts the emphasis to recall. We also discuss the TREC collection (Harman, 1992) and why this collection does not solve the problem of assessing recall. In the section *comparing two systems for recall*, we describe an information retrieval mechanism based on semantic signatures and compare it with a keyword retrieval mechanism. We re-evaluate the relevance judgments for part of the CACM test collection (Buckley *et al.*, 1988) and use the new judgments for comparing the two mechanisms.

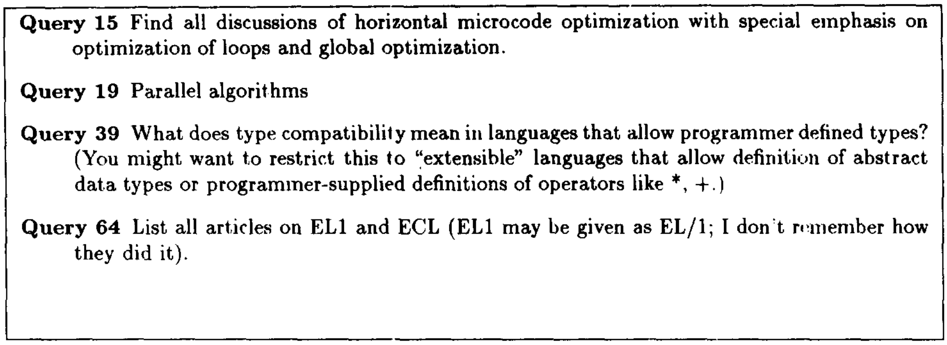
RELEVANCE JUDGMENTS AND INFORMATION RETRIEVAL

Information retrieval systems are usually considered to contain a static set of documents from which a user is wanting to extract those documents he or she will find interesting. Some documents are relevant to the user, and others are not. The effectiveness of a system for a particular query can be measured by computing recall and precision figures based on a list of documents that are considered to be relevant. For a given query and information retrieval system, *recall* measures the number relevant documents retrieved as a proportion all relevant documents, and *precision* measures the number of relevant documents retrieved as a proportion all documents retrieved. The perfect system would return a set of documents to the user containing all the relevant documents, and only the relevant documents. Such a query would have recall of 100% and precision of 100%. Critical to computing recall and precision is determining the set of relevant documents for a query. This is not as easy as it may seem.

Intuitively a relevant document is one that satisfies some requirement in the user's mind. Saracevic (1975) describes this approach to relevance as "a primitive 'y' know" concept, as is information for which we hardly need a definition". This concept of relevance was the basis of early relevance judgments in which the person who formulated the query chose the relevant documents.

> Only one person (the requester) was asked to collect the judgments for each request, and dichotomous assessments were made to declare each document either relevant or not (Lesk & Salton, 1969)

The position that the requester knows what he or she wants, and is therefore the person who knows what is relevant, is entirely reasonable, but it is not necessarily a good means of assessing the effectiveness of a retrieval system. What a user wants may not be what he or she describes. There are several approaches taken to this problem. The simplest is to ignore it; recall and precision can only be used comparatively anyway and Lesk and Salton (1969) have argued that user judgments can be effectively used to compare the performance of different systems. At the other extreme, many have tried to find a formal definition of relevance that would allow us to say definitively whether a text was relevant or not. A good example of this approach is Cooper's (1971) definition of *logical relevance*. More recently the TIPSTER and TREC projects (Harman, 1992) employ specially trained relevance assessors who, it is assumed, can make consistent and



Query 15 Find all discussions of horizontal microcode optimization with special emphasis on optimization of loops and global optimization.

Query 19 Parallel algorithms

Query 39 What does type compatibility mean in languages that allow programmer defined types? (You might want to restrict this to "extensible" languages that allow definition of abstract data types or programmer-supplied definitions of operators like *, +.)

Query 64 List all articles on EL1 and ECL (EL1 may be given as EL/1; I don't remember how they did it).

Fig. 1. Four queries from the CACM test collection.

accurate assessments of relevance. It is unrealistic to expect a third party, be it a machine or an information officer, to find what the author of a query wants rather than what the author actually requests. This especially so for poorly expressed queries. In many cases users do not express their desires clearly and the same query can be given by two users with significantly different meanings. Consider the second of the sample queries shown in Fig. 1. These queries are selected from the *Communications of the ACM* (CACM) test set of 64 queries and 3204 documents provided with the SMART information retrieval system (Buckley *et al.*, 1988). Query 19 is quite ambiguous. Does the author of this request want examples of parallel algorithms, or information about parallel algorithms? And does the author want everything on parallel algorithms, or simply a few examples? Figure 2 illustrates the difference between user judgments and those of a third party. A user in making relevance judgments will assess the relevance of documents based on an information need in the user's mind, whereas a third party making relevance judgments will assess the relevance of documents based on the text of the query as expressed by the user.

The problem of judging relevance has been around for a long time and the idea of third party judges is not new. There can be significant disagreement, not only between a third party and the author of the query, but also amongst third party judges working on the same query. In 1953 a large scale experiment was designed to compare the retrieval effectiveness of two information retrieval systems developed by different institutions. Two sets of relevance judgments were provided by the separate groups for a test collection consisting of 15,000 technical documents and 98 questions. There were 1390 documents that the two groups agreed were relevant, and another 1577 that one but not both thought relevant—"a colossal disagreement that was never resolved" (Swanson, 1988).

