THE EXTENSION AND APPLICATION OF SWETS'S THEORY OF INFORMATION RETRIEVAL

M.H. Heine MSc, MSc(Inf.Sc.)

Thesis Presented for the Degree of Doctor of Philosophy

Computing Laboratory
University of Newcastle upon Tyne

June 1981

PAGE
NUMBERING
AS
ORIGINAL

Contents

Acknowledgements

The author acknowledges and thanks the following persons and organizations for assistance in the preparation of this thesis: Dr. P.E. Lauer of the Computing Laboratory, University of Newcastle upon Tyne, who supervised the thesis and gave constructive advice on the scope and structure of this document; Miss Elizabeth Barraclough, Executive Director of the Computing Laboratory, who discussed very helpfully many of the details of the research described, and facilitated the acquisition of data for the experimental part of the project; the MEDLARS indexing staff of British Library Lending Division who gave incidental but indispensable advice on various aspects of MEDLARS use; and the staff of the National Library of Medicine, Washington, D.C., who also gave indispensable assistance in supplying machine-readable copy of MEDLARS records.

The work described herein is the original work of the author (except where earlier work of other persons is explicitly cited) and has not previously been submitted for a degree.

Abstract

The thesis comprises (1) a critical interpretation of Swets's contribution to information retrieval, (2) development (i.e. "extension") of the formalism, as so interpreted, and (3) a description of an experiment that identifies hypotheses consistent with the extended formalism. The early sections of the thesis place the original contribution by Swets in the contexts of both signal-detection theory and information retrieval theory. It is then argued that as the original theoretical contribution is ambiguous in key respects, an interpretation of it is necessary. The interpretation given constitutes an initial development of Swets's work but other developments, not simply a consequence of the interpretation of the original description by Swets, are also put forward. The major one of these is the explicit incorporation in the formalism of logical search expressions. Elementary logical conjuncts of search terms are seen as (1) being weakly ordered by "document ordering expressions", and (2) having probability-pairs attached to disjunctions of them defined by the ordering. A major part of the thesis is the identification of novel hypotheses, expressed within the extension of the original formalism, which relate to triples of: (1) instances of information need in medicine, represented by prespecified partitionings of a medical-literature data base (MEDLARS), (2) an analytical document ordering expression, and (3) an algorithmically-derived set of terms characterising the information need. An enhancement is suggested to data base management programs that at present employ only user-specified logical search expressions by way of search input, this enhancement stemming directly from the extension of the original formalism. The broad conclusion of the thesis is that when the original contribution of Swets is suitably interpreted and extended, a robust, hospitable conceptual framework for describing information retrieval at the macroscopic level is provided.

1. INTRODUCTION

1.1 Context of the present study and summary of its contribution

The term 'information retrieval' in the title of this thesis refers to the problem and process of identifying, in a set of records of objects (the 'data base'), a subset that matches as closely as possible some prescribed subset. The latter is agreed to be specifiable only through enumeration of its members, not through its members bearing an attribute not borne by members not in the subset. (Were this not the case, the problem would be a trivial one.) An example makes this clearer. A set of records relates to car components, the records containing information on such attributes as colour, cost, size, supplier, etc., but not on the nature of the material, say. An information retrieval problem, as distinct from a data base management problem, would then be that of forming an inventory of components made totally of copper, say, using only the information actually recorded to do this.

The above problem has been given much attention by workers in the areas of computing and librarianship/information science where the records of interest are descriptions of documents. Here the problem is that of selecting from the attributes of documents as have been assigned to them by an 'indexing' process those that will best identify the documents actually sought. The latter are referred to as 'relevant' documents. Just as individual inspection of car components would indicate whether each was made of copper or not, so it is assumed that inspection of all document records would indicate which was relevant or not - notwithstanding the possibility of inconsistency in such judgements if the process were repeated. The problem of information retrieval has assumed greater importance

since (1) the actual search process was 'delegated' by human to computer, with the advent of computer-accessible data bases in the 1960s; (2) the coded descriptions of documents used in the search process became more complicated; and (3) the difficulty of distinguishing relevant from non-relevant documents increased as the sizes of data bases increased. This greater importance is perhaps due most to a perceived need for less ad hoc, i.e. more controlled, retrieval procedures, and to a related need for a scientific knowledge as to the accuracy of such procedures.

This thesis looks at one approach to information retrieval contained in the increasing body of literature on the subject put forward in the 1960s. This was a contribution by J.A. Swets, put forward in 1963. Swets provided, it is maintained, a simple, hospitable, conceptual framework for information retrieval which, when suitably interpreted and extended, allows rigorous, controlled investigation of the phenomena it describes. The main value of the framework (or 'formalism' as we shall refer to it) is in (1) the tight distinction it makes between 'information need' and 'query', with 'relevance' being seen as attaching only to the former; (2) the fundamental importance it attaches to the partitioning of the data base by the information need; and (3) its joint treatment of (a) information retrieval as such and (b) the matter of retrieval effectiveness: both process and result are treated together. It will be maintained that the formalism as a formalism has two main features. First, hypotheses expressed in it are capable of 'falsification' in the classical positivist sense: it is not in any sense a metaphysical theory. (Put another way, prediction is possible using hypotheses expressed in the formalism.) Secondly, it is a macroscopic formalism in that although it describes information retrieval

in a particular way, it does not seek to account for the properties of that procedure by appealing to the existence of laws or relationships at a deeper level. (It offers a description of how things are, rather than why they are.) This is not to say either that Swets's basic formalism is not hospitable to concepts other than those originally introduced by him, or that it would not be profitable to combine Swets's formalism with other more 'microscopic' formalisms with benefit, just that the formalism in both its original form and the extended form has a natural macroscopic character.

This thesis is concerned not only to clarify the contribution made by Swets, although it is concerned to do this where the original presentation was ambiguous or insufficient. It also attempts, as implied by its title, to extend and apply the formalism. The main extension offered by the writer is the re-expression of the formalism in terms of discrete random variables, rather than in terms of the continuous random variables used by Swets. These are related to elementary logical conjuncts of search terms. A second extension is created by introducing more than one type of record attribute. The applications of the formalism that are introduced are to the following problem-areas. First, the identification of optimum logical search expressions (from optimum search queries expressed as sets of record attributes); secondly the use of document age as an indicator of documentary relevance in weighting expressions; and thirdly the use of the formalism in providing improved search queries when partial information on the success of a predecessor query is known.

This thesis also includes a description of an experiment, consistent with the extended formalism, designed to generate hypotheses (expressed in the formalism) relating to a particular data base, to particular ways of defining relevance and search query, and to

particular forms of another variable that we shall introduce. Since the hypotheses are *generated* by the experiment they cannot at the same time be *evaluated* (falsified) by it, but they are expressed in this thesis in a form that will allow them to be contested in later work. The experiment is novel in methodology in that the sets of relevant documents with which it works are defined by objective, behavioural evidence, and in that the queries that form part of the retrieval processes examined are generated in a controlled and algorithmic manner.

1.2 The Structure of this Dissertation

The first substantive part of this dissertation, Section 2, is a literature review. This has been restricted to the key theoretical papers and experimental reports which prompted the project, and such other papers as relate closely to the present study. The approach in this section attempts to be indicative rather than analytic. Section 2 is followed by the part of the dissertation that relates to the first major objective of the research: the analysis and extension of Swets's theory. In that the analysis there includes detailed analysis of Swets's published work (as Section 3.2), and in extending the theory attempts to cite all relevant work by other authors, this section is in part of an analytical-review character. The reviewing function has thus been apportional between Sections 2 and 3 in what seems to be the most useful way. The 'kernel' of the thesis, so far as the extension of Swets's formalism is concerned, is contained in Section 3.3.2.3.

The third substantive part of the dissertation, Section 4, is a description of an experiment undertaken to generate hypotheses expressed in the formalism developed in Section 3. The results of this experiment and their analysis are presented in Sections 4.4 and 4.5 respectively.

The general conclusions of both the theoretical and experimental investigations are given in Section 5. This is expressed in terms of the concepts and notation developed in Section 3, and as such is more accurate and complete than the intuitive summary given in Section 1.1, and the interim summaries at the end of Section 3.2, and forming Sections 3.4 and 4.5. Some suggestions for further research are also given in Section 5, along with a brief discussion of the technological implications of the findings.

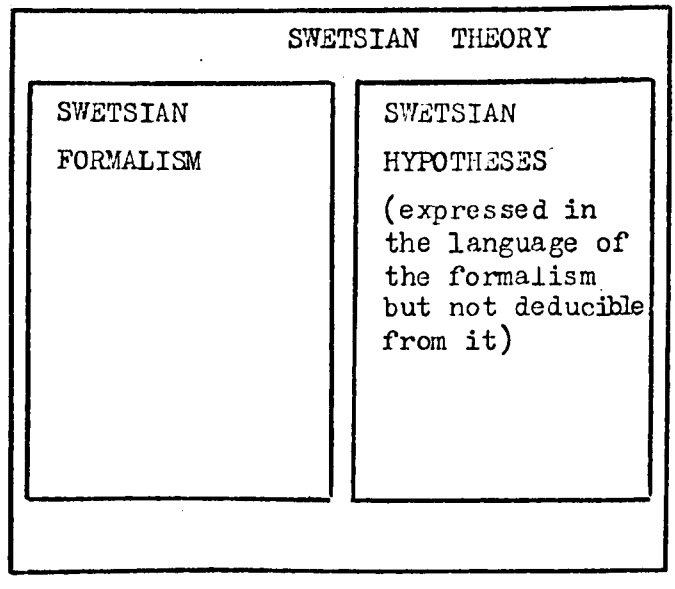
1.3 Terminology and Notation.

The concepts of 'information' and 'information need' will be treated as primitive entities. This stance is consistent with the approach implicit in Swets's work, but it is also one that is independently accepted as a basis for describing and extending his theory. The phenomenon of interest will be the assertion of relevance: the marking of documents as sources of information in reference to some perceived information need. This phenomenon is observable in principle and may be observable in practice or in an experiment. The notion of information need will not be explored beyond this - the observed behavioural fact embodied in an assertion of relevance will be taken to be the object of interest. This fact may be taken by the reader to be evidence of something deeper, but it is the fact alone that provides the main variable of the theory. In that sense and others, Swets's theory should be regarded as a macroscopic theory.

The phrase "information retrieval" has been kept to. It might be argued that document-handling systems do not present the user with information of interest (e.g. documents) but instead only references to documents (and possible abstracts). This seems a pedantic objection however, and the established term is freely used.

