Gedanken experimentation: An alternative to traditional system testing?*

William S. Cooper

Technological progress is generally brought about through a combination of intelligent theorizing, experimentation, and inspired tinkering. The technology of literature searching is no exception, and elements of all three have contributed to recent progress in the information retrieval field. However, without disparaging in any way the research that has been carried out, it might fairly be observed that information retrieval is an area in which by the very nature of the subject matter the theory is thin and large-scale experimentation cumbersome and often inconclusive. Perhaps, therefore, the time is ripe to start giving more attention to the third of the aforementioned alternatives, the scientifically disreputable but often surprisingly successful course of 'inspired tinkering'.

I have in mind especially a kind of 'tinkering' with retrieval system parameters through educated guesswork rather than careful full-scale experimentation. It involves the making of estimates, or 'guesstimates', based on very little data-gathering or even on nothing more than human intuition and experience combined with the few available shreds of theory. Since the process involves the vigorous use of the imagination, it might in the phraseology of the older sciences be called 'thought experimentation' or 'gedanken experimentation'. (*Gedanken experiment:* An experiment carried out by proposing a hypothesis in thought only—Webster.) Physicists, for instance, sometimes analyse at length what would happen if they were to carry out certain experiments in a freely falling elevator, but for some reason never seem to get around to the actual execution of the experiments.

As it applies to document retrieval, gedanken experimentation amounts to thoughtful, theory-guided guesswork about what is likely to make a system work most effectively. The guesswork may concern any of various decision problems or parameter-setting tasks which arise in setting up, maintaining, and using a retrieval system. Making the guesses may be the responsibility of any of several agents including the system designer, the analyst in charge of implementing the system design, the indexer, or the end user. The unique contribution of the information scientists is to suggest ways of making the

* This material is based upon work supported by the National Science Foundation under Grant IST-7917566.

guesswork more thoughtful and theory-guided than it might otherwise be—to make each guess resemble the physicists' analytical, theory-derived expert surmise more than it does a layman's initial hunch. The hope is that enlightened guesswork of this sort, though far from infallibly correct, is less likely to be mistaken than it would be in the absence of the aids proposed by the information scientist.

Gedanken experimentation is not incompatible with classical, full-scale system testing, but if successful should reduce the need for it. Ideally, the two might be combined; that is, classical retrieval tests would be made to confirm that the theory-guided guesswork proposed by the information scientist does indeed yield better retrieval results than the traditional guesswork it is intended to replace. Perhaps a modest amount of full-scale testing of this sort is called for when a particular form of gedanken experimentation is first introduced. However, at least three considerations might cause a researcher to hesitate before undertaking extensive testing along these lines. First, since it is the results of 'theory-guided' guesswork that are to be tested out, there is *a priori* reason to suppose, even in the absence of any tests, that the guesses are probably superior to traditional ways of guessing which have no explicit theoretical underpinnings. At least, the method is probably superior if its underlying rationale is sound, suggesting that it may be easier and more appropriate to undertake a critical examination of the theory than a large-scale empirical test. Secondly, as a practical matter there is likely to be a trade-off between test effort and effort spent in devising better techniques of gedanken experimentation: resources invested in the one will be lost to the other. Since gedanken techniques usually offer hope of improvement with little danger of making things worse, it would seem sensible to put the emphasis there. Finally, full-scale retrieval tests are difficult, expensive, unreliable, and often inconclusive. This suggests that the research effort devoted to them might be better spent simply in developing design ideas (such as gedanken experimentation techniques) which have a theoretically defensible basis and implementing these ideas operationally, without bothering to test them out empirically at all. From a scientific point of view that may be a heretical suggestion, but information retrieval is more a technology than a science and, as has already been pointed out, technologies often progress faster *via* a process of inspired tinkering than through programmes of formal experimentation.

In a broader perspective, what may be called for is a shift in the information retrieval field's research priorities—a shift which may already be under way—from conventional trial-and-error testing of plausible but somewhat *ad hoc* systems to the generation of theoretically more soundly motivated design ideas. The sounder the theory behind a design idea, the less the need to test it out empirically. The development of gedanken techniques is an area which seems ripe for such idea generation. We classify it here as a design rather than a testing activity, because gedanken experiments are essentially attempts to make rational judgements about design parameters of various kinds. But carrying out a gedanken experiment often involves envisioning a trivial retrieval experiment of some sort, and in that special sense might be regarded as an (imaginary) testing activity moved back in time into the design stages.

