Reassessing and extending the Precision and Recall concepts

M. H. Heine

Department of Information and Library Management, University of Northumbria
Newcastle upon Tyne, England
michael.heine@unn.ac.uk

Abstract

The contrivances of ‘Recall’ and ‘Precision’ are customarily used to assess the effectiveness of document retrieval systems. Despite their extensive use in experiments, including the recent TREC experiments, and their dominance in mathematical discussions of system performance, there has been continual questioning of their validity since their introduction with the Cranfield experiments. This has largely centred on critical analysis of the (pre-mathematical) concept of ‘relevance’. Those analyses have now led to the near-consensual view amongst relevance theoreticians that two types of relevance require to be distinguished, namely (1) document ‘topicality’, where cognition acts as a largely passive receiving agent of knowledge that has an objective, *a priori* or public character, and (2) ‘psychological relevance’, where cognition is more actively and creatively involved in the knowledge encoded in the document, i.e. where relevance is non-public (subjective) and conditioned by the user’s context and experience (etc) at a particular time. (Various synonyms for these terms have been suggested.) The continued use of *P* and *R* in their initial, largely Cranfield, form in document retrieval experiments, especially the continued, uncritical, use of ‘Recall’, suggest that experimentalists have largely failed to internalise this distinction, since (1) Recall is meaningless under one of these viewpoints, and (2) Precision is ambiguous when the separate validities of each are recognised. Several ways in which this distinction should influence the design of future experiments on document-retrieval/cognition interactions are suggested, involving the choice of, and amendments to the definitions of, these basic measures. Lastly, a generic 3-valued vectorial approach is suggested as a means of integrating both perspectives on ‘relevance’ within a common evaluative framework.

1 Introduction

Human life is governed by personal and public knowledge, and the feedstock and determinants of knowledge are the processes we term ‘information’. Our individual and social systems have found it necessary (or at least efficient) to embody aspects of the knowledge in individual minds in, and as, ‘documents’, which may be regarded as representing semi-permanent, coded, and public representations of knowledge. It is the proliferation of the objects we call documents, coupled with our perceptions that we need access to them, that have led to the discipline we call ‘information storage and retrieval’ and to the technologies and systems•computer based, organisation-based, and tacit knowledge based•that attempt to facilitate such access.

The research topic we term ‘the evaluation of IR systems’, which has attracted so much experimental and theoretical work, is motivated by the individual and social realities just referred to. It is, however, a party to various as yet unresolved tensions and uncertainties, perhaps inevitable in any young discipline, and even a defining property of one? At a broad level, these uncertainties relate to the *kind of thinking* that we should be bringing to the topic. Should we, for example, turn to the social sciences or computer science, say, for our thinking models, or should we develop these *de novo*?. At a more specific level, there has been a longstanding concern as to what the most suitable *variables* are with which to characterise search effectiveness. Some of the issues here are:

1. Whether numeric variables (‘measures’) should be used *at all* to measure system effectiveness, and whether the whole field has not been prematurely mathematised.
2. Whether just one or two optimal variables (whether numeric, logical or categorical) should be used to characterise search effectiveness, or many variables? Given the subtleties, complexity and dynamics of knowledge and information•and people•should we not be taking more of a ‘horses for courses’ stance in choosing our evaluation variables? Perhaps we have too quickly turned our back on the social sciences, with their concerns for contexts, openness in system definition, and recognition of system uniquenesses and non-

replicability, because of our loyalty towards ‘scientific’ (i.e. reductionist) thinking? Have we been lured into self-consistent but possibly inapplicable, closed, systems of logic?

3. Whether we are evaluating the delivery of *documents* to the human searcher, or (more ambitiously) ‘information’ (whatever that might be).
4. Whether the description of the system being evaluated should admit subjective descriptions *by the experimenter* of the user’s mental processes or whether to do this would risk taking us into a new ‘phrenology’ of the brain, or even (in one complexion of ‘theory’) to a [useless] ‘spinning of wheels’, to use Schamber’s phrase, that simply distracts from the more fundamental needs for objective, empirical data.
5. What heed our system descriptions should pay to the ‘real life’ contexts of searching: to users’ ‘real worlds’, leading under one reaction to that question to a doubting of the value of *any* experimentation *not* linked to the complex, dynamics of real lives.

In this paper we take the view that the *P-R* paradigm, although now almost the conventional wisdom of the discipline, as evidenced by its domination of both post-Cranfield experimentation, and the treatments given by the standard monographs in the field, requires radical reconsideration. In effect, we will argue for (1) the future rejection of the *unqualified* use of Precision, and (2) the replacement of ‘Recall’ by an alternative measure—the number of retrieved documents that inform the user—in *one* category of experiments.

Such reconsideration need not be destructive, but should require us:

1. To revisit what we mean by ‘information’ in the contexts of document use by the real-life user, and to ask whether the process of *learning* is not a more appropriate interpretation of this concept (rather than—as custom and practice imply—‘fact assimilation’.)
2. In particular, to question the *total* validity of the fundamental Cranfield assumptions that information should be regarded as having an ‘objective’, i.e. ‘extra cognitive’ and directly observable character.
3. Further to the latter, to ask why, given that it was the human mind that created the coded representations of knowledge we call documents in the first place, we have ignored creativity by the human mind *at the point of use of those representations*?
4. To ask (this time rhetorically) whether the term ‘information *retrieval*’ has not bewitched our thinking (to use a phrase of Wittgenstein’s) as to what is really happening, at the cognitive level, when people search for and use documents, and indeed bewitched our thinking in designing what we have boldly called ‘experiments in IR’.
5. To ask what we can do, at a basic scientific level, to make the Cranfield paradigm less inhospitable to ideas of knowledge interaction (rather than simple ‘fact transfer’, say) at the point of the use, and hence, if at all possible, amend or extend the *P-R* paradigm so as to accommodate this idea.