Unfortunately no common unambiguous usage attaches to the terms 'theory', 'model', 'formalism' or 'hypothesis'. In the face of this ambiguity reference is made to the conceptual framework put forward by Swets, and its extensions, as the Swetsian formalism. This is in sympathy with the notion that a formalism is a conceptual language in which statements about some area of phenomena of interest are described. Though a formalism must be self-consistent (or at least not obviously inconsistent) it is not in principle refutable by experiment. It does not predict. Thus a formalism is regarded

simply as a framework of definition and deduction and not, by virtue of its isolation from phenomena that prompted its formation, a device capable of prediction. Any assertions made within a formalism, but not deducible from it (not tautologies in it) will be referred to as hypotheses. If made within the Swetsian formalism, we will refer to them as Swetsian hypotheses, even if historically they were not made by Swets. The combination of Swetsian formalism and Swetsian hypotheses, i.e. the science of information retrieval put forward by Swets, will be referred to as Swetsian theory. The term 'model' in its usage as either theory or hypothesis will be avoided, but we will use the term in its narrower sense of approximating function, in later sections. In summary then, we have:



At the risk of repetition, we emphasise that a hypothesis, unlike a formalism in isolation, is vulnerable to experimental testing as well as requiring experimentation for its identification. Accordingly, an experiment that failed to support hypotheses (expressed within a formalism) would not be evidence of weakness in the formalism. It would simply be evidence that better hypotheses were

called for. Only a demonstration of logical inconsistency in the formalism, or a demonstration that all possible hypotheses expressable in it were unsupportable experimentally, or a lack of simplicity in the formalism, can lead to a formalism being rejected.

A further terminological point is that the description of the Swetsian formalism in Section 3 has been made as simple as possible. In particular, the term 'document' has been used as a convenient shorthand for 'reference to a document', and 'term' has been used freely when 'value of a document attribute' (or just 'document attribute', depending on the entity being regarded as a variable) would be more satisfactory in principle. (An exception to this practice is discussed in Section 3.3.2.6 when different types of attribute are considered.)

In the text single quotes ('...') have been freely used in an attempt to add clarity: emphasising that ambiguity or insufficiency in meaning attaches to a term or phrase, to introduce an important term, or to give an instance of a variable. Double quotes denote literal quotations. Context should make clear what is intended.

A common mathematical notation has been followed throughout the text, even where this entails recasting the notation of other workers. This has been made as simple and conventional as possible, e.g. by signifying all random variables by upper case letters (though not all upper case letters denote random variables), and instances of variables in lower case. Although 'F' has been used to denote 'Fallout', the starred version 'F_i^*' refers to one of several cumulative probability distribution functions. The probability function has been denoted by 'Pr' to avoid confusion with 'Precision', 'P'.

2. LITERATURE REVIEW

As mentioned in Section 1.2, this section provides an indicative review of the key theoretical and experimental papers which prompted, or have run parallel with, the present study.

The seminal paper which introduced the Swetsian formalism into information retrieval was published in Science (AAAS) in 1963 (Swets, 1963). In this paper, Swets prefaced his introduction of the formalism by a lengthy review of measures of retrieval effectiveness, and it seems clear from the structure of the paper, from his discussion, and from the full title of his paper: "Information retrieval systems: statistical decision theory may provide a measure of effectiveness better than measures proposed to data", that his prime objective was 'measuring retrieval effectiveness' rather than offering a complete theory of the retrieval process. That is, the formalism was apparently an accessory to this goal, not a goal in itself. Nonetheless, his approach to his prime objective did involve setting up a novel formalism of information retrieval (which he referred to as a "model") based in fact on one branch of statistical decision theory, namely signal-detection theory. Swets made 17 citations to earlier work, but none anticipate his own work in the specific matter of applying the signal-detection approach to information retrieval. No other papers have been identified by the author that do so. (There were occasional references to "signal" and "noise" in earlier literature on documentation (for example Maron and Kuhns had mentioned "semantic noise" in 1960, and Moss has reported that the Classification Research Group in England discussed "signal to noise" prior to this date (Moss, 1973),) but these terms were invariably used metaphorically. No structured, analytical

approach along the lines of signal-dection theory, and preceding the 1963 paper of Swets, has been found.)

Swets made reference in his 1963 paper to the possibility that: "An extensive testing program, originally designed for the study of signal detection in psychology, could be directly translated and applied to retrieval systems." (Swets, 1963: 250) This remark in fact anticipated his undertaking such a program, the results of which, together with a repetition of the basic formalism, were later published in American Documentation (Swets, 1969). These results were reported slightly earlier in a research report and a published symposium paper (Swets, 1967a and 1967b).

A paper by B.C. Brookes (1968), published in the Journal of Documentation at about the same time as Swets's American Documentation paper, both applied the formalism of the 1963 paper to a sample of data obtained from the Cranfield experiments, and suggested a modified version of one of the novel measures of retrieval effectiveness that that paper had proposed. Robertson, as part of an extensive review of measures of retrieval effectiveness, offered analytical commentary on Swets's work, and in particular stated a theorem that proved one of Swets's measures to be equivalent to a modified version of another of his (Swets's) measures - the modified version introduced by Brookes (Robertson, 1969).

Apart from the papers cited above, no analytical commentary on Swets's work existed up to 1973. This is not to say that the 1963 paper had not been widely read or cited: it was in fact frequently cited in the literature of the later 1960s, and occasionally the basic formalism was repeated. One early British review entitled "Information retrieval and the computer" published in 1964 as a research report was the first to do so, including in it an amended

version of one of Swets's figures (Barnes, 1964). The 1963 paper was also reprinted in several source books (Kochen, 1967a; Saracevic, 1970a). But, surprisingly perhaps, Swets's work did not attract the substantive analytical criticism from information scientists, librarians and data-base managers that might have been expected.

The above set of papers represents the area of published knowledge that was accessible to the author at the commencement of his study. This study itself led to several contributions which may possibly have served to draw further attention to Swets's work (Heine, 1973a, 1974, 1975, 1977a).

More recently, papers have appeared that concentrate on specific parts of the Swetsian formalism. An exception is the paper by Farradane (1974), a critical review again oriented to the problem of measuring retrieval effectiveness, and including both indicative and analytical comment on the Swetsian formalism. Papers by Bookstein (1974, 1977) have explored further the effect on retrieval effectiveness of ordering the basic events (the possible values of the weighting function) by likelihood ratios, thereby improving understanding of the role of such functions and of the attainable limits of retrieval effectiveness. A paper by Yu et al (1976) attempted to extend the formalism so that it could accommodate the notion of "relevance feedback", i.e. it introduced information transfer as a heuristic process into the formalism. Robertson (1977a) in a further major review paper on "Theories and models in information retrieval" offered criticism of the mathematics and structure of the Swetsian formalism, and on the compatibility of the formalism with the theoretical approaches of other workers. Very recently, a paper by Hutchinson (1978) has extended the formalism by introducing two

bivariate probability distributions in place of the univariate distributions originally suggested by Swets. (The idea was said by him to have been stimulated by discussion with B.C. Brookes and S.E. Robertson.) The extension is essentially based on the supposition that 'relevance' can be construed as a quantitative variable, i.e. that 'degrees of relevance' can usefully be recognised. As such, a fairly major amendment to both the formalism and the Swetsian hypotheses is involved which still awaits experimental investigation. The author's own 'bivariate generalization' of the Swetsian formalism does not anticipate Hutchinson's fundamental modification, since it is concerned with fixed sets of relevant documents and two (or more) weighting variables defined by observable document attributes of different type - not by variety in the marks used to denote documents as relevant or not. (Heine, 1977a)

As a further category of literature, the textbooks/monographs in the subject area have to date variously portrayed the Swetsian formalism. The works of Stamper (1973), Vickery (1975), and Paice (1977), surprisingly make no mention of Swets's work, although chapters by these authors are offered headed "Signal transmission", "Conceptual and mathematical models" and "The retrieval process" respectively. A monograph by Kochen (1974a) entitled "Principles of Information Retrieval" cites the 1963 paper of Swets, but neither summarises nor criticises Swets's theory. The works by King (1971) and Salton (1975a) give concise expositions of the formalism, the former without critical commentary but effectively linking some aspects of the formalism with that of others, the latter offering a brief criticism of it. The fullest treatment is in the monograph by van Rijsbergen (1979a) which gives both a summary of the theory and analytical commentary on it. Four criticisms of the theory are

offered by van Rijsbergen but it is perhaps fair to say that no distinction is made between Swets's formalism and hypotheses that may be expressed in that formalism, and that the emphasis is on the formalism as a tool in evaluation rather than as a device for portraying the roles of relevance decisions, questions and weighting functions in the retrieval process seen as a whole. Placing Swetsian theory in the context of 'evaluation' (or 'testing') is in fact a general feature of all the above works and in the author's view an unnecessary one in view of the capacity of the Swetsian formalism to describe what information retrieval 'is' in fundamental terms.

A last category is that of the annual review literature. The theory attracted various indicative or expository commentary in the early volumes of "Annual Review of Information Science and Technology", but has not been discussed to date in "Advances in Librarianship" or "Progress in Library Science". A review paper by van Rijsbergen in "Progress in Communication Science" (1979b) also treats the Swets theory briefly, again within the context of retrieval effectiveness.

Looking back at the literature on Swetsian theory at the time of writing, several features stand out. First, there has been an increasing although still rather tentative interest in the theory, an interest perhaps reflecting a growing concern that an adequate science of information transfer appropriate to document-handling in general and data-base design in particular is still lacking. The increasing reliance of document users on information retrieval when implemented through computer-based systems (rather than on shelf browsing in local library collections) may be one cause of this. Secondly, there seems to be an increasing appreciation that the

formalism offered by Swets should not simply be seen as an apparatus for defining novel measures of retrieval effectiveness, but as a hospitable, concise theoretical framework for describing and understanding the retrieval process in its entirety with the possible exception of the relevance-judgement process. This feature is perhaps implicit in the growing number of papers on diverse aspects of the formalism, and perhaps because of this and the general synoptic power of the theory it has been referred to as "The most highly developed of the theories on information retrieval " (Robertson, 1977a: 131). Thirdly, despite the simplicity, hospitality and falsifiability of the theory, there have been only two published accounts (Brookes, 1968; Swets, 1969) of experimental attempts at testing hypotheses expressed in the Swetsian formalism. Both attempts were however, in the author's view, based on an insufficiently rigorous methodology, so that apart from the work later described in this thesis, one can say that the Swetsian hypotheses have simply not been tested experimentally. (To say this is to give only provisional admission to hypotheses put forward by Swets: we shall later argue that Swets did not put forward hypotheses that were unambiguous.)

The points made above will be justified in the following text, which seeks not just to criticise the Swetsian formalism but to extend it so as to remedy the inadequacies and ambiguities it originally had.

3. SWETS'S THEORY OF INFORMATION RETRIEVAL

The formalism describing information retrieval and proposed by Swets had already been widely accepted in psychological research. This is evident from Swets's previous and later writings in the latter area (for a bibliography of same, see Green and Swets (1974)), and also from citations made by Swets in his two key papers on information retrieval (Swets, 1963, 1969). For example, Swets's 1963 paper cites one of the classical papers on signal detection theory in psychophysics by Tanner and Swets (Tanner, 1954), and his 1969 paper cites two of the main reference works in the area (Green and Swets, 1974 [1966 edition]; Swets, 1964).

Since the signal detection formalism is the basis of Swets's theory, and since his two papers on information retrieval give a relatively brief account of it, the following section has been included to fill this gap. An understanding of it is both necessary in order to be able to see where the information retrieval formalism departs from the basic psychophysical formalism, and desirable in pointing to assumptions that could be re-examined in any future refinements of the information retrieval formalism. Some of these refinements will be developed in later sections, or have already been made or touched on in published information retrieval literature.