If relevance judgments are so unstable, how can they be used as the basis for objective measurement of information retrieval system performance? Cuadra and Katter think they cannot, and say (as quoted by Lesk & Salton, 1969)

> the first and most obvious implication is that one cannot legitimately view 'precision' and 'recall' scores as precise and stable bases for comparison between systems or system components, unless [appropriate controls are introduced].

Lesk and Salton (1969) have argued that, on the contrary, for the purposes of assessing information retrieval systems it does not matter whether the relevance assessments of either the author or a third party are used. They found that although their volunteers gave inconsistent relevance judgments, these judgments could still be used to *compare* the relative effectiveness of information retrieval systems. They describe experiments in which the relevance judgments of the author of a query, and a second person were compared over 48 queries. On average there was about a 30% overlap between the two sets of judgments (ranging from an average of about 10% for one author to 53% for another—the actual queries are not provided). Similar differences between user judgments and second-person judgments have been reported elsewhere in the literature (Janes, 1994). Lesk and Salton (1969) show that as long as relevance judgments are used to compare systems, it does not seem to matter which person's relevance judgments are used, and a technique for information retrieval that performs well on one set of judgments will perform well on others as well. The explanation they provide for this phenomenon is that there is substantial overlap on the set of documents that are "most certainly relevant to each query". They conclude

> that although there may be considerable difference in the document sets termed relevant by different judges, *there is in fact a considerable amount of agreement for those documents which appear most similar to the queries and which are retrieved early in the search process* (assuming retrieval is in decreasing correlation order with the queries). Since it is precisely these documents which largely determine retrieval performance it is not surprising they find that the valuation output is substantially invariant for the different sets of relevance judgments (Lesk & Salton, 1969, p. 355).

In other words, since everyone agrees which documents should be found first, and the first found documents are most important, the relative performance of information retrieval systems can be tested on anyone's judgments. Why the first found documents are most important is discussed below, but here it should be noted that attempting to create a system which gives 100% recall is futile. The system would only need to be finding 30% of the material a given user thought relevant to perform as well as a human relevance assessor, and to perform better than this would seem to require the system to read the user's mind.

This problem with measuring recall is addressed in the next section, but first it is necessary to show why the first found relevant documents have more influence on average performance than the relevant documents found later. An information retrieval mechanism employing ranking provides the user with a list of documents that the system considers relevant. The list is

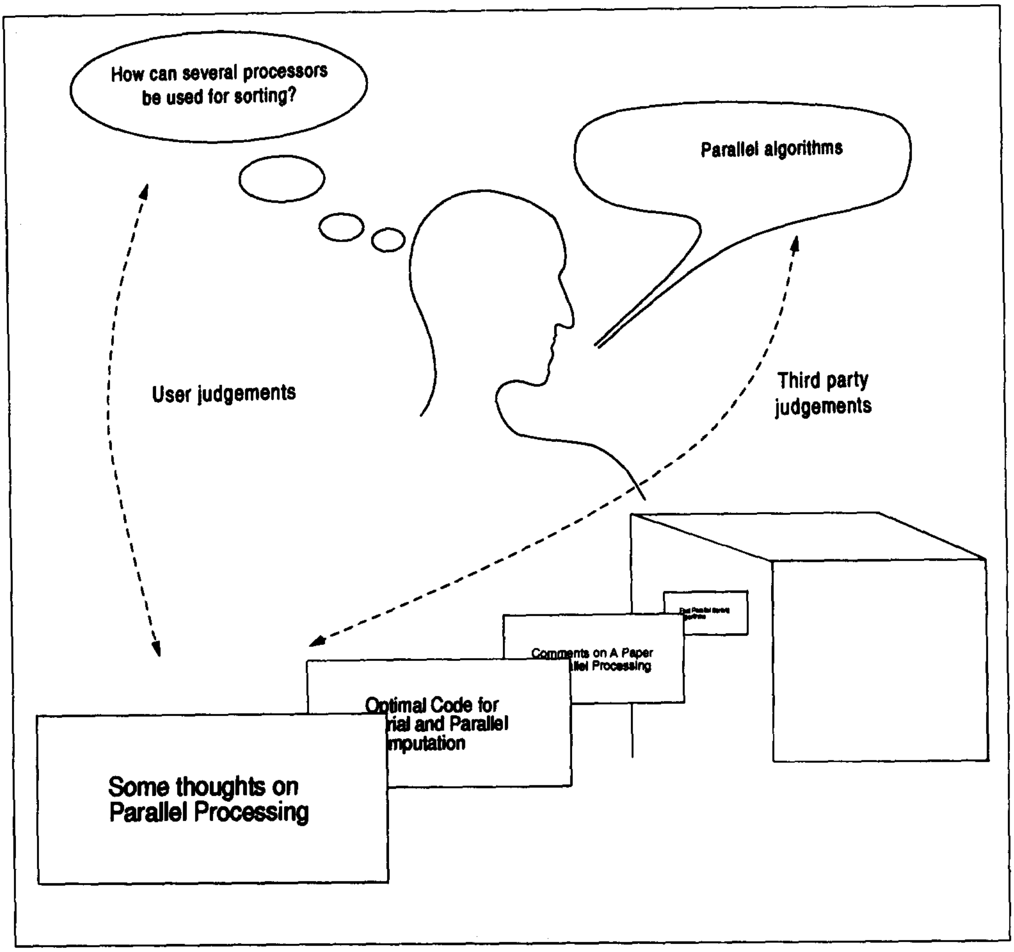

How can several processors be used for sorting?
Parallel algorithms
User judgements
Third party judgements
Some thoughts on Parallel Processing
Optimal Code for rial and Parallel mputation
Comments on A Paper llel Processing