The views put forward here have been influenced by many earlier papers published on search-system evaluation and relevance. General, recent, critical discussions or reviews that were found to be particularly useful were [1-4]. More specific reviews and discussions examined were, or were included in, refs. [5-10]. The literature on experiment design in particular is instanced by [10-11] and commentaries on recent experiments include [12-13]. The discussions [3,5-6,14-15] were particularly stimulating. The review by Salton [16] was useful for its detachment and oversight.¹ More specific references are given below. There is no implication in the citations that any of these authors necessarily agrees with the views here expressed, although citations attempt to give credit for prior work. There is also no implication that citation is exhaustive, and comments, critical or otherwise, on that or other respects would be welcomed.

2. Pre-mathematical concepts

2.1 Primitive concepts: ‘information’ and its ambiguity, and ‘relevance’

To reduce the confusion that so easily arises when discussing relevance and information, we *prescriptively* define ‘information’ for our present purposes *as a process that is created by the human mind in response to sensory data*. Harter [7] may have been the first author to construe information in process terms, in the document retrieval context, although the concept may arguably be traced to earlier writings of Shera, Swanson, Buckland, and others. (Harter cites Buckland [18] in regard to the ‘process’ term.) When emphasis seems appropriate, we will refer to ‘information’ as ‘an information process’. The ‘sensory data’ of interest to us are, of course, the data perceived

¹ Salton’s note that Fairthorne [48] proposed the *number of documents retrieved* was interesting, although apparently not the *number of informing documents retrieved*, as we will draw attention to here. See also [44-45].

when a searcher reads (i.e., assimilates, studies) a *document*. The objective evidence for an information process resulting from an inspection of a document, is *verbal behaviour by the searcher to that effect*. Theory in our field would surely be strengthened, not weakened, by our adopting such a user-centred approach.

However, there seems to be no need for a modeling of processes that can be perceived *only* subjectively, i.e. intra cognitively, by (higher-order) cognition. It is appreciated that some authors do see such modeling (i.e. models resulting from introspection) as valid, but clearly we cannot build a *scientific* discipline on unobservables such as 'need' even if this intuitive idea, like 'context', remains a useful primitive in our thinking, i.e. a useful but necessarily undefined root concept. (An analogy could be 'life' for a biologist, who does not *define* the term but rather deals with well-defined terms such as DNA, genes, habitats, etc.) 'Information' in these terms, is thus equivalent to, or at least a close synonym for, 'learning' taking the latter term at its widest meaning, and it is (to repeat) objective verbal behaviour expressive of learning effects that provide us with the basis for assessing 'information' in these terms. If we say something is 'informative', then we mean that we have *learned* from it, i.e. adapted ourselves. Not just 'plugged it in' as a 'fact' (i.e. acceptable datum) to a predefined cognitive database.²

We see 'relevance', on the other hand, as a strictly superfluous (i.e. redundant) concept, since in effect it simply refers to those documents that are party to a searcher's information process. It is *retrospectively* defined by observed user behaviour. However, it plays a useful intuitive role, and there seems to be no harm in retaining it in pre-formal discussions, given the present, evolving state of the discipline.

Granting the concept of 'relevance' a provisional validity, we note that there is now a near consensus, in writings on this concept, that two different sub-concepts can usefully be distinguished. Just what these sub-concepts are, from a *document-use* point of view, we shall attempt to review below, but we note in passing that the *terms* used to describe one of them include 'situational relevance' (Wilson [15]) or 'pertinence' (Kemp [19]) or 'psychological relevance', 'subjective relevance' [7], and in the case of the other sub-concept: 'topicality', 'weak relevance' or 'objective relevance' [7], 'aboutness' (Hutchins [20]), although Belkin [21] has pointed to ambiguity in usage of the latter term, 'logical relevance' (Cooper [22]), although Cooper here used the term in a rather more restricted way than this, or (simply) 'relevance' (as in the Cranfield experiments.) The 'default' meaning of 'relevance' in the discipline, at the present time, could be said to be the latter, perhaps again underlining the tall shadow cast by Cranfield thinking on experimental design.

When we use the implicitly complex phrase 'search for information' we accordingly mean, under our definition of information-as-process, a search for documents that would *contribute* to a learning process. The reason for emphasising the word 'contribute' here is that information *qua* process will clearly be influenced by other factors than the recorded knowledge in the document, for example the searcher's memory, cognitive skills, age and genetic-endowment, state of fatigue at the time of search, social and physical context, disposition to reassess views presently held, etc. We elaborate further on this matter below.

The problem of *defining measures that represent the information process* in regard to its support through document provision, is then perhaps one of the most central in our science. Sub-problems at the micro level include those of choosing suitable *performance* measures, and the macro level include the representation of *overall* user satisfaction following the document *aggregation* that is determined by retrieval. The classical probabilistic measures of performance ignore the latter, the whole is seen as simply the sum of the parts.

Many persons in the IS&R field—we may perhaps, with respect, refer to them as 'the Cranfield School'—have *not* regarded 'information' in the above 'process' way. Rather, they have seen information as an inherent (i.e. observer-free, or 'public') property of a document, conditional on a verbal statement of need but not otherwise conditional on the circumstances of the searcher. To such authors, information has the character of objective *data*, just as recorded knowledge itself—e.g. the flat page of text—less contentiously has this character. This definition or concept is clearly 'the Cranfield' one, and the widespread usage of the term 'information retrieval' rests on it, perhaps so much so that few of the experimenters in our field appear to have recognised the magnitude of the question that the term 'retrieval' begs. Only Rees, perhaps, in the foundational days of our discipline, expressed scepticism in this regard [23] through his preference for the term '*document retrieval*.' The mainstream of monographs appears to have kept to the Cranfield view, implicitly supporting the concept of 'information as data', i.e. as an objective, search-question specific, document-embedded, attribute, rather than as a cognitive process owned by an individual searcher at a particular time, and with which retrieved documents interact. With hindsight, and with respect to earlier researchers,

² The closest parallel to the information concept, when we see information in this way, is 'education'. The debates in this discipline between education as a 'transmission process', akin to Cranfield thinking, and education as something mutually interactive between an adaptive cognition and recorded knowledge, then parallel our own. See for example [49] where the latter is referred to as the 'studio model'.