Whether it is regarded as a design activity or an unconventional

experimental one, gedanken experimentation is an approach which offers considerable hope of supplementing, and perhaps in many cases rendering less necessary, classical retrieval testing.

11.1 Theory and experiment in information retrieval

If one were pressed to describe the central 'theory' underlying document retrieval, it would be hard to do much more than list the obvious conceptual elements of the retrieval situation. A typical list would note that there must be a collection of documents or records of some kind; a population of potential searchers; that to provide them with search assistance it seems necessary to isolate certain search properties of the documents (the 'descriptors' or 'index terms') and of the searchers' information needs (usually specified in the form of 'requests' or 'queries'); that rules for matching information need properties against document properties (the 'match function' or 'retrieval strategy') are also needed; and so forth. Although some might be willing to dignify such an account with the name 'theory', it is really not so much a theory of retrieval as a review of the problem setting with suggested terminology for discussing it. Occasionally a powerful bit of real theory might surface, as for instance the theory of syntax in a scheme for automatic indexing, or Boolean Logic in the specification of certain request languages, but these have to do with special kinds of retrieval systems or their components and do not constitute an overall theory of retrieval. In fact, in the search for a general theory it is hard to do much better than to give some elaboration of the vague rule that *a system should retrieve for the user those documents most likely to satisfy him.* As scientific theories go this truism is not very impressive, but it is the only wisp of general theory we have. What was said of a recent political candidate can be said of document retrieval theory: Deep down inside it's shallow.

Perhaps partly in recognition of this paucity of theory, many researchers have turned to experimentation, and especially laboratory experimentation. As might be expected, the classical experimental approach has been fairly theory-independent, consisting essentially in the trying out of various competing retrieval schemes (including indexing methods, etc.) to see which seem to work best. The methodology involved has been ably documented in other chapters of this book, and so need not be reviewed here except to note that the difficulties to be met in drawing useful conclusions from a retrieval experiment of classical design have turned out to be much more numerous and serious than had been expected. There are sampling and other statistical difficulties; difficulties in generalizing results obtained in just one or a few test collections; difficulties in generalizing the needs of the test user population, or in the absence of a real user population difficulties in assuring the realism of manufactured requests; difficulties arising from the variability and sensitivity to test conditions of the judgements of document relevance or usefulness; difficulties in extrapolating results to real situations where something about the system or the environment is bound to be different; and difficulties arising from the interaction of various available features of the retrieval rules under test which, if at all numerous, cannot as a practical

matter be tried out in all possible combinations. There are logical and conceptual difficulties involved in choosing a measure of retrieval effectiveness from among the dozens that have been proposed, and though from the experimenter's point of view the problem can be sidestepped to some extent by the simple expedient of reporting all results in terms of several different measures, the nagging question of which measure to put faith in is not thereby answered but merely passed along from the experimenter to the ultimate decision-maker. This is only a partial listing of the hazards of experimentation of the traditional sort. It is hardly surprising, then, that on the rare occasions when the methodology used in a large-scale retrieval test has been subjected to careful independent scrutiny, the results have been far from reassuring (see especially Swanson's and Harter's critiques of the Cranfield experiments[1–3]).

Thus there is a serious question whether full-scale retrieval experiments are worth their high cost. True, it often seems possible to glean at least some hint of a useful generality from reports of such experimentation. However, the danger of mistaking an experimental artifact for a generalizable conclusion is great, and the likelihood that a test result will eventually affect the design of future information systems for the better is small. No single obstacle seems insurmountable in itself, but in combination they are formidable. I raise the question of whether traditional experimentation is worth while here partly in order to provide a foil for the other chapter writers, but partly too because it seems to me that 'inspired tinkering' may be an alternative path to retrieval progress which is both easier and likelier of success.

11.2 Probability and utility theory in system design

The truism was mentioned earlier that a system should be designed to retrieve those documents most likely to satisfy the user. This being so, a retrieval system may be regarded as a device for estimating probabilities of satisfaction, and possibly also degrees of satisfaction, the aim being to lead the user to examine first those documents with a high probability of providing a high degree of satisfaction. If there is any general underlying theory of retrieval, then, it would appear that we must seek it in a theory of probability, and in order to quantify the notion of 'satisfaction' possibly also in the theory of utility, or decision theory. The theory of retrieval *per se* may be thin, but by regarding the retrieval problem as a problem in probability (and possibly utility) estimation, we may at least endow it with the structure of a branch of applied statistics.