Fig. 2. Users, queries, and relevance.

presented to the user in decreasing order of relevance—according to the system. Ranking systems can deal with what are often called "noise words". This feature allows users to enter their query in natural English. Consider phrases in four queries shown in Fig. 1 such as "I don't remember" in the CACM query 64, "Find all" in query 15, and "what does" in query 39. As far as keyword retrieval is concerned, these add no useful information to the query (and would not be used as part of a boolean query), but they do not seem to hinder the performance of ranking systems that incorporate stop-lists and term weighting. Such phrases can thus be left in and enable users to express their desires in a form that comes naturally.

Given an information retrieval system that ranks documents, performance can be assessed using recall and precision figures by comparing the list of documents returned with the set of documents judged as relevant in the test collection. For instance, a keyword retrieval system may return documents in the following order for query 64.

2651 1307 2513 793 . . .

According to the relevance judgments provided, the document 2651 is relevant, and indeed it is the only relevant document. The system has done well and the first document the user looks at will be useful.

When a ranking information retrieval system does not perform perfectly however, assessment is more difficult. Consider a ranking system executing Query 15 from the CACM collection. It has 10 relevant documents, and documents are returned by the system in the following order.

820 2835 3080 2685 307 2929 2616 2344 3054 1466 113 1461 658 1231 . . .

Of these the underlined document identifiers are the only relevant ones in the top fourteen documents. The precision at the 10% level of recall is calculated as if the system had stopped after the fourth document, giving 25% precision. The next relevant document occurs at position 14, and the next at position 20. We plot the precision at 10% increments of recall as the broken line in Fig. 3. The performance of information retrieval systems is fairly uneven, and so results are usually presented as the average precision at fixed levels of recall across all queries in a collection.*

Calculating a precision value for an information retrieval system is relatively simple: the user looks at the texts and decides which texts she or he actually want. The precision is the number the user wants, divided by the number the user has looked at. Finding recall is more difficult in that the calculation requires knowledge of *all* relevant documents in the collection. Having the user exhaustively search the collection to find all relevant material is an expensive task in

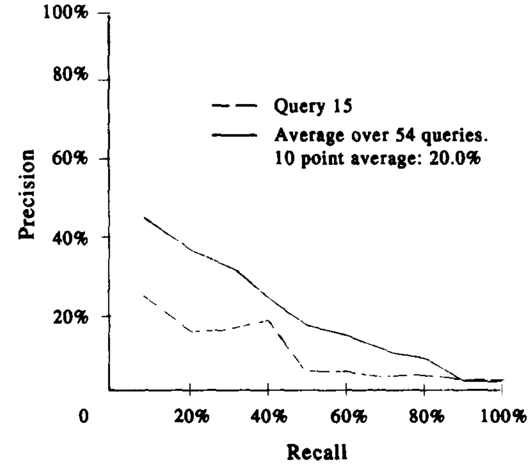


Fig. 3. Precision at various levels of Recall for Query 15.

* Variations on this are sometimes used, e.g. one of the evaluation figures produced for the TREC experiments interpolates for each query high precision figures to all lower levels of recall. This evaluation procedure, trec_eval(), is available from the ftp site ftp.cs.cornell.edu with the SMART information retrieval system.

document collections of significant size. Methods for predicting the number of remaining relevant documents using statistical methods have not been successful (Miller, 1971). An alternative to exhaustively searching the document collection is what is known as the pooled method. Using this test method, two or more information retrieval systems are used on the same query with the same document collection. The top *N* documents from each system are pooled, and judged for relevance. The judge does not know which system found which documents. The relevance judgments can then be used to compare the *relative* recall of the various systems. This is the approach taken in the TIPSTER project and the related TREC project (Harman, 1992). In some cases researchers have claimed that, by using a range of different information retrieval systems, the pooled method has found "the vast majority of relevant items" (Salton *et al.*, 1983). If this assumption holds, then the collection and relevance judgments can be used to assess other systems without having to reassess retrieved documents.

Often the performance of a system is summarized with a single figure that is the average precision at all levels of recall. This provides a crude but easy way to compare the performance of different information retrieval methods. Consider the performance of the information retrieval system plotted as the solid line in Fig. 3.

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.44	0.36	0.30	0.24	0.18	0.15	0.10	0.08	0.06	0.06

When these are averaged, the performance of this particular system can be summarized with the single figure, 20.0% average precision on a 10 point scale.

Averaging precision figures introduces a bias towards precision. First consider our keyword system ranking documents for query 15. The top 50 documents in the ranked list is represented below, from left to right, with dots representing non-relevant documents and hashes representing relevant documents.