Figure 1: A schematic for the Cranfield paradigm

one feels that reappraisal of the whole area, as facilitated by the Mira meetings for example, is now overdue.

To illustrate the above definitions, if that is not 'overkill', we might consider a document in the form of a library of music CDs. These each record human knowledge, albeit knowledge of a particular kind. The *effects* on a human user (i.e. listener) of each and any CD constitute 'information' as we use the term in this paper, and we note that *these effects do not reside on the CD tracks*. (The CDs do not know the listener's age or musical experience, or the context in which he or she is hearing them, which are presumably factors influencing the searcher's reaction to them.) Historically, however, in the classical 'IS&R experiments', information has *not* been used in this way. Rather, it has been seen as an objective property of the CDs, so that if, for example, a searcher said 'Which CDs contain Radiohead tracks?', then those CDs with Radiohead tracks would be asserted to 'contain information relevant to the searcher' whether the searcher liked them (i.e. 'saw them as relevant') or not. There is no question here of the *effect* on the searcher of their being played when retrieved, being taken into account. One approach to relevance is process oriented and user-specific, the other is objective or topical.

2.2 The CRAN Model of Retrieval

To summarise our view, we note that the dominant paradigm of the discipline to date stems from the Cranfield experiments [24], and is one where the evaluation of a search for information appears to be based on the following key assumptions. In stating these, we keep to our *present* usage of the term ‘information’, and *not* to the Cranfield concept.

- The searcher knows, in advance of the search, *exactly what he or she wishes to know*, i.e. that the information process to which he or she wishes to be party is accurately perceived by the searcher.
- The information process concerned is *stable*, i.e. does not influence the searcher’s subjective perception of it when it is invoked (as evident by verbal behaviour)
- The searcher’s own knowledge of that information process *can be rendered externally to his or cognition*, through verbal behaviour expressed to an experimenter (or intermediary) and so provide a basis for reliable experimental procedures by which documents in the store being searched can be labelled as either contributing to that information process, or not contributing, when delivered to the searcher

As we saw in the preceding section, this paradigm sees ‘information’ *not* as a cognitive process, but rather as an entity capable in principle of operational definition *prior* to retrieval taking place, and as such observable outside the searcher’s mind. (We repeat that the terms ‘storage’ and ‘retrieval’ within the phrase ‘information *storage* and *retrieval*’ impose this meaning!) ‘Information’ in the language of the Cranfield experiments, is thus seen as an observable entity which is somehow embedded in or imbues documents. It assumes that it is something that is ‘there for the taking’, that it is user-independent or, to use a computing term, that it is not searcher ‘aware’. The CRAN paradigm may then be reduced to the assumption that the knowledge growth of the searcher (i.e. the searcher’s ‘learning’) is of the character of simple fact gathering.. Interaction between cognition and document at the point of use is minimal. Information is seen as concretised, i.e. separated from the database searcher’s mind, and treated much as a fisherman might think of his potential catch, or a child might think of an object he has been told is in a bran tub (prior to digging in same.) The paradigm assumes a non-adaptive or passive (even ‘dumb’) learner, via its stipulation that the verbal criteria betokening ‘success’ in a document search *are declarable in advance of the search*, whether the documents that match those criteria can be found or not. The CRAN model is illustrated in Figure 1.

But to avoid a baby+bathtub scenario, let us immediately acknowledge the undoubted *part* validity of the CRAN model. As a practical example that *supports* this model, we might imagine a situation in which a searcher declared a need “to know the main functions of commercially-available computer-based learning tools” and then tries to ‘retrieve’ this information via a store of document records (or full-text documents.). His or her need would then be seen as met (for that particular store of documents) when all the documents that contained one or more descriptions of such functions were retrieved, and those that did not were not retrieved. The Cranfield paradigm seems unassailable here, and one might reasonably turn to the workhorses of *P* and *R* for evaluating the effects of whatever search expressions our user might input to the store. We are also absolutely clear that if a document could be found that (say) *questioned the sufficiency* of computer-based learning tools, this would be dismissed, under this paradigm, as ‘outside scope’, i.e. ‘not relevant’. Similarly, a document in which its author pointed to functions of computer-based learning tools that *would* be useful if they were commercially implemented, but noted that they were not yet commercially available, would (by definition) also be deemed ‘non-relevant.’ One might feel rather sad for the searcher, that he or she had cast their need in such rigid, dogmatic terms, but if in conversation it was confirmed that he or she had indeed asked for *exactly* what they wanted, and that they were not interested in negotiating their need or the possibility of creatively responding to new knowledge, then any adverse consequences of the rigidity of their criterion would be their problem (if it is a problem) not ours.

2.3 The searcher adaptation and response (SAR) paradigm

Here, the view maintained by the paradigm is that information *qua* process is the only admissible view. Relevance is not about ‘topics’ or ‘subjects’, but is a responsive process situated in the mind of the individual searcher at search time, and influenced by the searcher’s individuality (memory, skills, experience, aptitudes, etc) and immediate context in the world (organisational, professional, private, social). Here, we do not see information as objects to be shunted around somewhat mechanically, to be lifted from a document and then ‘input’ to a mind, but rather as an entity ‘created on the fly’ by human intelligence, and involving creative, adaptive partnership between mind and recorded knowledge, i.e. ‘document’. Adopting this paradigm then leads ineluctably to one conclusion: ‘*Recall*’ under this paradigm is a meaningless construct. The Recall concept views knowledge as *a priori*, i.e. definable prior to document delivery, as the everyday meaning of the phrase ‘to recall something’ conveys. The numerical

evaluation of ‘Recall’ in experiments accordingly depends on it being possible to say which documents *would* contribute to the information process *were they to be retrieved*. But within this paradigm information is seen as a process private to (or owned by) the individual, context-situated, searcher, *so that such documents are unobservable in principle*. (The searcher sees only *retrieved* documents.) To repeat, ‘Recall’ is then by definition meaningless.