The probabilistic approach is of course implicit in all sensible system designs insofar as it is expected that documents in the retrieved set are more likely to satisfy the user than the rest, or in the case of ranked output that documents higher in the ranking are more likely to satisfy than those of lower rank. However, in most present-day systems there is no explicit computation of numeric probability estimates. Rather, crude qualitative criteria are applied which are thought to produce an approximate probability-ranking

effect. And when a number of qualitative clues have been combined in order to determine the rank of each document, it is sometimes far from clear that this effect is actually obtained. As a simple example, suppose that in a system accepting weighted requests document A is indexed with one term appearing in the request with weight 0.6, while document B is indexed with two terms each appearing in that same request with weight 0.3. Under a retrieval rule commonly used in such systems (the 'vector product' rule), the two documents would be given equal priority in the output ranking. Yet it is not at all clear that the two have an equal probability of satisfying the user, for the request weights do not necessarily represent estimates of probabilities or functions of probabilities, nor does the retrieval rule constitute an appropriate probabilistic computation.

In a system based more firmly on probability theory (or utility theory) this problem would be alleviated. Let us call an information retrieval system **explicitly probabilistic** if it has the following characteristics: (1) all numeric system parameters, including any request term weights, index term weights, constants used in the ranking algorithm, numbers used as linkage strength indicators in a thesaurus, etc., have clear probabilistic interpretations as estimates of values of algebraic expressions within the standard probability calculus; (2) all 'binary' system parameters, e.g. the assignment or non-assignment of an index term to a document in a system with unweighted indexing, have clear interpretations as judgements of whether the value of one probabilistic expression exceeds that of another; (3) the retrieval rule is essentially to rank the documents of the collection for the user in order of decreasing estimated probability of satisfaction to him, where the probability estimates in question are calculated from the various numeric and binary parameters already mentioned, possibly with the aid of appropriate probabilistic independence assumptions, and all on the basis of formulae derivable within the probability calculus. (An **explicitly utility-theoretic** information retrieval system would have a slightly more general definition admitting terms for expected utilities as well as probabilities.) So far as I am aware no explicitly probabilistic (or explicitly utility-theoretic) systems have ever been put into operation and exploited as such, though there is by now a growing literature bearing on various aspects of how such systems might be designed[4–8]. An explicitly probabilistic system is presently being programmed for experimental and demonstration purposes at the School of Library and Information Studies, University of California, Berkeley.

The significance of explicitly probabilistic systems for us here is that, since the system parameters have clear probabilistic (or utility–theoretic) interpretations, the task of estimating them becomes susceptible to techniques of gedanken experimentation. The fact that the retrieval rule is based on the probability calculus guarantees that these parameter estimates will be exploited to produce the best output ranking obtainable from the data available to the system. True, the output ranking will be no better than the input estimates, but neither will it be any worse. From this it may be seen that by comparing parameter estimation methods it may be possible to replace the comparative testing of whole systems by restricted data-gathering aimed directly at the question of how good the estimates are. In cases where one method of estimation is obviously more accurate than another, the need for experimental comparison is eliminated entirely.

11.3 Examples

A few simple examples of hypothetical retrieval systems based on the thought-experiment approach may help clarify what is involved.

Example 1: The indexer as gedanken experimenter

Consider a simple retrieval system capable of responding only to single-term requests, but in which the indexing of the documents is weighted—each term assigned to a document has an associated numeric value indicative of its suitability as a descriptor for the document. When a request is received, the system simply ranks for the requestor all the documents to which the request term has been assigned, in descending order of the weight of the assignment.

To make this system explicitly probabilistic one need only instruct the indexers to restrict the numbers they use as weights to the interval between 0 and 1, and to think of these weights as probabilities. Thus if the indexer thinks there is one chance in ten of the document at hand satisfying a user submitting the term under consideration, he should assign that term to the document with weight 0.1. The gedanken experiment he performs in order to arrive at such a figure might run somewhat as follows. The indexer imagines all future system users whose request is the term in question to be transported backward in time and gathered together into a room. They are then in imagination asked to read or examine carefully copies of the document to be indexed and to raise their hand if it would satisfy at least partially the information need that caused them to submit their request. The proportion the indexer thinks would raise their hands is the weight to be assigned.