...#.........#.....#.#............................

The average precision for the system on this query is 9%. Now consider two hypothetical systems. The first system finds a high ranking relevant document sooner (i.e. given an even higher rank) so improving precision. Let us assume it moves the first relevant document in the list from position 4 to position 1:

#............#.....#.#............................

This improvement almost doubles the average precision of the system to 16.2%. The second system improves recall by making low ranking relevant documents, which are unlikely to be seen by the user, more accessible. Let us assume it finds all relevant documents in the top 50.

...#.........#.....#.#......................######

This improvement once against almost doubles the average precision of the system to 16.1%. This system is far more use to the author of the query who wanted to find all material and, even if the user is only after the first relevant document, the other system is of questionable benefit. The difference in utility of the two mechanisms is not shown when average precision is used to summarize system performance, and the averaging process introduces a distinct bias toward the relevance of the first few documents.

INFORMATION RETRIEVAL AND ASSESSMENT OF RECALL

Lesk and Salton's argument is based on some documents being *certainly* relevant, and another set of documents being *possibly* relevant. There is little doubt that any user will want to see the documents which are certainly relevant first, and this problem has been the focus, often unwittingly, of the last 40 years of information retrieval research.

But what about users who want to find *all* relevant documents? One approach has been to get assessors to provide judgments at various levels of relevance (Sparck Jones & van Rijsbergen, 1976), and then incorporate these judgments into an assessment process. However, degrees of relevance are not a simple thing to express to judges, nor does there seem to be any reason to choose two levels of relevance over say five. Frei and Schauble (1991) have proposed a test mechanism that takes this to its extreme and compares two ranked lists of documents.

Given the long history of assessment using a binary classification, and well understood methods of assessment using such a classification, we propose the use of conventional assessment methods, but with a set of relevant documents specifically designed for testing recall. From the discussion above, a reasonable choice for this set contains all documents *possibly* relevant to the query. The method for attaining such a set, advocated here and used later in this paper, is to have several people make independent judgments of relevance for each query, and then take the union of these judgments as the set of possibly relevant documents. Testing for recall is then simply a matter of using standard measures of recall and precision on the new relevance judgments. Note that the perfect high recall information retrieval system would not be considered perfect by any individual user, but that is in the nature of the problem, and not a fault of the mechanism or the assessment process.

Note also that given multiple judgments, an explicit set of certainly relevant documents for each query can be had by taking the intersection of the various sets of documents judged relevant. This would provide a less comparative measure of precision and provide researchers with some idea of the scope for improvement there is in a (precision based) retrieval system.

Lesk and Salton's (1969) conclusions were incorporated in the creation of test collections such as CACM, and supported the use of other collections. The importance of these collections has receded given the recent TIPSTER initiative. TREC has several appealing features including significantly more documents, unambiguous queries, and relevance assessment done by professionals. We believe however that these advantages introduce problems that must be considered. The TREC project grew out of an interest in routers and filters sponsored by DARPA, the U.S. *Defense Advanced Research Projects Agency.* Although there are considerable similarities between such mechanisms and information retrieval systems, the people using filters can have considerably different interests to those wanting an information retrieval system. We have already suggested that many doing information gathering are, unlike Lesk and Salton, going to be interested in finding more than just the first few relevant documents.

The TREC project's professional relevance assessors, in combination with the extended "topics" rather than queries, give the meaning of "relevance" a conciseness unattainable with test collections assembled with volunteer judges. Although the TREC project provides an environment with fewer disputes about what is relevant and what is not, this is not a feature of the average library catalogue search. *Ad hoc* queries are not only short (a point that has recently been addressed by the TREC organizers with the use of the summary field but the same relevance judgments) they are also, we claim, inherently ambiguous. "Parallel algorithms" is a genuine *ad hoc* query, no matter how scientifically unappealing it may be. Proper evaluation of *ad hoc* queries can only be carried out with multiple judgments and naive users.

The size of the TREC document collection also introduces problems as making exhaustive relevant judgments on a collection this big is not feasible. Instead the pooled method is used to compare competing systems. This is expensive in that each new comparison requires more documents to be judged, and thus the number of groups participating in the TREC competition must be limited. However, researchers who are not able to participate can use the TREC documents and queries, and the released relevance judgments. A problem with this is the status of unjudged documents. There is a tacit assumption that any documents that have not been found by the contractors and participants are not going to be relevant. If one believes that keyword retrieval is effective at finding significant numbers of the actual relevant documents, then, with enough participants, all the relevant material will be found. But this line of argument relies on one's faith in keyword retrieval being good for high recall. The more divergent new systems become from the participating systems, the more likely it is that a new system will be finding relevant, unjudged documents. Thus, care must be taken when using the released relevance judgments to test novel systems.

Even when a system is participating in TREC, there is an inbuilt bias toward conservative experiments. Each participating system gets to contribute 100 documents to the pool of documents to be assessed. Each system is then given a performance rating based on the top 1000 documents it retrieves. A system which finds radically new relevant documents will be fairly compared with the other systems for the first 100 documents, but what happens if many of the remaining 900 are also relevant? Unless the set of found documents of the new system is similar to that of the other participating systems, the new system will be at a distinct disadvantage. Although, once again, average recall/precision figures will not be unduly affected, if one is interested in recall, then the lower ranked relevant documents are important and taking the average will not reflect the true performance.