The effect of seeing relevance and information in these terms is thus to translate what has possibly seemed to be a somewhat academic analysis of the concept of relevance into both (1) a determinant of experimental design (pre-tagging of documents as ‘relevant’ or ‘non-relevant’ is then a meaningless procedure; relevance judgement by proxy is also meaningless; verbal behaviour in the searcher at the point of use needs to be sought in an appropriate way), and (2) a disenfranchising of one of the key and traditional variables that experimenters and others have used to describe system effectiveness.

In case the argument for the SAR paradigm seems thin, the following quotations are offered in support of it. All appear to support the view that ‘information’ should not be seen (or at least seen *solely*) as an object, but rather as a ‘process’.

“Part of our knowledge we obtain direct; and part by argument....” J. M. Keynes, 1921 [25] quoted by C. J. van Rijsbergen [26].)

“Within the framework of a theory of thinking, concepts and the process of concept-formation appear not as independent entities but as aspects of the thought process. Concept-formation... is intimately connected with memory-organisation, anticipation,, cognition and other mental processes and functions.” (Shera, 1957 [27])

“An ‘information need’ is thus revealed to be a dynamic entity whose times of greatest dynamism and change may come in the very process of interacting with a retrieval system.” L. B. Doyle, 1963 [28] quoted by D. Ellis [29: p.18]

“Unless we predicate discussion of system effectiveness on criteria of effect, we shall continue to be, in Alan Rees’s phrase, “busy people spending large sums of money, designing•or attempting to design•phantom systems for non-existent people in hypothetical situations with unknown needs.... a criterion does not stand or fall on its measurability, nor are incidental system attributes valid criteria of system value simply because they can be quantified. Users’ satisfaction, a difficult concept to operationalise because of its two undefined terms, is none the less, in our view, the ultimate criterion for evaluating an information retrieval system...” (Paisley and Parker, 1965 [14])

“Only the user himself may judge the relevance of documents to him and his uses...” (Saracevic, 1970 [5].)

“...for traditional measures there is no coherent explanation of why hidden document properties should be taken into account at all.” (Cooper, 1976 [6])

“The ‘relevance of a document is...taken to be a piece of new knowledge constructed by the requestor in the light of some information need or deficit. This knowledge can be described only as a subjective experience of the requestor. In this sense, relevance is not something already present which is to be measured or judged; it is created. It is not and cannot be a property of a document and a request...” (Swanson, 1977 [17])

“There is no such thing as the relevance of a document to an information requirement, but rather the relevance judgement of an individual in a specific judging situation recording his judgement... at a certain point in time.” (Rees, 1966 [23] quoted by Schamber et al. [8].)

“A definition of relevance that relies on fixed-for-all-time, unchanging relevance judgments•such as those characterizing nearly all retrieval tests that have been conducted to now•must be seen as wrong.” (Harter, [7].)

Earlier work which has (explicitly or implicitly) moved the paradigmatic viewpoint away from that which sees relevance as an *a priori* construct towards an *a posteriori* one, also notably includes [6], where Cooper sees the set of retrieved documents as the only “reality” of the 2x2 table in operational systems; and the author’s construction of ‘Retrievability’ [30] as the retrieval analogue of Salton’s ‘Generality’. Although apparently not rejecting the idea of information as having a *a priori* status, Sparck Jones [31] has noted that different ‘functional perspectives’ on information can be recognised, wherein [information need] can be prompted by “quite distinct reasons, ...as in e.g. writing a paper, indexing a paper, refereeing a paper, or seeking a paper.”

However, the early criticisms of Cranfield, such as those by Swanson [32] and Harter [33], largely focus on *the methods used to assess Recall*, rather than on *the validity of the Recall concept itself*, i.e. on the meaningfulness of tagging non-retrieved documents as ‘relevant’. Various recent writings, e.g. [34-37], have also addressed

Figure 2: The SAR paradigm in which learning and relevance decisions are linked

themselves largely to the same question, *and therefore have again assumed the validity of the CRAN paradigm*. From our point of view, this shows insufficient awareness of the literature on relevance, which has increasingly cast doubt on the CRAN paradigm, seeing it as, in a sense, ‘lying alongside’ another one, which we have called SAR. (The author claims no innocence in this respect, his own experiments [39-40] having also treated Recall uncritically.) Returning to the bathwater problem, we nevertheless need to accept that in some search circumstances the Recall concept may be regarded as valid, so that highly focused discussion on it (as is offered by Su [41] for example) is then justifiable. Patent searching is usually mentioned in this regard, but even under CRAN, a patent searcher may be looking for just one prior patent that might contest the eligibility of a claimed novel invention, and under SAR a patent database searcher doing a novelty check on a firm’s proposed patentable invention, might identify a patented invention that does *not* contest the novelty of the proposed patent but *is* clearly of interest to the research team that created the invention. Buckland [42] provides examples where just one retrieved relevant item will apparently suffice (e.g. where the system user needs to know “the population of Klagenfurt in 1900, the cosine of 36°, Nelson’s last words...”). (I have added ‘apparently’ here since there may be disagreement between sources as to what the population of Klagenfurt was in that year, and perhaps the angle of 36° is on a sphere, and perhaps a document contains Nelson’s last words but they are in Bulgarian and the user wanted them in English.)