3.1 The Historical Origins of the Signal-Detection Formalism

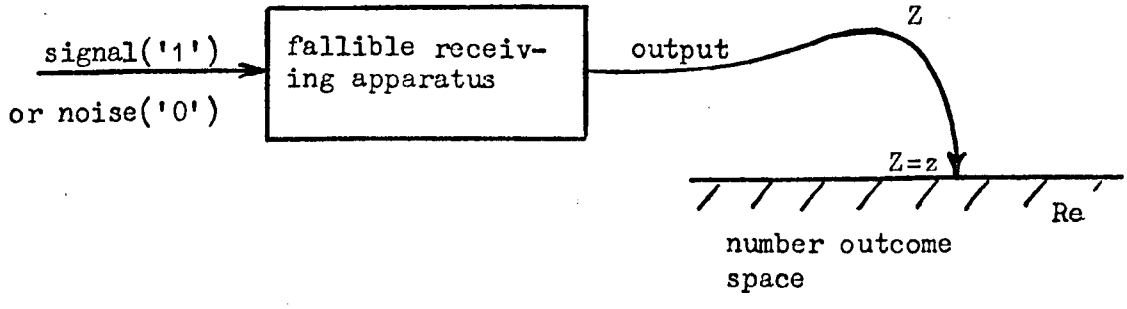
The signal-detection formalism has origins in electronic engineering (especially in regard to the receipt of electromagnetic signals in the presence of noise) and statistical theory (in hypothesis testing), in the 1940s and 1930s respectively, as well as in psychophysics in more recent years. These origins are discussed more fully in a review paper by Swets (1973) which gives as the key source papers for hypothesis testing the works by Neyman and Pearson (Neyman, 1933) and Wald (1950). For a list of source references in the electronics area, the reader is referred to Green and Swets (1974:1). The review paper by Swets just referred to also relates signal detection theory to previous theoretical work (not couched specifically in signal detection terms) in psychophysics. The key papers on the formalism as originally applied in psychophysics appear to be by Tanner and Swets (Tanner, 1954), Smith and Wilson (Smith, 1953), and Munson and Karlin (Munson, 1954). The literature in this area, now extensive, has been thoroughly reviewed in the monograph by Green and Swets and, in one specific aspect, by Egan (1975). Key papers are reprinted in Swets (1964). Perhaps indicative of the widespread acceptance of the theory, and of its apparent success in describing observational data, is the fact that about 50 papers per year are currently being published in the area as a whole, divisible into about 12 more-specific areas of application such as memory, vigilance, the diagnostic process, and recognition.

The basic situation that statistical decision theory, and more particularly signal detection theory describes is that where an observer receives, from some fallible device, stimuli or observations which relate to one of two possible events. The 'observer' is human or animal of course in psychological research,

but can be construed more abstractly as an observing and decision making process. In view of the objective of our discussion, we adhere to the latter construction even though this represents a slight change to a portrayal of the psychophysics literature. The two input events may be labelled 'signal' and 'noise', the underlying intuitive notion being that events labelled 'signal' are somehow of greater importance to the observation process - even if the importance is introduced arbitrarily in an experiment. These events are, however, labelled rather confusingly by many writers as "signal+noise" and "noise" respectively. This is apparently in recognition of two facts. One of these is that in most psychological experiments in which the theory is introduced the signal chosen is deliberately corrupted or complicated in some manner, and to a variable extent. For example, an audible signal transmitted to a human subject in an experiment will usually not simply consist of a waveform of one frequency and amplitude, but will have added to it other waveforms of varying frequency, amplitude and phase. Since this complicated signal is to be compared by the human observer with instances of noise (other transmitted information so labelled), trivialization of the experiment is prevented. The other fact is that there are, inevitably, small random variations within the instruments generating the signal which add a stochastic character to it, variations which might more properly be referred to as 'noise'.

The formalism of 'observation' is then advanced as follows. A 'receiving apparatus' extracts information from the events to which it has (through its design) access. These events are said to be 'transmitted' to it. It does so with two characteristics: (1) the extraction of information is transient or Markovian (the device is assumed, ideally at least, not to be an integrating device: receipt

of an event causes it to 'forget' its response to the previous event); and (2) it is 'imperfect' or 'fallible' in its extraction of the information. By the latter is meant that an event (transiently) recorded by it cannot be predicted with certainty from a knowledge of the event input to it. We also suppose that the output of the receiving apparatus, for input of binary character (i.e. a stream of events each of which can be labelled '1' or '0'), is a real number. It is emphasised that the mapping of a sample event to a real number is both non-deterministic and Markovian. The diagram below captures this simple idea. Z denotes the function mapping input events to values in the real line, denoted Re, and z is a sample value of Z.



The receiving apparatus forms only part of the observation process. The essential remaining parts of the process are as follows:

(1) a data structure, in which information is stored on (a) the estimated relative frequency of occurrence of the input events $\{1\}$ and $\{0\}$; (b) estimated values of the likelihood ratio, $\ell(z)$, of the individual events $Z = z$; and (c) following Coombs et al (1970) the maximum-likelihood "utilities" of four events (at present undefined) denoted by U_{mn} $(m, n = 0,1)$.

(2) a signal-detection algorithm which performs, for each event input to the receiving apparatus, the evaluation of the logical expression $\ell(x) \lesseqgtr \beta$ using the information in the data structure. The parameter (or 'threshold value') β is also calculated from the data structure. When the expression evaluates to 'true', the algorithm asserts that the input event conforms to one hypothesis (H_o say), otherwise that it conforms to the alternative hypothesis (H_1). (The utility values U_{mn} referred to above are attached to the four outcomes: H_m true and asserted to be true (false), $n = 0$ ($n = 1$), $m = 0,1$.)

The concept of 'observing process' may also admit, as either a complication to the description just given, or as an alternative description, the notion of variation in the character of the input events beyond the simple '1' and '0' classes. Instead of an input signal '1', we might consider inputs labelled '1.1', '1.23', '0.99', etc., these appearing in place of '1' in an unpredictable manner. The latter 'stochastic randomness' can then be regarded in one of two ways, determining two characterisations of the observation process. On the one hand we could assert that stochastic randomness in signal is not 'knowable' to an observational process, by definition of the latter. The process cannot then distinguish between non-binary variations in signal, since (by definition) only the response of its receiver is accessible to it. If an assumption as to the existence of stochastic randomness in the input signal is built into the observing process, the effect can simply be seen by the process as one determinant of the random function Z, i.e. it adds to the random behaviour of the receiver that exists in any case for binary input. On the other hand, the observation process

not necessarily equivalent, in the sense of 'implying the same assertion for each input event', as will later be discussed.* Thirdly, the likelihood ratio $l(z)$ can be expressed as the ratio of two probability densities, $f(z \mid s)$ and $f(z \mid n)$:

$$l(z) = f(z \mid s)/f(z \mid n) \qquad \text{(defn)}$$

where s denotes 'signal received' (e.g. the character '1' in a non-stochastic input stream), and n denotes 'noise received' (e.g. the character '0'), when continuous probability density functions are used to model receiver output behaviour. The observation process then has, in its data structure, either an array $l(z)$ (one real-number value for each recognised output value (sampled from Re) of the receiver, z); or a set of 3-tuples (i.e. three arrays): $(z, f(z \mid s), f(z \mid n))$. The requirement of economy of storage in the data structure suggests the former structure be used. A fourth point is that the observation process also has stored, as we indicated earlier, an estimate of the value of $Pr(s)/Pr(n)$ - the so-called prior odds of signal to noise - and four values of the variables U_{mn} $(m,n=0,1)$. These five values allow the process to fix the parameter β in the following way. (The approach follows Green et al (1974: 20-5) and Coombs et al (1970: 168-71), the latter also being recently cited by Bookstein (1977).) Our problem is: given a set of values of the utility variables:

U_{00} = utility of event 'signal received and 'signal' asserted' (> 0)

U_{01} = utility of event 'signal received and 'noise' asserted' (< 0)

U_{10} = utility of event 'noise received and 'signal' asserted' (< 0)

U_{11} = utility of event 'noise received and 'noise' asserted' (> 0),

and of $Pr(s)/Pr(n)$, what is the optimum value for β? The criterion

---

* This point is taken up in recent information retrieval literature by Bookstein (1974, 1977), though perhaps with inadequate emphasis given to earlier work in signal detection theory (e.g. Helstrom, 1960).

for optimality may be chosen to be 'the value of β that maximises the expected total utility'. Denote expectations by E(...), and the total utility for the assertion '...' by U('...'). Then for a set of input events labelled s or n we have*:

$$E(U('signal')\,|\,z) = U_{00}Pr(s\,|\,z) + U_{10}Pr(n\,|\,z)$$

and $$E(U('noise')\,|\,z) = U_{01}Pr(s\,|\,z) + U_{11}Pr(n\,|\,z).$$

- by definition of E(...). We require a decision process that is such that if $E(U('signal')\,|\,z) \geqslant E(U('noise')\,|\,z)$ then 'signal' is asserted, else 'noise' is asserted. This inequality is equivalent to:

$$\frac{Pr(s\,|\,z)}{Pr(n\,|\,z)} \geqslant \frac{U_{11}-U_{10}}{U_{00}-U_{01}} \qquad (1)$$

But from the definition of conditional probability,

$$Pr(s\,|\,z) = \frac{Pr(z\,|\,s)\,Pr(s)}{Pr(z)}, \text{ and } Pr(n\,|\,z) = \frac{Pr(z\,|\,n)\,Pr(n)}{Pr(z)}$$

so that '(1)' can be rewritten as:

$$\frac{Pr(z\,|\,s)}{Pr(z\,|\,n)} \geqslant \frac{U_{11}-U_{10}}{U_{00}-U_{01}}\,\frac{Pr(n)}{Pr(s)}\,. \qquad (2)$$

The right hand side is a constant, now identifiable with β, and the left hand side is by definition the likelihood ratio for the event z, l(z). Accordingly our decision rule is equivalent to: assert 'signal' if the inequality '(2)' is true, else assert 'noise'.

The above likelihood ratio rule is workable since it is based on 'knowledge' or 'information' that we have prescribed to be part

---

* Strictly, since probability functions are associated with <u>sets</u>, we should write '{z}' for 'z' in what follows, '{z}' denoting a Borel set - a subset of the Borel field, which contains all possible subsets of the real line. This is avoided here for simplicity in presenting the 'utility argument'. Induced density functions on the other hand are functions of z, not {z}.

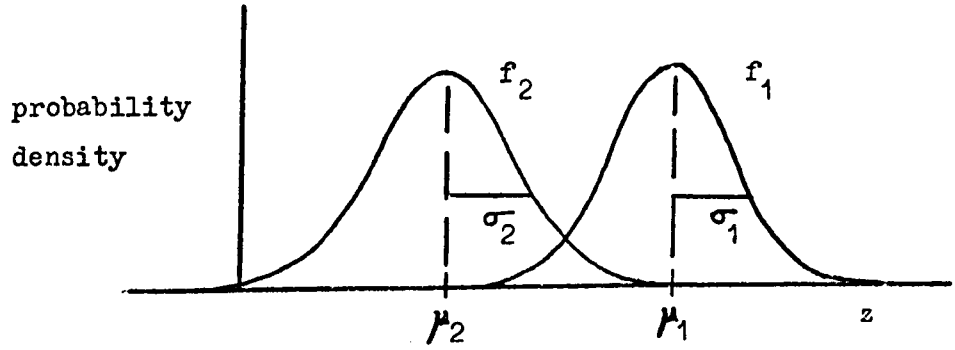
of the observation process. We also add that in the psychophysics area, when a decision is based on the above inequality, the observation process is referred to as "an ideal observer". Various alternative decision rules can also be defined, all of which involve a test of the form $l(z) \gtrless \beta$ (Green and Swets, 1974: 20).