In summary, it is not possible for an information retrieval system to achieve perfect retrieval (100% precision at 100% recall) because different users have different ideas as to what is relevant. Without looking into the minds of each user, the ideal information retrieval system would retrieve all documents certainly relevant (those in the intersection test set) followed by all those thought relevant by at least one judge (those in the union) followed by the rest. The performance of such a system would be difficult to assess in absolute terms, and so we advocate using two distinct relevance sets for testing purposes: the intersection of the relevance judgments for those research projects focusing on precision, and the union of these sets for those interested in high recall. The ultimate system would perform well on both sets of relevance judgments.

COMPARING TWO SYSTEMS FOR RECALL

The remainder of this paper illustrates the proposed assessment process by comparing two retrieval mechanisms using a new set of relevance judgments designed for assessing recall. In this section we describe a new information retrieval mechanism that we expect to give better recall.

The semantic signature mechanism is based on the assumption that relevant documents will *partially paraphrase* the users query. There is thus room to improve information retrieval systems by having them recognize when the same idea is being expressed in different words. Natural languages such as English allow great diversity in the way ideas are expressed. Syntactic variations can be dealt with formally by, for instance, converting all texts to their active form. In information retrieval, a less formal and quicker mechanism is used and all word order is removed and texts are treated as sets of words. This works, we argue, because the semantic structure of a text is primarily carried by the lexical preferences associated with the words themselves. As an example of this process, there are only two ways to assemble a meaningful sentence from the words "break", "with", "police", "door", and "sledge-hammer"— and only one is likely. Variations in word meanings have been tackled using thesaurus-like mechanisms. This approach, however, does not capture paraphrases in which single words are replaced with longer texts. The semantic signature mechanism attempts to deal with variations in the meaning/text mapping at the text level by constraining the vocabulary in which ideas are expressed. Keyword retrieval of documents written in a restricted language, by queries written in the same restricted language could significantly reduce the number of misses caused by variation in language use. The constrained vocabulary in which semantic signatures attempt to capture document content is that used in the definitions in the Longman Dictionary of Contemporary English (LDOCE). These definitions have been written from a vocabulary of approx. 2000 words.

Information retrieval (unlike for example machine translation) is a forgiving process, so accuracy is not essential in the translation of documents and queries to their new representation. The technique used relies on the assumption that the definitions in LDOCE have been written using Liebniz's substitutability criteria for a good definition, and that the definition of a given word can be substituted for the word in a text without changing the meaning of the text.* Naturally substituting LDOCE definitions for words in text does not result in a particularly readable text but this does not matter. The aim is to imitate a system that paraphrases both

* Liebniz actually wanted the *truth* of the text to remain unchanged rather than its *meaning*.

documents and queries in a language with a restricted vocabulary, and then use conventional information retrieval on the new representations of the documents and queries. As the information retrieval mechanism will ignore all word order anyway, there is no need to generate syntactically correct texts in the new representation of the document or query.

In terms of the vector-space model, conventional keyword retrieval places documents and queries in N-dimensional space, where N is the number of unique words in the document collection. When the cosine similarity measure is used, the *similarity* (*relevance*) of a document to a query is the cosine of the angle between the appropriate vectors. When a keyword information retrieval mechanism is used, the *features* (or dimensions) are the words used to describe the concept. The features need not be defined by a one-to-one mapping with the set of distinct words, and several attempts have been made to improve the mapping from words to features before the documents and queries are placed in the vector space (Deerwester *et al.*, 1990; Gallant *et al.*, 1992; Wallis, 1993). The major problems with implementing this technique are in devising a reasonable set of better features, and devising a mapping from the words appearing in the text of documents and queries, to a suitable feature-based representation. The semantic signature mechanism uses LDOCE to solve both these problems.

Mapping words to dictionary entries is not straightforward. Two filters are used: one to remove affixes, and another to select the required sense of the word. Both these procedures have been observed to sometimes degrade retrieval performance* and so the keyword mechanism we use has the affixes removed and words are tagged with their homograph number. This process is compared with a conventional stemming filter in Fig. 4. The keyword test takes the original text and replaces each word that appears in LDOCE with its sense-selected root word. As an example, query 25 from the CACM test set asks for

"Performance evaluation and modeling of computer systems".

This text is passed through the above filters to become the set of sense-selected words:

$\{ \textit{performance}_2\ \textit{evaluate}_1\ \textit{model}_8\ \textit{computer}_1\ \textit{system}_4 \}$.