A variant of this mental experiment would have the indexer imagine a future searcher under the term to be drawn at random. The indexer would then ask himself, 'If forced to make a small wager, what odds would I be (barely) willing to give in a bet that this searcher would, when the time came, find the document I am about to index to be satisfactory, given that I index it in such a way that he is led to examine it?' It is a simple matter in probability theory to translate a betting odds into an approximate subjective probability estimate; in fact for unlikely events the odds and the corresponding probabilities are almost equal. Thus if the indexer found himself willing to give 1:10 odds for satisfaction (i.e. ten to one odds *against* satisfaction), he would again be led to attach to the term a weight of approximately 0.1.

Several points are worth noting about this example. First, for the sake of a simple procedure all considerations of degree of satisfactoriness (that is, all utility-theoretic considerations) have been omitted. There are elaborations of the foregoing gedanken experiments which could take such considerations into account, if they were deemed worth while. Second, the retrieval rule—to rank by the indexing weight of the term submitted as request—is so simple that no numeric computation whatsoever would have to be carried out by the system, which could in fact be easily implemented manually. Third, although one might expect an indexer to make more useful guesses under the suggested interpretation of the weights than under no interpretation at all or a vague one about term 'importance', it would be desirable to provide him with a

little training. Some possible forms which such training might take have been proposed elsewhere[9]. Fourth, for some purposes it would be desirable to have the capability of testing an indexer's skill in making the requisite guesses; this has also been discussed elsewhere. Finally, it might be objected that since indexers cannot foretell the future they would be unable to make the required probability estimates with any high degree of accuracy, either with gedanken experimentation or without. But it was never claimed they could. It was merely suggested that there is likely to be a tendency for the numbers they come up with under the gedanken approach to be less inaccurate as probability estimates than the numbers they would otherwise come up with. And since in the last analysis an output ranking is always either explicitly or implicitly a ranking by estimated probability, any improvement in the accuracy of estimation is a step forward.

Example 2: The indexer as gedanken experimenter, unweighted indexing

Next consider an even simpler retrieval system in which the indexing is unweighted (or 'binary'), where the searcher submits a single term as his request, and where the system responds by retrieving for him as an unranked set all documents indexed under the request term. The common subject card catalogue is a system of this sort with minor elaborations.

To decide whether or not to assign a term to a document, an indexer indexing documents for use in such a system can make use of what I have elsewhere called the 'Odds-Payoff' decision rule[9, 10]. Three steps are required. First the indexer estimates the odds against satisfaction after the fashion of the mental experiments of the previous example. Second, he performs another thought experiment whose result is a judgement of how many unsatisfactory documents a typical requestor submitting the term under consideration would be willing to examine and discard as the penalty to be paid to obtain the document to be indexed. Finally, he compares these two numbers and assigns the term if and only if the latter exceeds the former.

A variant of this procedure involves substituting a standard average value for the figure obtained in the second step, thereby eliminating that step and greatly simplifying the indexing process. The price of the simplification is that variations in degrees of predicted usefulness among the documents are ignored.

Example 3: The requestor as gedanken experimenter

Retrieval requests containing user-weighted terms have been in use for some time, but the weights are usually regarded vaguely as indicators of 'importance' rather than as estimates of probabilities or functions of probabilities. Moreover, the weights are not manipulated by the system as though they had a probabilistic interpretation. Might it be possible to regard the weights as probabilistic estimates of some kind, and reformulate the retrieval rules so that the weights are treated as such and used to compute explicit estimates of the final document probabilities?

A crude system using ordinary unweighted document indexing but capable of handling request-term weights probabilistically might be designed somewhat as follows. A request consists as in most ordinary weighted-request

systems of a list of descriptors accompanied by numeric weights and separated by commas as for example

IRON 50, MANUFACTURING 20, POLLUTION 5

The positive number following each term is to be interpreted as that term's 'probability change factor' or 'precision-boosting power'—that is, as the multiplicative factor by which a document's probability of satisfaction is changed by the presence of the term on the document. Suppose for example that in a random draw from the entire collection the probability of obtaining a useful document is 0.001. Then the presence of the weight 50 on the term IRON indicates that the requestor, if he were to learn that the randomly drawn document had been indexed under IRON, would raise his personal estimate of that probability from 0.001 to 0.05. In other words, in a gedanken experiment in which he compares the density of useful documents in the whole collection to their density in the set indexed by IRON, he guesses the latter density to be some fifty times the former. The weights on the other request terms are interpreted independently in similar fashion.