So far the discussion has been in decision-theory terms - except insofar as the events of interest, the z values, have been ordered. Signal detection theory is a more specific form of decision theory in which probability distributions determined by one or other analytical expression are assumed to describe (i.e. model) the distributions of $Pr(z \mid s)$ and $Pr(z \mid n)$. One very common assumption in applying the theory is that the two latter distributions are describable as normal density functions, i.e.

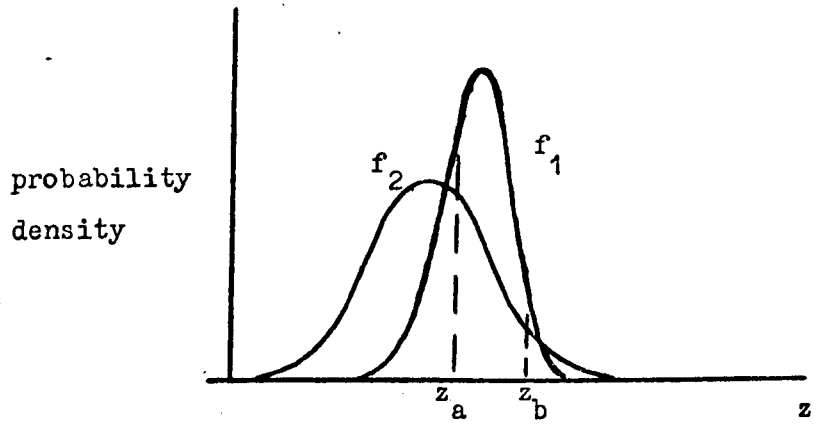
$$Pr(\{ z \in (z, z+dz) \} \mid s) = f_1(z) = (2\pi\sigma_1^2)^{-\frac{1}{2}} \exp(-(z-\mu_1)^2 / 2\sigma_1^2)dz$$

$$Pr(\{ z \in (z, z+dz) \} \mid n) = f_2(z) = (2\pi\sigma_2^2)^{-\frac{1}{2}} \exp(-(z-\mu_2)^2 / 2\sigma_2^2)dz$$

- the subscripts '1' and '2' referring to the signal and noise distributions respectively. (The subscripts 's' and 'n' would be more suggestive at this stage, but the ones given are consistent with later notation used in information retrieval.) The following diagram illustrates these functions:



The figure assumes (1) that z is continuous (as the concept of a Normal distribution does also); (2) that the receiver records higher values for input signals than for input noise ($\mu_1 > \mu_2$); and (3) the variances of the two functions are the same ($\sigma_1^2 = \sigma_2^2$). The latter point leads to an important further consideration. To assume that the (ideal) observation process decides to output 'signal' if and only if $l(z) > \beta$, is not necessarily to assume that a value z_c exists such that $l(z_c) = \beta$ and the observation process can output 'signal' when $z > z_c$. In other words a decision based on likelihood ratio values is not necessarily equivalent to a decision based on values of the receiver. (By 'equivalent' here we mean 'partitions the outcome space in the same way'.) That this is true is intuitively evident from the following diagram:



At two points, z_a and z_b, f_1 and f_2 are such that f_1/f_2 = constant ($= \beta$ say). By the likelihood-ratio decision rule, the observation process should identify as 'signal' all those events that map to the interval $z \in [z_a, z_b]$; whereas by the receiver-value decision rule the process should identify as 'signal' either (1) all events mapping to the interval (z_a, ∞), choosing the lower value of z_c or (2) all events mapping to the interval (z_b, ∞), choosing the higher z_c-value. Two conclusions follow. First that if the receiver really

does behave like this, then the likelihood ratio rule is preferable. Secondly that there would appear to be no advantage to an observation process in having a receiver that responded in the receiver-value rule way: discrimination between input events of the two types can be weakened in the region of higher z-values. There will be no difference between the two algorithms if and only if $l(z)$ increases monotonically with z. It follows that if the two decision rules are equivalent, the process must constrain f_1 and f_2 so that that is true. A summary of the constraints on f_1 and f_2, for various common analytical forms (Normal distribution, binomial distribution, gamma distribution) is given by Egan (1975, Appendix E, Section E.3). In passing the author offers the view that there is no adequate and coherent discussion in the signal detection literature as applied to psychophysics, on the optimality of observation processes for which equivalence between these decision rules obtains. This is so notwithstanding an immense amount of discussion on the monotonicity of $l(z)$ and z in solely mathematical terms. The author has not found constructive comment on this point, but the matter is in any case not central to this thesis. Possibly in modelling human or animal behaviour in general (i.e. other than in respect of relevance-judging) it is a reasonable hypothesis that information on z, $Pr(z \mid s)$ and $Pr(z \mid n)$ is not stored separately, but instead stored more economically (with two-thirds the storage space) as z and $l(z)$. In that event the decision would <u>necessarily</u> be based on an '$l(z) > \beta$' test, rather than a '$z > z_c : z_c = \ell^{-1}(\beta)$' test. It is also plausible that such $l(z)$ values are stored sequentially in increasing order, on the grounds of economy, which would reduce the storage requirements to the values of $l(z)$ themselves, i.e. reduce storage by a half again. If

this were so it follows that the (unstored) z values associated (mathematically) with the $l(z)$ values are again monotonically related to $l(z)$ as a matter of necessity. It is then immaterial whether we view the (ideal) observation process as one making decisions on the basis of one rule rather than the other since they are equivalent. However, if the process stores both $l(z)$ values and z values, and these are not monotonically related (in which case the process is, by definition, 'non-ideal'), then decision is better when based on the $l(z)$ values rather than the z values. (The process's storage and decision mechanisms could not be bettered by storing the z-values as well, and deciding on the basis of a '$z > z_c$: $z_c = l^{-1}(\beta)$' test.)

Before proceeding we summarise the signal detection theory as it has been presented. Binary information is transmitted to an observation process or observer, which (or who) receives such information via an imperfect receiving device, the output of which is a numeric value. The process is required to generate a binary output for each input event. From information as to (1) the estimated relative frequencies of occurrence of the two characters in the input stream, (2) the estimated utilities of decisions made, and (3) the estimated response characteristics of the recording device for each type of input event, the process determines a further numeric value, β, and then makes a decision as to the nature of the input on one of two grounds. One of these is: 'if $l(z) > \beta$ then choose one hypothesis as to the character of the input event, else choose the alternative hypothesis'; the other is 'if $z > z_c$ (where $z_c = l^{-1}(\beta)$), then choose (etc.)'

There are still further fundamental points that need clarification, in the author's view. First, although we have presented the

observation process and in particular its receiving apparatus as non-observables, this is simply a reflection of the way in which they have been (and apparently must be) treated in psychophysical experiments. (The mapping function of the ear, for example is not directly observable, let alone stored utility values for each possible auditory signal.) In other experiments however, these could be observables. The observation process could be an algorithm, possibly under human control, for example. (See Coombs et al (1970: 167) for an instance of psychological evidence, presented in terms of signal detection theory, but without a human observer.) Secondly, despite widespread misunderstanding or at least arbitrary assumption on the point, signal detection theory does not require one to assume that the distributions of $Pr(z \mid n)$ and $Pr(z \mid s)$ are Normal. The literature frequently describes these distributions as Normal, partly to avoid unduly abstract discussion, and partly in acceptance of the hypothesis that the random action of the receiver will be dictated by the random sum of many unknown, equi-distributed, and independent random variables. (In such cases the "Normalising" probabilistic effect described by the Central Limit Theorem would take effect.) But it must be stressed that many other analytical forms of distribution have been considered in the literature, as described in the standard reference sources already cited. (Interestingly, however, no literature has been identified in which the analytical forms used to describe f_1 and f_2 are different for a given observational process.)*

* That this was true in the case of Swets's application of the formalism to information retrieval is one of the criticisms of the application made in van Rijsbergen's monograph (1979a:158; and first edition, 1975: 109).

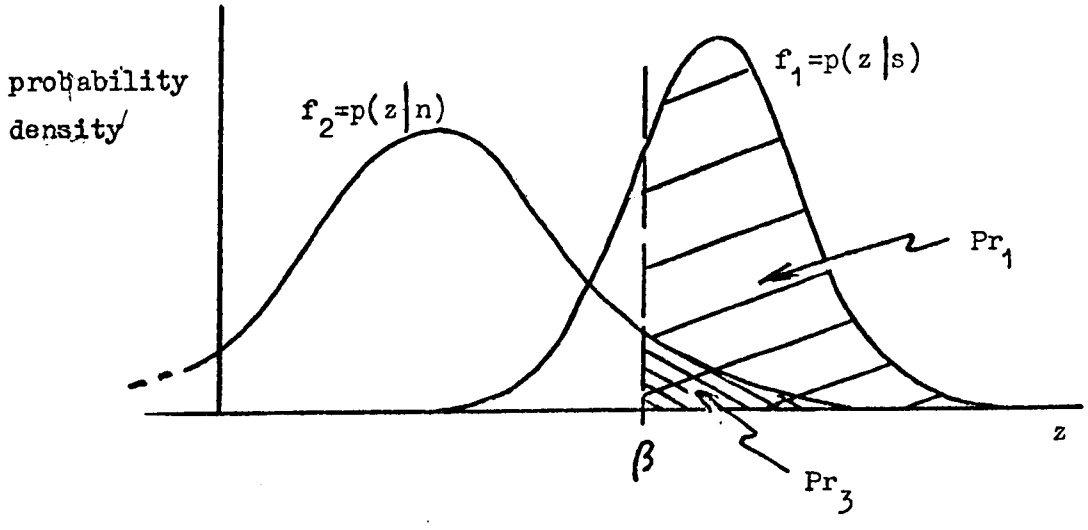
Thirdly, although we presented the view that the decision was made on Bayesian grounds, using the a priori information we have described, this notion was introduced in order to avoid the notion that β was chosen arbitrarily by the observation process. One alternative approach is to assume a 'guessing observer' who employs a value for β that is randomly distributed (in some way) about its assumed optimum value. The point being made here is that signal detection theory does not need to assume that β is computed in any fixed way, only that decisions are made on the basis of a (relatively) fixed β value while external (e.g. experimental) conditions are fixed. This is simply a consequence of the definition of observational process: it does not preclude our defining more general entities.