There is now a one-to-one correspondence between the terms in the text and definitions. When the words are replaced with the appropriate definition for each term, the text becomes the set of "primitives":

act action before character music perform piece
calculate degree value
model
calculate electric information machine make speed store
body system usual way work

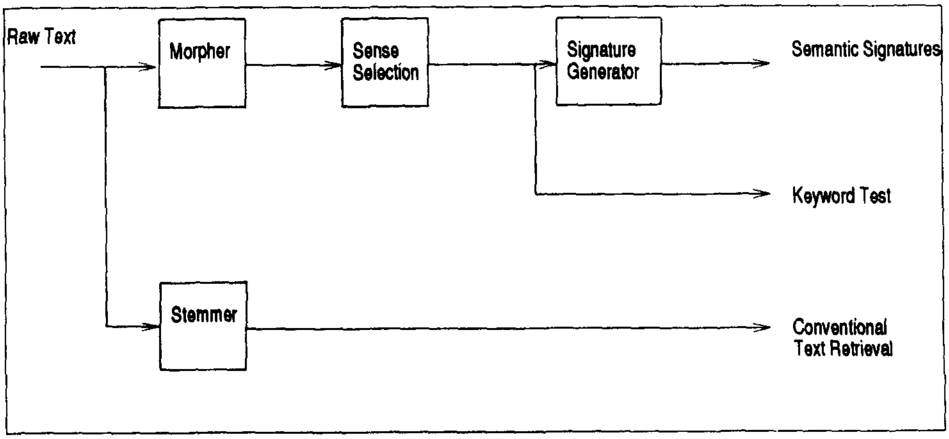


Fig. 4. Text processing requirements for semantic signature tests.

* Recent tests by the authors on TREC suggest this is not always the case.

To illustrate the utility of combining relevance judgments in different ways, we would have liked to take a collection of documents and a set of queries and have several people make relevance judgments on each query and all the documents. This is an exceedingly labour-intensive task and so we have restricted the size of the experiment. We chose instead to use the existing CACM document and query test collection because it is widely known and used by many researchers. The aim of the experiment is to create relevance assessments suitable for information retrieval systems that emphasize recall, and to test the provided relevance judgments for CACM against those documents that are certainly relevant.

There are 3204 documents in the CACM collection; these documents consist of titles and, in most cases, abstracts. In order to limit the number of documents each person would need to examine, we have followed the TREC procedure and used the pooling method of system assessment. That is, each system is run on the collection with each query; the best *K* documents according to each system are collected and judged as relevant or not. The precision for each system is then calculated in the usual manner, and a comparative result for the recall can be given. In the TREC experiments the number of documents examined from each system, *K*, is 100 for TREC participants, and 200 for contractors. The relevance assessments are made by an expert provided by The National Institute of Standards and Technology. In our experiments *K* is 80, and the judgments are made by the authors of this paper. The information retrieval systems we compare are the semantic signature mechanism described above, and the keyword mechanism using sense-tagged words. There are 64 queries in the CACM test collection, and between 80 and 160 documents per query requiring consideration, which was still too many relevance judgments to make. We set out to select approx. 10 queries that had similar performance for each system when compared using the supplied relevance judgments. We also wanted the chosen queries to be representative of the overall performance of each system. Both systems achieve around 20% overall average precision on a 10-point scale, and so all queries that had an average precision of 20±5% on both systems were chosen for these tests. This gives seven queries: 6, 7, 19, 20, 25, 36, and 61. The queries themselves are shown in Fig. 5. This selection of queries, although not a random selection, provides a relatively diverse range of query styles, and a significant variation in the amount of relevant material.

The results of the relevance judgments reported here and characterized below were attained by having the authors of this paper make the relevance judgments. We believe this does not introduce a significant bias because, using the pooled method, it is difficult for the judges to know which system provided which documents. Although Judge-1 and Judge-2 found significantly more relevant material than the SMART assessors, the overlap between Judge-1 and the SMART judge is empty for query 6, and neither new judge identified all documents

Query id		number of relevant documents (provided with SMART)
Q6	Interested in articles on robotics, motion planning particularly the geometric and combinatorial aspects. We are not interested in the dynamics of arm motion.	3
Q7	I am interested in distributed algorithms — concurrent programs in which processes communicate and synchronize by using message passing. Areas of particular interest include fault-tolerance and techniques for understanding the correctness of these algorithms.	28
Q19	Parallel algorithms	11
Q20	Graph theoretic algorithms applicable to sparse matrices	21
Q25	Performance evaluation and modeling of computer systems	51
Q36	Fast algorithm for context-free language recognition or parsing	20
Q61	Information retrieval articles by Gerard Salton or others about clustering, bibliographic coupling, use of citations or co-citations, the vector space model, Boolean search methods using inverted files, feedback, etc.	31

Fig. 5. The 7 queries from CACM used in these experiments.

judge(s)	Qry 6	Qry 7	Qry 19	Qry 20	Qry 25	Qry 36	Qry 61	sum
smart	3(3)	16(28)	8(11)	3(21)	21(51)	13(20)	21(31)	85(147)
*jdge*1	3	34	40	13	48	18	55	211
*jdge*2	7	40	38	11	61	22	47	220
*jdge*1 ∩ *smart*	0	15	8	2	20	9	19	73
*jdge*2 ∩ *smart*	3	16	8	2	21	11	17	78
*jdge*1 ∩ *jdge*2	1	32	37	9	45	15	43	182
*jdge*1 ∩ *smart* ∩ *jdge*2	0	15	8	2	20	9	17	71
*jdge*1 ∪ *smart* ∪ *jdge*2	9	42	41	16	64	27	61	260

Fig. 6. Agreement of relevance judgments—Wallis and Thom.

judged relevant by the SMART assessors even though both chose more than twice as many relevant documents.