If, in response to the relevance theorists, we agree that the assumptions of CRAN are unwarrantedly strong, i.e. acceptable in some instances of information need but not all, and see cognition as *actively* involved in *creating* information in at least some document searches, then the results of evaluative studies that use these assumptions are of little value, as Swanson has suggested. Similarly, discussions that focus on the mathematical properties of Precision and Recall can be seen to be of only limited use to the managers of real-life information management systems•since they deny the reality of a learning searcher. The SAR is illustrated in Figure 2.

Lastly, in case the reader remains convinced that human cognition *does* always act according to the Cranfield assumptions, i.e. that it can always accurately foresee what it needs rather than (at least sometimes) creatively responds to what it beholds, we would draw analogies from the social sciences. The demand for medical services

tends to be strongly influenced by the *availability* of medical services, e.g. the number of available hospital beds, i.e. is ‘supply driven’ at least partly. The demand for books from public libraries has been shown to correlate strongly with the amount spent on books by the local authority. Again, demand (need) appears as strongly supply driven. We might also turn to advertising, where the whole of the industry is based on the principle that people’s perceived needs for products and services (including many information products and services) are based on information *about* those products and services proactively disseminated to them. Again, demand (i.e. perceived need) is supply driven. Last, we might turn to technology in general. How many individuals, surveyed in the 1970s, say, as to their needs for new personal technology, would have expressed needs for the items we now call ‘pagers’, ‘mobile phones’ or ‘Walkmen’? The perceived need (relevance) of such items seems almost totally supply driven. Such analogies readily convince one (if personal experience of document database searching does not!) that there will be documents in a database that are *not* perceived to be relevant *at the time of query articulation*, but which will be perceived as relevant if brought to the searcher’s attention. Ergo, the SAR paradigm is not just a speculative possibility, but a very real facet of individuals’ cognitive dispositions, and indeed one wonders why the alternative CRAN paradigm has become so entrenched in our thinking.

2.4 The CRAN and SAR paradigms and feedback systems

It is interesting to ask how the CRAN and SAR paradigms stand in regard to systems that support ‘relevance feedback’. (For a recent review, see Harman [43], the concept, of course, having originated with Salton.) The answer would seem to be complicated, and require distinctions between a ‘learning searcher’ and a ‘learning search expression’. In some instances of cognition/database interaction, the feedback may be directed at search expression optimisation. The searcher in this situation can be either a learning searcher (the SAR paradigm applies) or one whose needs remain fixed with successive iterations of the search. If the feedback is directed at clarification of the need that is driving the search process, then by definition we are within the SAR paradigm.

3. Revision of *P* and *R* under the CRAN and SAR paradigms, and experimental realities

3.1 Reassessment of Precision

In general, subscripts should be given to effectiveness variables to signify the paradigm under whether the experiment is conducted. For example, if the experimenter has asked the searcher to flag all those retrieved documents that inform him, irrespective of their similarity or not to the initial (pre-search) statement of need by the searcher³, then this would determine a Precision value with a particular character that might be denoted by, say, P_{SAR} . If, on the other hand, the experimenter asked the searcher to flag all the retrieved documents that were in conformity with the original (pre-search, verbal) statement of, and recognition of, need, then this would determine a Precision value with a different character, which could be denoted by P_{CRAN} . With ‘Precision’ in general, i.e. under either viewpoint (learning searcher or ‘dumb’ searcher), we are attempting to capture the intuitive notion of ‘noise’ in the retrieved documents.

3.2 Reassessment of Recall

Under CRAN we can, more or less comfortably, capture the notion of *informativeness in the set of retrieved documents*, using ‘Recall’. This seems reasonably uncontentious where the search for knowledge (or rather ‘the enablement of informing’) is at the low level of fact-searching, however sophisticated the expression of the wanted fact may be. In that case, the paradigm admits no notion of information support other than which is rendered through the capture and scrutiny of pre-flagged (or rather ‘pre-search flaggable’) documents. Accordingly we can continue to use *R*, or if we wish ‘ R_{CRAN} ’ for emphasis, in this situation. The more important problem is rather one of identifying a substitute for *R*, i.e. a substitute measure of ‘that which supports the information process’, under the SAR paradigm. The author suggests that, with SAR, we move to a simple frequency measure rather than a probabilistic one, namely: *the number of documents in the set of retrieve documents that inform the searcher*, and

³ Obviously, in a real experiment, the nature of the experimenter’s cue that prompted searcher behaviour would need to be developed less casually than this. The experimental design in this regard would need to balance the wording of the cue against prompting for ‘Skinnerian operant’ search behaviour (‘What interests you today?’ perhaps) and prompting for search behaviour against known, recorded, workplace or livingplace objectives. Our wording is only meant to be suggestive.

denote this by N . We suggest the term ‘informativeness’ and will use this in the formal definitions below. *We repeat here for emphasis the earlier point that Recall is a meaningless construct within the SAR paradigm.*

The reader may object that N is a redundant variable since if we know the Precision (under either paradigm) and also know the size of the set of retrieved documents, then N is immediately determined. (In other words, N equals $P_x k_i$, under either paradigm, where k_i is the size of the set of retrieved documents on the i th search.) However, *this is true only for a specific search instance* (i.e. for a given searcher, fixed-in-time arbitration process, and search expression.) In other words, k_i is *not* a constant across different search instances but varies from one search to another, as indeed the subscript to k emphasises. It is a random variable, just as N is, for a fixed value of Precision. N thus serves to capture a concept of informativeness that, in a sense, is parallel to ‘Recall’ but does so without having to assume that an ‘entity’ one might otherwise have labeled as information has a pre-search existence. An early reference to the idea of counting retrieved and informing documents, rather than moving to probabilities, is given by Swets [44] to a paper by Bornstein.[45].