We have not so far said what evidence there is to support hypotheses expressed in the signal detection formalism in the case where the receiver response values z and the likelihood values l(z) are unobservable - as is the case with experiments on human subjects, say. The evidence used is in fact the manner of variation of the paired data: (probability of response 'signal' given that the input event is a signal; probability of response 'noise' given that the input event is noise), or in fact any other set of paired probabilities relating to the four experimental events. The variation is associated with a set of fixed experimental conditions, and a specified observation process, with each condition allowing for a sufficiently large number of input events that values for these probabilities can be measured with relatively small estimated error. (The variation of experimental conditions can itself take various forms, e.g. that of informing a human observer what the value of Pr(s)/Pr(n) will be, and varying this value, or altering the

'motivation' of the observer. Then if hypotheses expressed in terms of a choice of f_1 and f_2 are true, the probabilities chosen (e.g. $Pr(s \mid$ 'signal' asserted), or $Pr(n \mid$ 'noise' asserted)) will vary in a way related to f_1 and f_2. The probabilities of interest are predicted by signal detection theory, under the 'l(z) increases monotonically with z condition' as:

experimental event for fixed experimental conditions	probability
signal input and 'signal' asserted	$Pr_1 = \int_{\beta}^{\infty} f_1(z)dz$
signal input and 'noise' asserted	$Pr_2 = \int_{-\infty}^{\beta} f_1(z)dz = 1 - Pr_1$
noise input and 'signal' asserted	$Pr_3 = \int_{\beta}^{\infty} f_2(z)dz$
noise input and 'noise' asserted	$Pr_4 = \int_{-\infty}^{\beta} f_2(z)dz = 1 - Pr_3$

Pr_2 and Pr_3 are identifiable with, respectively, the probabilities of Type II errors, and Type I errors, in statistical hypothesis testing. Pr_1 is also identifiable with the 'power' of the statistical test used. The following, conventional diagram illustrates the nature of the Pr_i as areas under a probability density curve (for z assumed to be continuous).



might be defined so that the data structure recognized such variation in the input signal. In this case the process asserts that the input is distributed over some set of intervals $\{\alpha_i, i \in I\}$, and the purpose of its signal-identification algorithm is to assert the truth of one of a set of hypotheses $\{H_i, i \in I\}$, perhaps using a more complicated store of threshold values B_i and utility values $\{U^i_{mn}, i \in I\}$.

We have described the influence of the random behaviour of the receiver on the observation process as a whole in the above terms, since it removes several common sources of confusion in respect of the theory, and since the formalism to be applied to information retrieval is of a relevant character. We add, looking ahead, that the first interpretation given above relating to a binary input was the one considered by Swets in applying the formalism to information retrieval.

Before proceeding, various points require further comment. In describing the outcome space of the receiving device to be the real numbers, we implied that the output was a continuous variable. This is not essential: the outcome space can be discrete. (See Egan (1975) for example, who devotes two chapters to discrete outcome spaces - the positive integers in fact.) Secondly, we suggested that the 'deciding part' of the algorithm was based on the truth of an inequality: $l(z) \leqslant \beta$. In fact although the overwhelmingly larger part of signal detection theory is concerned with criterion-inequalities of this form, it may be the case that in some observation processes the simpler criterion: $z > z_c$, $(z_o = l^{-1}(\beta))$, is tested and its logical value used as the basis of the assertion as to which hypothesis is correct. The two criteria are

Under the assumptions that f_1 and f_2 (1) remain fixed during an experiment, and (2) have prescribed analytical forms, it is further assumed that the effect of varying conditions during an experiment is such as to produce change only in the value of β selected by the observation process. This is taken as evident (i.e. observable outside the process) in the way that variation in the four probabilities Pr_i is constrained. Thus experimental evidence will tend to confirm the assumption (for a particular choice of forms for f_1 and f_2) if variation in experimental conditions causes the experimental values of Pr_i and Pr_j (any i,j, $(i \neq j)$) to vary in a way similar to the variation in the values of Pr_i and Pr_j that are predicted by f_1, f_2 and a varying value of β. The question then is just how much variation in the discrepancies between experimental and predicted values of such pairs of probabilities should be permitted before the assumptions are regarded as failing. Discrepancies will exist in practice due to (1) uncontrolled randomness in the experimental arrangement, (2) guessing and 'non-ideal' behaviour in the observation process,and (3) 'learning' by the observation process (i.e. changes in the stored values of l(z), etc.) and (4) sampling errors in the measurement of Pr_i and Pr_j due to the experiment's sampling of z values; as well as due to insufficiencies or inaccuracies in the assumptions made, as expressed in the language of the signal detection formalism.

The statistical problem of testing hypotheses expressed within the signal detection formalism does not appear to have been treated in any depth by psychologists or statisticians, the treatise by Green and Swets devoting only 11 pages (of an appendix) to the topic. (Green, 1974: Appendix III, Section III.3 "Data analysis"). The usual approach to demonstrating the validity of such hypotheses

is instead semi-intuitive, with the probabilities Pr_1 and Pr_3 plotted onto what is called a "ROC graph", ROC standing historically for "receiver operating characteristic". (Swets (1973) has recently suggested "relative operating characteristic" as being more in keeping with the related mathematics of hypothesis testing.) We note also the term "proper ROC graph" used by Egan to denote ROC graphs arising from a varying $l(z)$ criterion, rather than from a varying z criterion. The ROC graph is usually scaled so that the probability values recorded on it are mapped to the equivalent standard score values using the inverse of the complement of the standard Normal probability integral. That is, an experimental probability value, Pr_i, is mapped to a value of z_c, say z_c^i, by means of:

$$Pr_i \longmapsto z_c^i = \Phi_c^{-1}(Pr_i), \text{ where } \Phi_c(z_c^i) = \int_{z_c^i}^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-u^2/2)du.$$

Fig. 3.1-1 and its caption elaborate on this notion. For scaling of this kind, the property that f_1 and f_2 are both Normal densities implies that paired (Pr_1, Pr_3) values will lie on a straight line when β is varied. When the slope of the line is unity, the functions f_1 and f_2 will (if both are Normal) have the same variance, not necessarily unity. Unfortunately however the straight line and unit slope properties for experimental data are not easily verified (see Green and Swets, 1974: 401) since even for samples of 600 events, the variations in the values of Pr_i due simply to sampling error are large. (For N=600, and Pr_1=0.1 say, the standard error in Pr_1 is approximately 0.012; i.e. approximately 68% of measured Pr_1 values obtained with samples of this size will lie within the interval 0.1 ± 0.012. Transforming to equivalent values of z_c, expressed as a standard Normal score, gives an equivalent interval

for z_c of [1.185,1.353].) The difficulties of testing a choice of analytical forms for f_1 and f_2 are accordingly formidable, not just because of high estimated standard errors in the parameters of f_1 at f_2, but because different analytical forms tend to produce very similar ROC graphs (see especially Green and Swets (1974:401 and chaps. 3, 5). This is due to the Pr_i being cumulative probabilities, in contrast to the more basic distributions of $Pr(z \mid s)$ and $Pr(z \mid n)$. A further variable element points to a weakness in the ROC-graph approach. This is that although Normality of both f_1 and f_2 implies a straight line ROC graph (when Pr_i values are transformed to standard normal scores and plotted in those scores), the converse is not true. For given any ROC graph, and any density function f_1, a density function f_2 can be found yielding that graph. The additional supposition that f_1 and f_2 have the same analytical form would thus be a necessary additional hypothesis in any claim that experimental evidence supports one of the distributions being Normal, when this is argued solely from the ROC graph, i.e. when the f_1 and f_2 distributions are not directly observable.

Two further and final concepts referred to in the psychophysics literature are now briefly described. First, we note that it is possible to define a distance between the distributions f_1 and f_2. One such distance is d' defined by:

$$d' \equiv \frac{\mu_1 - \mu_2}{\sigma} \qquad \text{(defn)}$$

for the case where f_1 and f_2 have a common variance σ^2. See for example Egan (1975). When the variances differ, d' is undefined. This distance can be identified with geometric distances appearing in Figure 3.1-1, namely GG_1, GG_2 and GG_3 (Egan: 68), and provides one measure of the capacity of an observation process to discriminate between signal and noise. The second concept is that of 'psycho-

metric function'. This is defined (e.g. Egan, 1975: 48; Green and Swets, 1974: 187), as a functional relationship between (1) sets of paired values of Pr_1 and Pr_2, i.e. $\{(Pr_1.Pr_2)\}_j$ defined by the process's chosen values for β, and for constant prior odds $Pr(s)/Pr(n)$, and (2) the latter values themselves. Intuitively, the psychometric function reflects the change in the discriminating power of the observing process (as evident in f_1 and f_2) for various levels of signal to noise.

Before leaving signal detection theory in its psychophysical context, the context which apparently prompted Swets's work on the description of the information retrieval process, we summarise its main features:

(1) The signal detection formalism, distinguishes, through its structure, between (a) the overall discriminating power of an observation process (expressed through the location of the ROC graph, or as a value of some measure of separation of f_1 and f_2), and (b) the specific discriminatory power of a process when a variable criterion used in the process and determining its response, is assigned a value. The distinction just given might be said to constitute the main 'explaining power' of the formalism.

(2) The formalism is flexible in allowing for an arbitrary choice of distributions f_1 and f_2. Statements about f_1 and f_2 are hypotheses expressed in the formalism.

(3) A fundamental feature of the formalism is its 'macroscopic character': it actually depends, for its definition, on unexplained random behaviour. It does not describe a deterministic situation. It is thus necessarily a probabilistic formalism, like those pertaining to quantum mechanics or gas mechanics say, and unlike, say, those of Newtonian mechanics or electromagnetic field theory.