The probabilistic computations needed to estimate a document's probability of satisfaction on the basis of such a request are involved and will not be presented here, though we hope to discuss them in a later publication. They require for their input not only the request weights but also some indexing statistics, specifically such data as the number of documents in the collection indexed under IRON, and the number indexed jointly under IRON and MANUFACTURING. An estimate of the total number of useful documents in the collection is needed too, but the final output ordering is not very sensitive to the value supplied for this estimate. An independence assumption of some sort is needed to circumvent the need for data on such higher order interactions as the degree of overlap in the collection among three or more terms. Paul Huizinga of the University of California has proposed that an independence assumption derived from the maximum entropy principle may be appropriate for this purpose[11].

Example 4: The system designer as gedanken experimenter

For systems where the users formulate their own requests without the aid of an information professional as intermediary, it might be unrealistic to hope for meaningful numeric values of the kind required for the query language of the previous example. A more workable system might merely require that the user attach to his request terms not numeric weights but non-numeric symbols indicative of his qualitative judgements of relative likelihood. Here for example is a six level scale of such judgements:

Symbol	*Interpretation*
A	Presence of clue would, other things being equal, make it vastly more likely that the document is useful. A clue of the strongest sort.
B	Presence of clue would make document a much more likely candidate. Clue is a typical 'good' clue.
C	Presence of clue would make it somewhat more likely that the document would prove useful.

D Presence of clue would make it slightly more likely that the document is satisfactory. A positive clue, but of the weakest sort.

E Presence of clue would make it a little less likely that the document is useful than if nothing were known about it. A mildly negative correlate of usefulness.

F Presence of clue would make document a much less likely candidate. A strong indicator of uselessness.

If a finer scale of judgements were desired, these 'grades' could of course be refined by the usual addition of plus and minus signs. Use of the scale would make the request in the previous example look something like this:

IRON B, MANUFACTURING C^+, POLLUTION C^-

Any unweighted terms in a request would be treated by the system as though they had been assigned weight B.

The qualitative weights must of course be translated by the system into numeric weights if they are to be manipulated probabilistically, and the rules of translation must be supplied to the program at the time the system is put into operation by the system designer, manager, or other analyst. It would be the responsibility of this analyst to apply for each grade on the scale a numeric value judged to be typical of the probability change factor that a user applying that grade would supply if only he had the time and understanding to do the required gedanken experimentation. The translation data supplied by the analyst might be, say, A: 200; B: 50; C: 10; D: 2; E: 0.5; F: 0.02. Such a table would allow any graded request to be transformed immediately by the system into a numerically weighted request, after which retrieval could proceed as in the previous example.

How is the translation table to be arrived at? The simplest option is for the analyst to play the role of user for a few typical requests, perform the necessary gedanken experiments to translate the grades into numbers, and note for each grade the typical numeric weight range he finds himself translating it into. However, since the translation table need only be constructed once, there is also a possibility of some limited 'real' experimentation at this stage. That is, the analyst might actually gather enough data to provide a crude empirical estimate of the probability change factors experienced for a sampling of graded clues. The data-gathering would consist in estimating for each clue in the sample the proportion of useful documents in the subset of the collection bearing that clue as opposed to the proportion of useful documents in the collection as a whole, and computing the actual probability change factor from this data.

This sort of data-gathering would, alas, resurrect *some* of the difficulties inherent in classical experimentation. The need to establish an empirical criterion of relevance or usefulness and to apply it to many documents would be chief among these, and may often be a sufficient obstacle in itself to discourage the effort. However, it is important to note that many of the worst difficulties of classical experimentation simply do not arise in the limited, focused kind of data-gathering envisioned here. For instance, since there is no comparison of retrieval performances, the problem of choosing a measure of retrieval effectiveness is avoided. In fact, it may be misleading even to

refer to such data-gathering as an experiment, since the aim is merely to obtain crude estimates of certain statistics rather than to test anything. Here is the kind of small-scale empirical investigation by which many conventional 'What-works-best?' tests might well be replaced, if indeed there is to be any experimentation at all beyond gedanken experimentation.

11.4 Further remarks

These examples by no means exhaust the possibilities inherent in an approach based on probability and utility theory and gedanken or small-scale experimentation. There are ways of combining gedanken weighted indexing with gedanken weighted requesting, of constructing thesauri which weight relationships among terms probabilistically by thought experiment, of translating boolean requests into probabilistically weighted ones, and so on.