One might reasonably argue that N , as defined, is in its way as simplistic a measure as R and P . But to accept this criticism, and reject N because of it (presumably along with R and P) would be to leave our science with an almost empty evaluative toolbox. One would then be left perhaps with innumerable tape-recording of conversations with searchers as to the effectiveness of their searches, and with little or no capacity to generalise from them in an objective manner. However, a more careful response would be to acknowledge the sense of the objection, and to say that N could be modified. We could ask the searcher to assign weights or (following Cooper) dollar-values to retrieved documents, and use aggregations of weight or dollar-value (perhaps normalised by the number of retrieved documents) as a measure to replace N . We use N notionally here in just that sense, i.e. as a placeholder measure that could (like R and P) be refined at will.

3.3 Bivariate measures

Measures such as D (further discussed by Shaw [46]) or van Rijsbergen’s E measure [47], perhaps require to be carefully handled. For example, both can be expressed in a form that ‘inherits’ the variable R , which is unacceptable under SAR. The solution would seem to be to start by re-expressing them in terms of the basic sets involved. This is noted below for D , when we see that for the SAR paradigm, D_{SAR} reduces to $1 - P_{SAR}$.

3.4 Summary of suggested new measures

The following table summarises our suggestions

	Measure of noise in the retrieved set	Measure of informativeness in the retrieved set
CRAN paradigm	P_{CRAN}	R
SAR paradigm	P_{SAR}	N

Formal definitions of our new measures, in procedural and set terms are then as follows. The definitions assume that the retrieved set is not empty. We use $n(\dots)$ to count the number of items in a set.

1. Define relevance-flagging procedures RFP-CRAN and RFP-SAR:
 - RFP-CRAN is applied *to the database as a whole*, including all and any retrieved documents. Its starting point will be verbal behaviour by a searcher, or else verbal behaviour by an experimenter who seeks to emulate searcher verbal behaviour. Searcher-proxy implementations of this procedure, for *non*-retrieved documents, will almost certainly be needed in experimental studies in view of the size of modern databases. The procedure is not altered by the searcher’s exposure to retrieved documents.
 - RFP-SAR is applied *only to retrieved documents*, and may vary from one retrieved set to another. Proxy implementations of it are not necessary in experiments. The searcher applies the procedure, private to his or her mind, directly to retrieved documents.
2. Let A be a set of *potentially informing* documents, defined in reference to a verbal statement by the searcher prior to the search. (*Meaningful only under CRAN.*)
3. Let B be a set of retrieved documents. (Meaningful under *both* paradigms)
4. Let B^* be the subset of B flagged as informing by RFP-CRAN.

5. Let B^- be the subset of B flagged as informing by RFP-SAR.
6. Then:
 - $P_{CRAN} = n(A \cap B^*)/n(B)$
 - $P_{SAR} = n(B^-)/n(B)$
 - R (or R_{CRAN}) $= n(A \cap B^*)/n(A)$
 - $N = n(B^-)$
 - $D_{CRAN} = 1 - [n((A \cup B) \setminus (A \cap B))/n(A \cup B)] = (P_{CRAN} + R - 2P_{CRAN}R)/P_{CRAN} + R - P_{CRAN}R$
 - $D_{SAR} = 1 - [n(B \setminus n(B^-))/n(B)]$, since B^- replaces A in the definition, and B^- is a subset of B
 $= 1 - P_{SAR}$

4. An integrated paradigm

In the ‘real world’, of which we hear so much, real people doing real searches can be observed to respond to searches in *both* a ‘CRAN’ *and* an ‘SAR’ way. An aircraft engineer may begin a search by looking for documents on fatigue in aluminium alloys, and yet have her attention drawn by a retrieved document that refers to bacterial corrosion of those alloys. A sixth former searching for information on the Gaia concept may find himself drawn to articles on plate tectonics. Such a searcher could be said to be operating in ‘joint mode’, interested in both the problem as it was originally expressed, and yet open to the possibility of a redefinition of that problem. The question then arises as to how we might describe search effectiveness in such a situation? The author suggests that one sensible approach would seem to be:

1. To retain set A as a *partially* valid construct, and to recognise informing documents in the retrieved set, B , as those that inform the searcher in *either* way. In formal terms, this ‘either’ is represented by the documents in $B^* \cup B^-$. Let us denote the latter set, the set of *all* documents that inform the search, whether by way of surprise and creative involvement, or by way of their matching a pre-set statement of need, by B^T (using ‘ T ’ for ‘total’.)
2. To retain Recall, R , just as it was defined under CRAN (since it remains meaningless except in that regard.)
3. To redefine N as $N^T = n(B^* \cup B^-)$
4. To re-define Precision as ‘Total Precision’: $P^T = n(B^T)/n(B)$
5. To represent document retrieval effectiveness by means of a 3-vector, comprising P^T , R and N^T .
6. Figure 3 illustrates such a 3-vector.

7. Diagrammatic representation of the three paradigms

The three paradigms we have discussed, namedly CRAN, SAR and the integrated paradigm described above, are illustrated in Figure 4.

5. Further discussion

It is appreciated that in this discussion we have both made simplifying assumptions and limited the factors being taken into account.

In the author’s view, the main factors that would need to be considered next are:

- The effect on relevance decisions of the *order* in which documents in a retrieved set are examined, when the experiment is conducted under the regime of SAR, e.g. controlled randomness in the order in which retrieved documents are presented to the subject. (With CRAN, of course, order of presentation is, by definition, immaterial.) It is difficult to see what else could be done. Presenting retrieved documents from the same search expression to the same searcher would be to ignore learning effects.
- Whether the searcher is making judgements on metadocuments (i.e., bibliographical descriptions + abstracts, say) or full-text documents (e.g. Web pages, full-text electronic articles.) But for full-text searching the assumptions are similar. But presumably this is not a ‘problem’ but simply a matter to be declared by the experimenter, who in fact may be interested in the differences in effects of such different presentations.