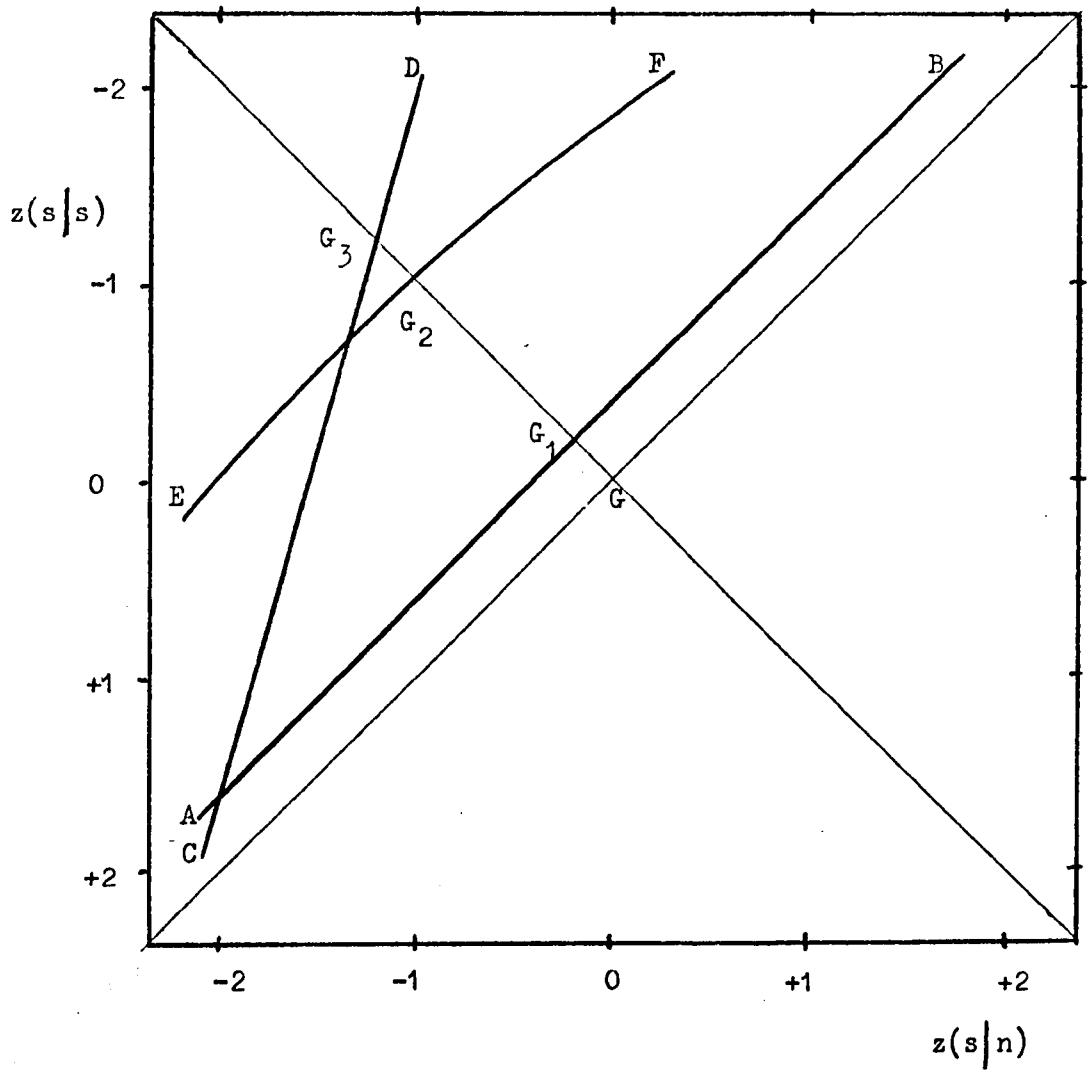


Fig. 3.1-1. ROC graphs in standard form. The ordinates $z(s|s)$ and $z(s|n)$ are standard scores corresponding to empirically-measured probabilities $Pr(s|n)$ and $Pr(s|s)$, on the assumption that the probability distributions underlying these are distributed $N(0,1)$. Three idealised observation processes, O_1, O_2, O_3 have each been the subject of a series of experiments, each experiment yielding data in the form a pair of values of Pr_1 and Pr_3 lying on the curves shown. The process O_1 yielded data lying on the straight line AB, indicating that the process is consistent with a signal detection hypothesis that f_1 and f_2 are both Normal and of the same variance. Observation process O_2 has yielded data lying on the straight line CD, indicating consistency with the hypothesis that f_1 and f_2 are Normal but of unequal variance. Process O_3 yielded data lying on the curved line EF, indicating the f_1 and f_2 cannot both be Normal.

(4) The formalism portrays observation as essentially a number-assignment process, rather than as a truth value-assignment process, i.e. as a logical process in the technical sense. Historically, this has probably been due to a need to describe phenomena that are naturally complex, i.e. where a portrayal of observation in terms of set-operations would be unfeasible, although the author has found no work which explicitly discusses this alternative approach and rejects it.

(5) Lastly, the emphasis in attempting to contest hypotheses expressed within the formalism has been on data portrayed as a 'ROC graph', i.e. as essentially a plot of Pr_1 against Pr_3. With a degree of hindsight from the author's own work, and from the emphasis in current work in information retrieval, it is perhaps surprising that data more oriented to the characterisation of the signal has not been used. In particular a graph of Pr_1 against another (conditional) probability: that for the event 'signal input given signal asserted', would provide a more natural focus of interest. For 'rare' signals, this would appear to provide a much more sensitive characterisation of the influence of observation upon signal transmission.

3.2 The Formalism used by Swets to describe Information Retrieval: Description and Interpretation.

The heading of this section is perhaps surprising, in that it might seem that a more satisfactory approach to Swets's work would be to separate a description of it from an interpretation of it. In the author's view however, this is in principle impossible owing to various major ambiguities in Swets's presentations.

As mentioned in Section 1.1, the theory described in Section 3.1 was first applied to the information retrieval process by J.A. Swets (1963, 1969). Swets was the first worker (1) to treat information retrieval from the point of view of signal detection theory, (2) to develop such an approach, and (3) to attempt to relate experimental results to the theory.

The signal detection formalism was introduced by Swets in his 1963 paper under the broader heading "statistical decision theory". The formalism appeared in the context of a review of methods of assessing retrieval effectiveness. It may seem merely a historical point that a formalism should have been introduced in the context of an "evaluation" paper, but this led to two weaknesses in his presentation which will become apparent: (1) the formalism was not introduced in a careful rigorous way, and (2) it was seen essentially as a basis for the assessment of system performance rather than as a pervasive theoretical structure providing unity to the whole retrieval process and serving, or having the potential to serve, as a means of optimising that process. This context (i.e. the evaluation context) is apparent in both the mini-abstract of the 1963 paper("statistical decision theory may provide a measure of effectiveness better than measures provided to date"), and by the

title of Swets's later American Documentation paper: "Effectiveness of information retrieval methods". Because of this, it has been unfortunate that Swets's work has usually been discussed in the context of evaluation rather than in the context of the formal characterisation of the information retrieval process qua process.

The 1963 paper began with a review of published quantitative measures of retrieval effectiveness. Swets then introduced the heading 'Proposal'. He briefly discussed statistical decision theory, citing Maron and Kuhns (1960) and Wordsworth and Booth (1959), and introduced signal detection theory as an "analogous" form of it. He wrote:

> "The measures taken of the input to a detector...must be assigned to one of two events - the detector system reports either that noise (random interference) alone existed or that a specified signal existed in addition to the noise. Similarly, a retrieval system takes a measure of a given item in the store, relative to a particular query, in order to assign the item to one of two categories - the retrieval system rejects the item as not pertinent or retrieves it." (p.247)

The basic signal detection formalism is thus introduced, not in the context of human assessments of the members of a stream (or set) of documents (which might have been expected perhaps from Swets's earlier work on human subjects as observers), but in the context of the more abstract process: a system identifying documents as relevant or non-relevant to a need, in response to a query. The term "pertinent" in the above quotation is clearly being used as a synonym for the present-day term "relevant", notwithstanding dis-

tinctions in meaning between these two terms introduced by other writers (e.g. Rees and Saracevic, 1963). An element of confusion is perhaps introduced where Swets wrote "...or that a specified signal existed in addition to the noise." It will become apparent that what was meant is that only input events of the form "noise" or "signal" are involved in the application of the formalism, not "noise" and "signal plus noise" as is usual, perhaps inevitable, in psychophysical experiments.

Having introduced the concept of a retrieval system as one describable in signal detection terms, i.e. as an observation process, Swets made the following disclaimer:

> "The primary aim here is not with a process, or with system design, but with the measurement techniques that accompany the process description. The process modelled is presented here, though very briefly, because it provides for the measurement technique. The model is described in the language of the retrieval problem to display one possible coordination between the elements of the model and the physical realities of retrieval. It is suggested, however, that the measurement techniques may be used to advantage whether or not this particular coordination seems entirely apt." (p.247)

It seems apparent, therefore, that the formalism was seen by Swets as provisional and not necessarily a representation of "physical reality" - to use his phrase. The formalism, he was suggesting, was provided simply as a background sketch to explain the meaning of certain quantitative measures that were to be described. With hindsight, this disclaimer seems surprising and unnecessary. Unlike the

human signal detection situation, the signal detection formalism as applied to information retrieval deals entirely with observables, and its "aptness" is direct and apparent, not a matter for conjecture. That Swets's investigation of the matter may not have convinced him of this may have several causes. First, he may not himself have clearly distinguished between human signal detection of relevance, and detection by an explicit retrieval process acting as a proxy for human decision-making, notwithstanding the definition adopted by him in the first quotation given above. Since this distinction may not be obvious, it is treated in more detail as follows. In the author's view, two signal detection processes relevant to information retrieval can be defined:

> Signal Detection Process I: A human observer, X, is confronted with document descriptions and has to decide, which descriptions denote documents that are (unknown to him) relevant, and which non-relevant, in respect of some information used. The identification of documents as relevant or non-relevant could be on the basis of either (1) another observer, Y, labelling the documents in just this way, or (2) X inspecting the documents subsequent to his decision making, with his own (non-verbal) notion of information need dictating his labelling of them as relevant or non-relevant. In this situation the decision process is hidden in that s, f_1,f_2 and β are not directly observable. The situation is therefore close to that involved in psychophysical experiments, but it is not a situation treated in Swets's writings, and to the author's knowledge the process has not previously been considered in the literature of information retrieval.

Signal Detection Process II: An abstract process (in fact an algorithm) examines each document description in a set of same. Some document descriptions relate to documents relevant to some information need, some to documents that are not relevant. On the basis of a query i.e. a search-statement in some form (which is input to, and forms part of the algorithm) the process labels documents as relevant or non-relevant. The labelling is fallible. No human intervention is involved other than providing to the algorithm a search statement (and perhaps other parameters). Given an identifiable subset of relevant documents, z, f_1, f_2 and β are all directly observable.

The two processes just described are conceptually different although a deeper, philosophical approach to the study of observers and the transfer of information, beyond the scope of this dissertation, might lead to some unifying view. Swets clearly addressed Process II, in our terminology. What is being suggested therefore is that his evident doubts on its validity may have been associated with his not distinguishing it from Process I.

A second cause of Swets offering the disclaimer quoted above, not independent of the first reason just given, may have been the imprecision attached by Swets to certain terms, in particular the terms "pertinence" and "query". This statement will be justified shortly.

A third reason is that Swets's approach was relatively informal in character. Had there been a careful, formal description of the process, the caution he attached to its introduction would not have been necessary: and what we have labelled as Process II could have

been introduced for its unifying and possibly predictive power, rather than as a somewhat covert background justification for novel measures of effectiveness. That it dealt entirely in observables, whatever deeper basis those observable might have from other points of view, would have been an additional reason for giving the formalism more emphasis than he did.