Figure 6 provides the same information as that provided for the Lesk and Salton (1969) experiment. The intersection of the two judges' relevance judgments is often larger than the original set of judgments, and this, once again, is presumably a product of the small number of relevance assessors participating.

We compared the semantic signature mechanism with the keyword mechanism using three different test sets of relevance judgments: the original CACM relevance judgments; the intersection of our judgments; and the union of our judgments. The second test set is appropriate for comparison of keyword and semantic signature mechanisms when precision is the emphasis, and the third test set is appropriate when recall is important. The results are presented as the amount of relevant documents after fixed numbers of viewed documents. We do not use precision at levels of recall because, using the pooled method, the performance of either system does not indicate anything about the overall number of relevant documents in the collection. We cannot, therefore, calculate actual recall. Figure 7 plots precision over the first ranked 50 documents for the seven queries, using the original CACM judgments, on the two systems. When the number of relevant documents is summed for the 10 first ranked documents for each system, the keyword mechanism finds one more document than the semantic signature mechanism. Over the first 50 ranked documents the performance is about the same. This is not surprising in that the parameters of the semantic signature mechanism were chosen to give the best performance using the original relevance judgments. Figure 9 shows the same comparison when the intersection of the two sets of judgments are used. This is the test aimed at high

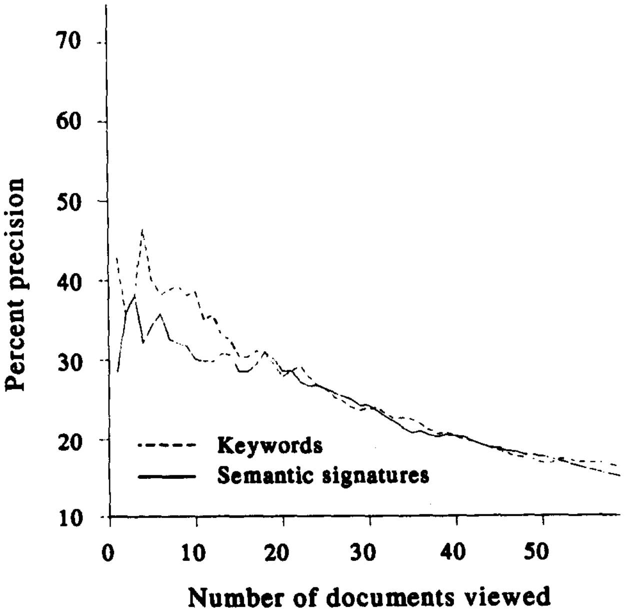


Fig. 7. Semantic signatures vs keyword: assessed with SMART package relevance judgments.

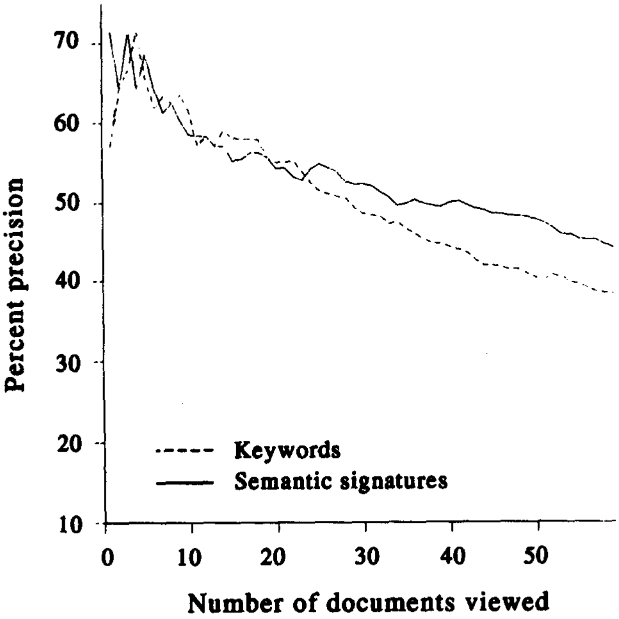


Fig. 8. Semantic signatures vs keyword: high precision test.

precision and, although the performance does not vary in accord with the SMART judgments, more relevance judges, would reduce the size of the intersection and presumably reproduce the results seen in Fig. 7.

Figure 8 compares the two systems for high recall. In this case the performance is about the same for the first 20 documents, but then the semantic signature mechanism starts to find more documents. By the time the user has looked at 50 documents, the user has found an average of 4 more documents per query. This represents about a 20% increase in the number of found relevant material. If the user is interested in finding more later, rather than a few sooner, the semantic signature mechanism has a significant advantage. This performance does not appear to drop off. We have examined the ranking up to the 80 document level and the semantic signature