Figure 3: Schematic of the 'integrated paradigm'. Here P^T represents the Precision value based on the set union of documents relevant against both pre-search expressed verbal criteria, and (possibly non-verbal) criteria invoked responsively by the searcher when exploring retrieved documents, i.e. with the searcher's knowledge enhanced by both 'capture' and 'creativity'.. The classical P - R view is in part a projection of the P - R - N vector onto the P - R plane, with P suitably redefined. See Figure 4 for further clarification using Venn diagrams.

6. Conclusions

The effectiveness of document identification systems, traditionally but presumptuously called information retrieval systems⁴, has been customarily evaluated using 'Recall' and 'Precision' values, although the validity of these measures has attracted long-standing controversy. One criticism of them is centred on the nature of document 'relevance', which Recall in particular sees as being knowable, and fixed in advance of the search, notwithstanding both creative involvement of the searcher in retrieved documents, and the problems in principle of knowing what is inaccessible to the searcher.

Published discussion on the nature of relevance (dating back to the 1960s) has attempted to disambiguate this concept into sub-concepts of (1) document 'topicality' (aboutness, subject), where cognition may be regarded as a largely passive receiver of objective knowledge in the document, and (2) 'situational' or 'subjective' relevance, where the interaction of document and cognition is creative, individual-specific, and context- and time-specific. However, experimentalists in the field of document retrieval effectiveness appear to have largely ignored this distinction, as is evident by the continued, unmodified, use of Recall and Precision.

We have offered refinements to the choice and use of these measures which incorporate the above distinction, and also offered a unifying paradigm that takes both into account jointly. If the basic distinction as to types of relevance, and the suggested use of refined variables, are accepted, then it follows that future evaluation studies should: (1) state clearly which of the two definitions of relevance (information) is recognised in the study, (2) subscript the

⁴ We say 'presumptuously' since they *retrieve* 'represented knowledge', and *contribute to* 'information' as we have defined it in process terms. However, if one regards the term 'information retrieval' as an elliptical usage for the latter, then perhaps we should continue to live with the historic term?

CRAN (knowledge 'capture') paradigm. Recall= $n(A \cap B^*)/n(A)$; Precision= $n(A \cap B^*)/n(B)$.

SAR (creative interaction) paradigm. Recall is undefined; $N=n(B^~)$; Precision= $n(B^~)/n(B)$.

Integrated (hybrid learning) paradigm. Recall= $n(A \cap B^*)/n(A)$; $N=n(B^* \cup B^~)$;
Precision= $n(B^* \cup B^~)/n(B)$.

Figure 4. The three diagrams show the main sets of documents representing the three paradigms discussed. The upper diagram shows the classical (CRAN) view of retrieval, in which a set of 'relevant' documents, defined in an *a priori* manner, and the set of retrieved documents, are recognised. The middle diagram shows the effect of seeing informing documents as definable *only* following retrieval, as with the SAR paradigm. The lower diagram shows an integrated view, in which (1) cognition is informed by documents in a way that could have been anticipated (CRAN) and also (2) learns in an unanticipated way from some retrieved documents. The notation follows the text.

variable 'Precision' appropriately so as to eliminate needless ambiguity as to what this measure means in that particular experiment, (3) either adopt or reject the 'Recall' concept, as is appropriate to the aim of the experiment, and (4) in the case of investigations of situation relevance, where 'Recall' is meaningless, adopt a *new measure* of relevance, namely 'Informativeness', N , in place of R . The latter variable, as would P_{SAR} , would appear to be particularly useful in studies of those sequences of searches where relevance is treated as if it has a dynamic property.

In summary, to answer the question posed in the title of this paper, it is suggested we should *not* abandon P and R but show much more care than we have done in choosing when and how to use them.

Acknowledgement: I am grateful to Heesop Kim for identifying and supplying several of the key papers that appear to be relevant to this discussion and which I had missed. However, the responsibility for the selection of papers cited remains my own. An anonymous referee requested clarifications of a draft of this paper, to which I have attempted to respond. A colleague Adrian Gordon kindly referred me to item [49].

8. References

1. Dervin, B., Nilan, M. Information needs and uses. *Annual Review of Information Science and Technology*, **21**, 1986, 3-33.
2. Schamber, L. Human behavior and relevance. *Annual Review of Information Science and Technology*, **29**, 1994, 1-33.
3. Katzer, J., Snyder, H. Towards a more realistic assessment of information retrieval performance, In: D. Henderson (ed.) *Proceedings of the American Society for Information Science 53rd Annual Meeting*, 4-8 November 1990, Toronto. **27**, 1990. pp 80-85.
4. Spink, A., Wilson, T., Ellis, D., Ford, N. Modeling users' successive searches in digital environments: a National Science Foundation/British Library funded study. *D-Lib Magazine*, April 1998. [www.dlib.org/dlib/april98/04spink.html (visited 26 Jan 1999)]
5. Saracevic, T. The concept of 'relevance' in information science: a historical review. In: T. Saracevic (ed.) *Introduction to information science*. Bowker, N.Y., 1970. pp 111-151.
6. Cooper, W. S. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Information Processing & Management*, **12**, 1976, 367-375.
7. Harter, S. P. Psychological relevance and information science. *Journal of the American Society for Information Science*, **43**, 1992, 602-615.
8. Schamber, L., Eisenberg, M.B., Nilan, M. S. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, **26**, 1990, 755-776.
9. Wilson, T. D. Information behaviour: an interdisciplinary perspective. *Information Processing & Management*, **33**, 1997, 551-572.
10. Sparck Jones, K. (ed.) *Information retrieval experiment*. London, Butterworths, 1981.
11. Tague-Sutcliffe, J. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, **28**, 1992, 467-490.
12. Harman, D. Overview of the Second Text Retrieval Conference (TREC-2), *Information Processing & Management*, **31**, 1995, 271-289.
13. Sparck Jones, K. Reflections on TREC. *Information Processing & Management*, **31**, 1995, 291-314.
14. Paisley, W.J., Parker, E.B. Information retrieval as a receiver-controlled communication system. In: Heilprin, L. B. et al. (eds.) *Proceedings of the Symposium on Education for Information Science*, Warrenton, Va., Sep. 7-10, 1965. Washington, D.C. Spartan, 1965. pp. 23-31.
15. Wilson, P. Situational relevance. *Information Storage and Retrieval*, **9**, 1973, 457-471.