Swets then introduced the formalism more fully as follows:

> "Let us assume that when a search query is submitted to a retrieval system the system assigns an index value (call it z) to each item in the store (an item can be a document, a sentence or a fact) to reflect the degree of pertinence of the item to the query. (Maron and Kuhns have described a particular procedure to accomplish this assignment, but let us regard such a procedure, in general, as a feature of all retrieval systems.) Now it may be that for a given need, or for the need as translated into a search query, the items in a given store do in fact vary considerably in pertinence, from a very low value (or no pertinence) to a very high value (or full satisfaction of the need). On the other hand, all of the items may in fact (according to expert opinion of the user's opinion) be either clearly nonpertinent or clearly pertinent to the need. In either case the retrieval system, being imperfect, will view the items as varying over a range of pertinence; indeed, because of the error which will exist in any retrieval system, the value of z assigned to a nonpertinent item will frequently be higher than the value of z assigned to a pertinent item.
>
> "Thus we assume that the retrieval system assigns a fallible index of pertinence, z, and that there exists,

apart from the retrieval system, a knowledge as to which items are 'in truth' pertinent and nonpertinent." (p.247)

The above quotation brings out several points. The first is an answer to the fundamental question: If, in signal detection theory, the basis is a fallible detection device coupled with a decision process, the two being regarded jointly as an observation process, what are we identifying with this concept in regard to the retrieval process? The answer, according to Swets, is that the "fallible detection device" is the combination of query and number-assignment process. Thus, in the sense of the treatment given in Section 3.1, each query (among other variables) defines a separate observation process. The fallibility, as Swets clearly describes it, is such that two documents, one relevant to the need, the other non-relevant, can have z-values suggesting the reverse. (Here, as elsewhere, Swets's term "pertinence" has been translated by the author into the currently more acceptable term "relevance".) A second point that the quotation brings out is that Swets is inconsistent in the meaning he attaches to the term "pertinent" (i.e. relevant). In the last section of the quotation (from "indeed, because of the error" to the end of the passage) he describes pertinence as a binary attribute of documents: an item is either "a nonpertinent item" or it is "a pertinent item". Again, items are " 'in truth' pertinent and nonpertinent." But in the earlier part of the quotation he attaches a different, almost quantitative meaning to the term. He uses the phrases "degree of pertinence of the item to the query" and "the items...do in fact vary considerably in pertinence". There is thus an inconsistency in his presentation, one expressed in signal detection language as

that between assuming simultaneously (1) a signal of constant form, and (2) a signal of variable form, (and even (3) a signal of randomly variable form, i.e. a stochastic signal). Possibly Swets was influenced in his description by copious (and unsubstantiated) references in the information retrieval literature to quantitative variability in relevance. But it seems clear from the subsequent development of the formalism by Swets that the binary input event was the one intended, so that the point is just a semantic one, not a scientific one. This is not to say that a revised version of the formalism Swets then described could not incorporate quantitative or qualitative variability in the relevance assigned to documents that are input to retrieval process.* It is just to say that Swets's formalism was originally not of this character. This statement is not affected by the fact that in his American Documentation paper Swets analysed the results of experiments in which variability in relevance was assigned to documents prior to processing. The analysis was in fact of the "Cranfield data" of Cleverdon and Keen (1960). In that analysis, represented graphically in Swets (1969) as Swets's Figure 12 and by a brief comment on p.79, Swets applied the binary input formalism to sets of relevant documents defined using different document "relevance levels". (These levels were "1" (most relevant), "2", "3" and "4" (least relevant). The four sets of relevant documents, for individual sets of data. were based in effect on inclusion criteria of the form: all documents with relevance levels from 1 to J, where J ranged from 1 to 4.) From a signal detection point of view however, this represented four separate implementations of the basic binary formalism, each 'signal' being defined in a different way. (In the case of signals defined as 'those documents of

* Hutchinson (1978) has more recently suggested this as described in Section 3.3.2.2.

relevance 1 to 2' for example, the set of signals is a subset of those defined by 'those documents of relevance 1 to 3', etc.) A prior labelling of documents as signal or non-signal (noise) is of course necessary if the binary response of a retrieval system is to be judged correct or not.

We now recall that the situation being described is that where there is binary input and binary output, the two together defining a 2X2 table of events. Signal detection theory exists in order to predict the frequencies of occurrence of input/output events assigned to each compartment of this table, for a given observation process. But in introducing Swets's application of this to the information retrieval process we have not accounted for the random variability in the detection device, a necessary feature of the formalism. If this variability is not (for a given observation process) due to variability in pertinence, to what is it due? Swets does not discuss this point explicitly, but perhaps because it is obvious: the variability in z is simply due to variability in the set of attributes assigned to each document by the producers of the data base. This variability exists irrespective of the choice of set of documents each member of which is regarded as a 'signal', and of whether such labelling is to incorporate some subjectively-recognised notion of 'degree of relevance', or some other more specific attribute sufficient to denote a document as 'signal'. To emphasise this point: the signal detection formalism introduced by Swets has randomness in it which is solely the result of action in the receiver, i.e. a stochastic-signal model was <u>not</u> intended. The receiver, being a fixed, non-stochastic algorithm assigning z-values to documents, has a randomness in its action due entirely to variability in the set of

attributes attached to documents for a given question and similarity measure.

The decision-making part of the observation process was then described by Swets as follows. A threshold value (or "acceptance criterion" or "cutoff") is defined such that documents with z values higher than z_c are retrieved, and documents having $z < z_c$ are rejected. (The physical form of 'retrieval' need not concern us: the document descriptions concerned could be brought to the attention of the user of the process, or the documents themselves could.) In a later statement (Swets, 1963: 249) Swets commented: "Strictly speaking, it is assumed in statistical theory that the z-axis ... is a scale-of-likelihood ratio ...". A formula in support of this statement is quoted, in terms of prior odds and utilities (anticipating the later accounts by Coombs, and Bookstein) showing how this criterion can be objectively determined. The formula differs only in notation from that described in Section 3.1 here. (The quotation just given can again be criticised as loosely worded: "statistical theory" can assume either type of criterion. The difference, as we saw in Section 3.1, is that use of a likelihood ratio criterion leads to superior performance (i.e. higher utility values) only when f_1 and f_2 are not such that l(z) increases monotonically with z.) The comment by Swets does clearly show however that he saw z as identifiable with <u>either</u> a value equivalent to the value of the receiver output (i.e. the value we have labelled z in Section 3.1), <u>or</u> a value equivalent to a value of the likelihood ratio of that z-value (i.e. the value we have previously labelled as l(z).) Notwithstanding the two possibilities here, i.e. the ambiguous status Swets gave to "z", all his subsequent discussion (and

diagrams) clearly identify z with the receiver output value, rather than its likelihood ratio. Before leaving this point we emphasise that Swets was clearly aware of the value of the likelihood ratio criterion, and that the best such criterion was "determined by the values and costs appropriate to a particular retrieval need" (Swets, 1963: 248, and formula quoted on p.249).

The effectiveness of the information retrieval process is then, according to Swets, expressible via four probabilities of the form: probability that a document is retrieved given that it is relevant or non-relevant, which we label by the conventional information retrieval terms 'Recall' (R) and 'Fallout' (F) respectively; and: probability that a document is not retrieved given that it is relevant or non-relevant , (1-R, and 1-F respectively). For comparison with Swets's notation we note:

Notation used	Swets's notation	Verbal equivalent
R	$Pr_p(R)$	Recall (= probability that a relevant document is retrieved)
F	$Pr_{\bar{p}}(R)$	Fallout (= probability that a non-relevant document is retrieved)
1-R	$Pr_p(\bar{R})$	Probability that a relevant document is rejected
1-F	$Pr_{\bar{p}}(\bar{R})$	Probability that a non-relevant document is rejected
$f_1(z)$	$f_p(z)$	Probability density of a variable z, usually the receiver output value in Swets's work, and always this in the present work. The probability is defined for the set of relevant documents.
$f_2(z)$	$f_{\bar{p}}(z)$	As for above, but defined for the set of non-relevant documents.

Then R and 1-R are related to $f_1(z)$, and F and 1-F are related to $f_2(z)$, as follows. (The names Pr_1 and Pr_3 of Section 3.1 are equivalent to R and F respectively.)

$$R = \int_{z_c}^{\infty} f_1(z)\, dz \quad ; \quad F = \int_{z_c}^{\infty} f_2(z)\, dz.$$

We stress that the above relationships are true only if z is interpreted to be the receiver output value. If z is construed as the likelihood ratio of the receiver value (call this temporarily x, so that $z=l(x)$), then Swets's definitions need to be replaced by the following:

$$R = \int_I f_1(x)\, dx \quad ; \quad F = \int_I f_2(x)\, dx \quad ; \quad \text{where}$$
$$I = \{x: z \equiv l(x) > z_c\}.$$

Swets does not describe or pursue the latter possibility beyond the extent indicated earlier.

In terms of the 2X2 table, which according to Farradane (1974: 201) was introduced into information retrieval by Swets, the probabilities can be described as ratios as follows.

	Retrieved	Not retrieved
Relevant	a	b
Not relevant	c	d

(a,b,c,d are document frequencies, so that a+b+c+d equals the size of the data base so analysed.)

Then: $R = a/(a+b)$ and $F = c/(c+d)$.

We note in passing that the 2X2 table has since become a standard

concept in the methodologies of those investigating retrieval effectiveness. Numerous measures based on the table, or expressible in terms of the frequencies in the table, have been suggested. Such measures have been extensively reviewed: see for example Keen et al (1972), King (1971), Salton (1975a), Robertson (1969), Vickery (1970), and van Rijsbergen (1979b). In view of the depth of the review work done on this subject, such a review is not repeated yet again here. We also note that it is widely acknowledged that probabilistic measures based on the table, i.e. based on the degree of coincidence of the sets of relevant and retrieved documents, represent only one approach to measuring or assessing retrieval effectiveness, a point also, incidentally, emphasised by Swets. Fuller approaches also involve (for example) operating cost, form of presentation of output, the retrieval system's speed of response, and the subject scope of the data base.

Two hypothetical families of ROC graphs arising from variation in the information retrieval process were then described by Swets. One family is characterised by pairs of density functions $(N(\mu_1,\sigma_1), N(\mu_2,\sigma_2))$ where $\sigma_1=\sigma=\sigma_2$, the variation being in the value $(\mu_1-\mu_2)/\sigma$. The other family is determined by pairs $(N(\mu_1,\sigma_1), N(\mu_2,\sigma_2))$ such that $\sigma_1-\sigma_2=(\mu_1-\mu_2)/4$, the variation being in $\mu_1-\mu_2$. No basis for defining such families was offered, although both types are considered in the literature on signal detection theory as applied to psychophysics. In the first-mentioned ROC graph family, Swets also illustrated the effect of plotting the family when the probability values are transformed into standard Normal scores, when a family of straight lines results. He demonstrated how values of $(\mu_1-\mu_2)/\sigma \equiv E$ (in his notation) can be used to calibrate the negative diagonal and

so uniquely characterise each curve by a number. The latter graphs, relabelled, are reproduced here in Fig. 3.2-1 (based on Swets, 1963, 249 and Fig. 8). The "Normal deviate" scores u_R and u_F are defined using the complement of the standard Normal probability function Φ:

$$\Phi_c(u) = 1 - \Phi(u) = \int_u^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-y^2)\, dy.$$

That is, for a probability value p, u is defined by $\Phi_c(u)=p$, or $u=\Phi_c^{-1}(p)$. For Recall probabilities, R, based on an unstandardised Normal density function, $N(\mu_1,\sigma_1)$, the value of $u=u_R$ is thus $u_R=\Phi_c^{-1}(R)$, where u_R is expressed in units of $(z_c-\mu_1)/\sigma_1$. Thus $u_R=0$ corresponds to $z_c=\mu_1$. Similarly Fallout probabilities, F, based on a $N(\mu_2,\sigma_2)$ density function are associated with a value of $u=u_F$ obtained as $u_F=\Phi_c^{-1}(F)$, expressed this time in units of $(z_c-\mu_2)/\sigma_2$. Thus $u_F=0$ implies $z_c=\mu_2$. Hence if $\mu_1\neq\mu_2$, as is usually the case in signal detection processes, $u_R=0$ does not imply $u_F=0$. On the other hand a linear relationship between u_F and u_R follows immediately from the relationship each variable has with z_c:

$$u_R = au_F + b \;;\quad \text{where } a=\sigma_2/\sigma_1,\quad b=(\mu_2-\mu_1)/\sigma_1 \equiv -\underset{\sim}{E}.$$