One of the most far reaching advantages of the probabilistic approach to system design is that it provides a natural means of combining large numbers of weak clues. Many kinds of evidence could be brought to bear in ordering system output that are not exploited in conventional systems, but which it would be natural to utilize in a probabilisitc system. Among them are the many kinds of relatively weak clues available even before a request is received, e.g. document recency, citedness, language, level of technicality, form of publication, and so on. These could all be used with low weights as part of the probability computations and would for many kinds of requests be apt to bring about greatly improved retrieval. Known-work searches on the basis of non-standard clue-types constitute another possible application[12]. There is much scope for further investigation in this area.

11.5 Summary

When a retrieval system design is explicitly probabilistic or utility-theoretic, its parameters are endowed with a clear meaning which makes their estimation a fit subject for gedanken experimentation or in some cases small-scale statistical estimation techniques. Since by virtue of the statistical theory embodied in them such systems are known *a priori* to make optimal or near-optimal use of the data at their disposal, comparative tests among whole systems of this kind may be largely replaceable by tests of the accuracy of their associated input data estimation methods, or in obvious cases by simple judgements of which of these estimation methods is probably most accurate. This suggests as potentially advantageous an approach to information retrieval research which (1) emphasizes the discovery of explicitly probabilistic or utility-theoretic retrieval system designs; (2) emphasizes the development of improved input estimation methods including gedanken experimentation techniques; and (3) de-emphasizes the role of traditional comparative system tests in favour of restricted data-gathering aimed at measuring error of estimation in the input data.

Gedanken experimentation, as opposed to actual data-gathering, is apt in general to be most valuable where decisions must be taken quickly, frequently, and with a minimum of fuss. Indexing and request-weighting

decisions would appear to be of this sort. Design decisions which can be taken in more leisurely fashion and are important enough may in some cases be worth in addition a little data-gathering effort. Experimental evaluations of the relative effectiveness of whole systems (or of aspects of systems tested within whole systems) along with some widely accepted approaches to system design should probably be rethought to see if they cannot be reformulated in an explicitly probabilistic or utility-theoretic way which reduces the need for full-scale experimentation.

Some may view gedanken experimentation with alarm, feeling that it is a retreat from scientific certainty to wild guesswork propped up by an occasional counting exercise. This attitude would be understandable, but I suspect it greatly overestimates the reliability and usefulness of classical experimentation in our field, and underestimates the potential value of theory-supported system design and theory-guided thought about its input.

11.6 Acknowledgements

I am indebted to M. Buckland, K. Sparck Jones, M. Maron, and P. Wilson for their incisive but constructive critical commentary on an earlier draft of this chapter.

References

1. SWANSON, D. R. The evidence underlying the Cranfield results, *Library Quarterly* **35**, 1–20 (1965)
2. SWANSON, D. R. Some unexplained aspects of the Cranfield tests of indexing performance factors, *Library Quarterly* **41**, 223–228 (1971)
3. HARTER, S. P. The Cranfield II relevance assessments: a critical evaluation, *Library Quarterly* **41**, 229–243 (1971)
4. MARON, M. E. and KUHNS, J. L. On relevance, probabilistic indexing, and information retrieval, *Journal of the ACM* **7**, 216–244 (1960)
5. BOOKSTEIN, A. and SWANSON, D. R. A decision-theoretic foundation for indexing, *Journal of the American Society for Information Science* **26**, 45–50 (1975)
6. VAN RIJSBERGEN, C. J. *Information Retrieval*, 2nd edn, Butterworths, London (1979)
7. ROBERTSON, S. E. and SPARCK JONES, K. Relevance weighting of search terms, *Journal of the American Society for Information Science* **27**, 129–146 (1976)
8. ROBERTSON, S. E. The probability ranking principle in IR, *Journal of Documentation* **33**, 294–304 (1977)
9. COOPER, W. S. Indexing documents by gedanken experimentation, *Journal of the American Society for Information Science* **29**, 107–119 (1978)
10. COOPER, W. S and MARON, M. E. Foundations of probabilistic and utility-theoretic indexing, *Journal of the ACM* **25**, 67–80 (1978)
11. GOOD, I. J. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, *Annals of Mathematical Statistics* **34**, 911–932 (1963)
12. COOPER, W. S. The potential usefulness of catalog access points other than author, title, and subject, *Journal of the American Society for Information Science* **21**, 112–127 (1970)