16. Salton, G. The state of retrieval system evaluation. *Information Processing & Management*, **28**, 1992, 441-449.
17. Swanson, D. R. Information retrieval as a trial-and-error process. *Library Quarterly*, **47**, 1977, 128-148.
18. Buckland, M. K. Information and information systems. N.Y., Greenwood, 1991.
19. Kemp, D. A. Relevance, pertinence and information system development. *Information Storage and Retrieval*, **10**, 1974, 37-47.
20. Hutchins, W.J. On the problem of 'aboutness' in information retrieval. *Journal of Informatics*, **1**, 1977, 17-35.
21. Belkin, N. J. Ineffable concepts in information retrieval. In K. SPARCK JONES (ed.) *Information retrieval experiment*. London, Butterworths, 1981. pp.44-58
22. Cooper, W. S. A definition of relevance for information retrieval. *Information Storage and Retrieval*, **7**, 1971, 19-38.
23. Rees, M. The relevance of relevance to the testing and evaluation of document retrieval systems. *Aslib Proceedings*, **18**, 1966, 316-324.
24. Cleverdon, C. W., Mills, J., Keen, M. *Factors determining the performance of indexing systems*. Cranfield, England, Aslib Cranfield Project, 1966. 2 vols.
25. Keynes, J. M. *A treatise on probability*. London, Macmillan, 1921.
26. van Rijsbergen, C. J. Information, logic and uncertainty in information science. *Proceedings: CoLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, October 13-16, 1996, edited by Peter Ingwersen and Niels Ole Pors*. Royal School of Librarianship, Copenhagen, 1996. pp.1-10.
27. Shera, J. H. Pattern, structure, and conceptualization in classification. In: *Proceedings of the International Study Conference on Classification for Information Retrieval*, Dorking, England, 13-17 May, 1957. London, Aslib, 1957. pp.15-27.
28. Doyle, L. B. Is relevance an adequate criterion in retrieval system evaluation? *Proceedings of the American Documentation Institute 26th Annual Meeting*. Washington, D.C., ADI, 1963, pp199-200. [reference unverified]
29. Ellis, D. *Progress and problems in information retrieval*. London, Library Association Publishing, 1996.
30. Heine, M. H. Distance between sets as an objective measure of retrieval effectiveness. *Information Storage and Retrieval*, **9**, 1973, 181-198.
31. Sparck Jones, K. Fashionable trends and feasible strategies in information management. *Information Processing & Management*, **24**, 1988, 703-711.
32. Swanson, D. R. Some unexplained aspects of the Cranfield tests of indexing performance factors. *Library Quarterly*, **41**, 1971, 223-228.
33. Harter, S. P. The Cranfield II relevance assessments: a critical evaluation. *Library Quarterly*, **41**, 1971, 229-243.
34. Frakes, W. B. Introduction to information storage and retrieval systems. In: *Frakes, W. B. and Baeza-Yates, R. (eds.) Information retrieval: data structures & algorithms*. Englewood Cliffs, Prentice Hall, 1992. pp.1-12.
35. Losee, R. M. Upper bounds for retrieval performance and their use measuring performance and generating optimal boolean queries: can it get any better than this? *Information Processing & Management*, **30**, 1994, 191-203.
36. Frické, M. Measuring recall. *Journal of Information Science*, 1998, **24**, 409-417.
37. Frants, V. L., Shapiro, J., Voiskunskii, V. G. *Automated information retrieval*. San Diego, Academic Press, 1997. (pp.279-281)
38. Wallis, P., Thom, J. A. Relevance judgements for assessing Recall. *Information Processing & Management*, **32**, 273-286, 1996.
39. Heine, M. H. Information retrieval from classical databases from a signal-detection point of view: a review. *Information Technology: Research, Development, Applications*, 1984, **3**, 95-112.
40. Heine, M. H. An investigation of the relative influences of database informativeness, query size and query term

- specificity on the effectiveness of Medline searching. *Journal of Information Science*, **21**, 1995, 173-185.
41. Su, L.T. The relevance of Recall and Precision in user evaluation. *Journal of the American Society for Information Science*, **45**, 1994, 207-217.
 42. Buckland, M. K. On types of search and the allocation of library resources. *Journal of the American Society for Information Science*, **30**, 1979, 143-147.
 43. Harman, D. Relevance feedback and other query modification techniques. In: *Frakes, W. B. and Baeza-Yates, R. (eds.) Information retrieval: data structures & algorithms*. Englewood Cliffs, Prentice Hall, 1992. pp.24-26.
 44. Swets, J. A. Information retrieval systems. *Science*, **241**, 1963, 245-250.
 45. Bornstein, H. *American Documentation*, **12**, 1961, 254. [Reference unverified]
 46. Shaw, W. M. Jr. On the foundation of evaluation. *Journal of the American Society for Information Science*, **37**, 1986, 346-348.
 47. van Rijsbergen, C. J. *Information retrieval. Second edition*. Butterworths, London, 1979.
 48. Fairthorne, R. A. Basic parameters of retrieval tests. *Proceedings of the 1964 Annual Meeting of the American Documentation Institute*, Spartan Books, Washington, D.C., 1964. 343-347.
 49. Andriessen, J., Sandberg, J. Where is education heading and how about AI? *International Journal of Artificial Intelligence in Education*, **10**, 1999 [in press]