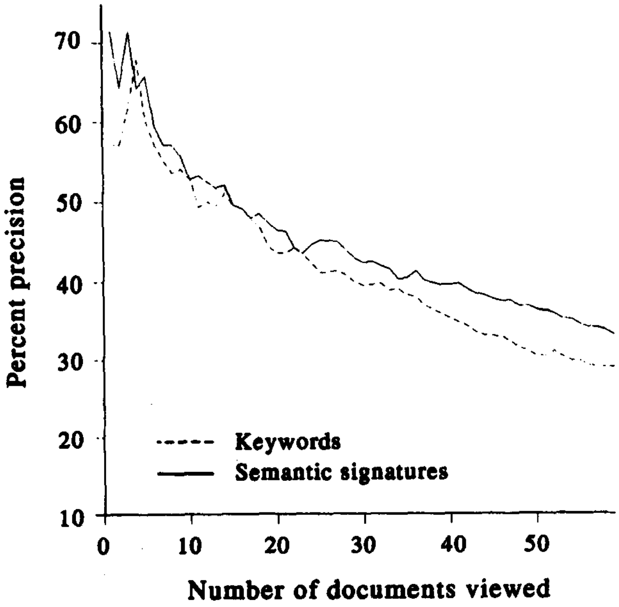


Fig. 9. Semantic signatures vs keyword: high recall test.

mechanism maintains a solid 20% advantage. Based on the new set of relevance judgments we conclude that the semantic signature mechanism is worthy of further investigation in the context of high recall. This contrasts sharply with the conclusion drawn from the results based on the SMART relevance judgments.

CONCLUSION

Keyword retrieval of texts is very effective at what it does, and the history of information retrieval suggests that it is very hard to find something better. There are things though that keyword retrieval cannot do and one of those is finding all relevant material. High recall systems are perhaps not as generally useful as systems which emphasize precision, but in several applications there is a need for better recall. There would appear to be considerable room for interesting reseach on the high recall problem. We have shown that there is a bias in the way information retrieval systems have been assesed, and that this bias means the effectiveness of high recall systems has not been apparent. We advocate that recall be treated as a separate problem with its own testing procedures.

The word "relevant" means different things depending on, amongst other things, which individual is asking, and the degree to which the individual is interested in finding all relevant material. The variation between individual ideas of relevance is objectified by combining multiple judgments, and different needs for recall determines the way the relevance judgments are combined. An ideal information retrieval system that found *all* relevant documents for an *ad hoc* query, would find all and only the documents in the *union* of the sets of all users' relevance judgments. Each individual user would of course consider the precision of such a system to be less than perfect, but this is a byproduct of the nature of relevance. It is indeed misleading to think that a system might simultaneously attain 100% recall and precision.

The relevance judgments in existing test collections are suitable for testing systems which emphasize precision, however for testing high recall the relevance judgments in these test collections cannot be used. In the future we hope there will be test collections available for testing high recall mechanisms, with diverse naive queries, and multiple exhaustive relevance judgments. Without such a collection, future work on information retrieval mechanisms for high recall must be run using the pooled method of assessment, with multiple relevance judgments on one of the existing *ad hoc* query test collections—a very expensive process.

Acknowledgements—We would like to thank Justin Zobel and the anonymous referees whose comments we found helpful in revising this paper.

REFERENCES

Buckley, C., Voorhees, E., & Salton, G. (1988). The SMART information retrieval system, version 8.8. Fetched from ftp site ftp.cs.cornell.edu.

Cleverdon, C. W. (1974). User evaluation of information retrieval systems. *Journal of Documentation, 30*(2), 170.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*(1), 19–37.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing & Management, 26*(5), 629–640.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

El-hamdouch, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal, 32*(3), 220–227.

Fagan, J. L. (1987). *Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods*. Ph.D. thesis, Department of Computer Science, Cornell University, Ithaca, N.Y.

Frei, H. P., & Schauble, P. (1991). Determining the effectiveness of retrieval algorithms. *Information Processing & Management, 27*(2), 153–164.

Gallant, S. I., Caid, W. R., Carleton, J., Hecht-Nielsen, R., Qing, K. P., & Sudbeck, D. (1992). HNC's MatchPlus system. *SIGIR Forum, 26*(2), 2–5.

Harman, D. (1992). The DARPA TIPSTER project. *SIGIR Forum, 26*(2), 26–28.

Janes, J. W. (1994). Other people's judgments: A comparison users' and others' judgments of document relevance, topicality and utility. *Journal of the American Society for Information Science, 45*(3), 160–171.

Lesk, M. E., & Salton, G. (1969). Relevance assessments and retrieval system evaluation. *Information storage and retrieval, 4*(4), 343–359.

Miller, W. L. (1971). The extension of users' literature awareness as a measure of retrieval performance, and its application to MEDLARS. *Journal of Documentation, 27*(2), 125–135.

Salton, G., Fox, E., & Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM, 26*(12), 1022–1036.

Saracevic, T. (1975). RELEVANCE: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*, 321–343.

Smeaton, A. F. (1987). *Using parsing of natural language as part of document retrieval.* Ph.D. thesis, Department of Computer Science, University College Dublin, The National University of Ireland.

Sparck Jones, K., & van Rijsbergen, C. (1976). Progress in documentation. *Journal of Documentation, 32*(1), 59–75.

Su, L. T. (1994). The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science, 45*(3), 207–217.

Swanson, D. R. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39*(2), 92–98.

Thom, J. A., & Wallis, P. (1992). Enhancing retrieval using the Macquarie thesaurus. In *1st Australian Workshop on Natural Language Processing and Information Retrieval*, pp. 111–117. Monash University, Clayton, Australia.

Wallis, P. (1993). Information retrieval based on paraphrase. In *Proceedings of the 1st Pacific Association for Computational Linguistics Conference*, pp. 118–126. Simon Fraser University, Vancouver, Canada.