Since $\sigma_2/\sigma_1>0$, the ROC graph (a straight line) determined by choosing different threshold values, z_c, has positive slope, and the slope is in fact unity when $\sigma_1=\sigma_2$. The value of u_F is in practice always less than the value of u_R for any point on such lines since $\mu_2-\mu_1<0$. In the case where $\sigma_1=\sigma=\sigma_2$ this reduces to saying that for any point lying on a line of value $\underset{\sim}{E}$, the two coordinate values differ by $\underset{\sim}{E}$, since:

$$u_F-u_R = \frac{z_c-\mu_2}{\sigma} - \frac{z_c-\mu_1}{\sigma} = \frac{\mu_1-\mu_2}{\sigma} = \underset{\sim}{E}$$

Swets discussed the problem of which value to choose for z_c as follows:

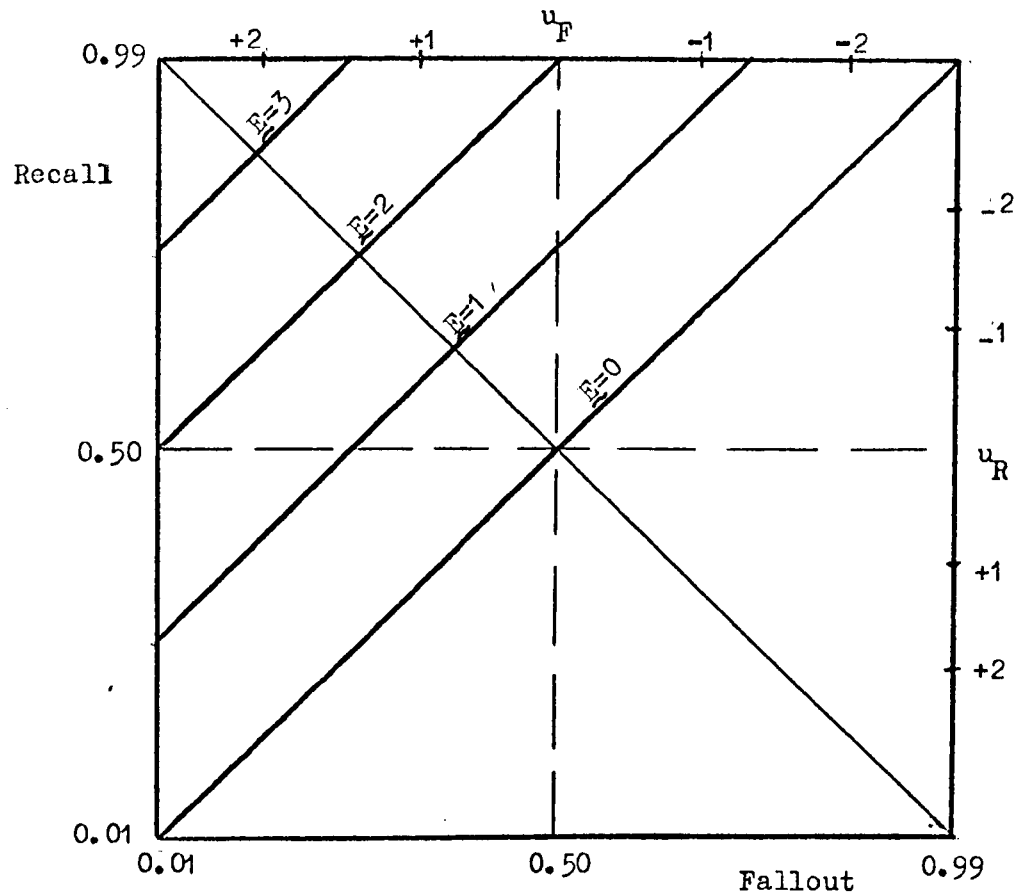


Fig. 3.2-1. ROC graphs for $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$ density functions, for the case where $\sigma_1=\sigma=\sigma_2$, and for various values of $(\mu_1-\mu_2)/\sigma$ ($\equiv \underset{\sim}{E}$). (After Swets (1963: Fig. 8).)

> "If the user is willing to examine a good deal of non-pertinent material in order to reduce the chance of missing a pertinent item, the cutoff should be low. Alternatively, if time or money is an important factor and a miss is not very serious, the cutoff should be high. Similarly certain a priori probabilities may affect the level of the desired cutoff. If the user has good reason to believe the store [i.e. data base] contains the item he wants, he may choose to make a relatively thorough search; if he is doubtful that the store contains the item he requires, he may prefer a token search, of only the items most likely to be responsive to his query." (p.248)

The latter seems fairly clearly in agreement with the decision to choose z_c on the evidence of prior odds and a set of utility values attaching to possible outcomes, as described in Section 3.1. The last sentence is a little obscure however in that it hints again at the notion of degrees of relevance. What seems to be intended is: '...he may prefer a token search, of only those items that are attached to z-values most likely to yield relevant documents.'

Swets then adds:

> "In practice, the level of cutoff may be set ... by the choice of a form of query. The choice of an 'and' or 'or' relationship among a set of key terms, and the selection of the number of key terms, are ways of determining the breadth of the query and thus the level of the z-axis cutoff."

The point being made seems a fundamental one. Previously the "query", expressing an enquirer's need for information, was introduced as the

basis for assigning z-values to each document. Now an additional role is suggested for it: that the query should form an input to the decision to calculate the threshold. (This is presumably in addition to the other input information we described in Section 3.1: prior odds, utility values and a stored array of likelihood values.) This represents a radical change in signal detection theory which conventionally sees the recording device as independent of the decision process. The suggestion seems a most useful one since the behaviour of the receiving device must be dependent on the form of the question. The introduction of Boolean operators to define the query as a Boolean expression was, as can be seen from the quotation, also touched on. It is regrettable that Swets did not develop this point, or indeed attempt to explore the range of meanings that the term "query" can connote. With hindsight, this seems understandable though, since at that time equivalences between Boolean expressions and functions based on comparisons between sets were not widely understood, not by workers in the information retrieval field at least. This point will be carried further in future sections.

We have now seen that Swets identified the range of response of an information retrieval process to a single query, with a ROC curve (and its E-value). The particular value of z_c (and thus a particular point on the ROC curve) would be chosen arbitrarily or would be influenced by knowledge as to prior odds, utility values, the form of the f_1 and f_2 distributions (or equivalently the likelihood values f_1/f_2 at each value of z), and the query form, the latter being left as a soft (undefined) concept. This question naturally leads one to further questions such as: "Will the ROC graphs (or indeed other graphs based on f_1 and f_2) vary from query

to query, for a given set of signals (i.e. set of relevant documents)?", "Will such graphs vary when the sets of relevant documents change, in reference to different information needs, but remain of constant size?", and "Will such graphs change when sets of different size are defined?" Such questions are largely experimental questions. The formalism alone, as will be demonstrated later, will partly predict the variation with query for a given query form: for example when the query has the form of a logical expression of varying logical structure. It does so through an analysis of the (immediate) outcome space of the receiver. But such questions are essentially hypotheses to be tested experimentally rather than questions about the formalism. There is room for adding practical constraints as well: Will the ROC graphs and other graphs vary with the 'subject' of the data base taken as a whole? Will they vary with the depth of indexing of documents (e.g. with the expected number of terms assigned to documents), with the size of the set of indexing terms, the manner of assigning terms to documents (human assignment versus algorithmic assignment), or the manner in which the terms comprising the query are chosen? The variability in all the components of the information storage and retrieval process will determine the variation of the distributions f_1 and f_2, and associated variables, and in particular the ROC graph. As a response surface or combined effect of such causes, the ROC graph thus provides (as do also other related graphs) one criterion by which optimisation of the retrieval process can be judged.

Returning to Swets's description, it is noted that Swets did suggest that changes in the form of a query [the writer's emphasis] were unlikely to influence the ROC graph for a given document

collection, retrieval language and depth of indexing (and, possibly, for a fixed information need to which different forms of query relate.) As previously noted, no precise meaning was attached by Swets to this term "form" however. This assumption was first introduced by Swets as follows:

> "Of course, the assumption that a real retrieval system has a constant effectiveness, independent of the various forms of queries it will handle, is open to question. It seems plausible, however, that the sharpness of the retrieval system's query language, and its depth of indexing, and also the heterogeneity of items in store, will determine a level of effectiveness that is relatively invariant over changes in the form of the query. In any event the assumption is subject to empirical test, and its importance is sufficient to justify the effort of testing." (p.248)

The "level of effectiveness" mentioned in the above was presumably intended by Swets to be "$\underset{\sim}{E}$", or some other variable characterising the ROC graph, but we note that for a fixed value of Fallout [Recall] the Recall [Fallout]value will increase monotonically with $\underset{\sim}{E}$, so that either one of these probabilities could be substituted. (That is, effectiveness could be assessed through the variation of R, at a fixed F value of 0.1 say.) Although, as emphasised above, Swets was suggesting that the ROC graph was invariant to the form of the query (by implication, for a given information used) it seems that he also viewed the graphs as invariant to the information need itself, i.e. for various queries pertaining to various needs. As this was not stated explicitly by Swets, this interpretation may

therefore be unwarranted, but his 1969 paper reports extensive analyses of experimental data in which data pertaining to different information needs* were "pooled" (i.e. confounded). So possibly in his using the phrase "form of query" he was mis-stating his own position, and invariance in the ROC graph in the face of simultaneous variation in information need and query was intended. The underlying presumption was, perhaps that the query pertaining to each need was formed in some more or less constant way. By way of clarifying the difference more fully, we note that two information needs may be characterised by the two queries:

$$Q1 = \{t_a, t_b, t_c, t_d, t_e\} \quad \text{and} \quad Q2 = \{t_i, t_j, t_k, t_l, t_m\}$$

both queries (so specified) being of the same "form" (i.e. both expressed as a set of attributes). Then one hypothesis would be:

H1: The ROC graphs for Q1 and Q2 are the same (are drawn from the same population of ROC graphs). This hypothesis embodies the thought that the ROC curve is invariant to queries of constant form and size, irrespective of information needs.

Another hypothesis would be:

H2: The ROC graphs for $Q1 = \{t_a, t_b, t_c, t_d, t_e\}$

$Q1' = \{t_a, t_b, t_c, t_d\}$

$Q1'' = \{t_a, t_c, t_d, t_e\}$

(etc.)

are the same; i.e. the ROC graph is invariant to query form for a given need, where 'query form' is defined to mean here 'of the form of a set of terms, of arbitrary size' (in distinction to a Boolean form, say), the sets being subsets of some parent set.

* We shall later have occasion to criticise the experimental design yielding this data: the statement here is generously worded in order not to introduce that weakness at this stage and obscure the point made here.

Yet another hypothesis, defining query form to be 'a set of terms of constant size' and, like H2, exploring the effect on the ROC graph of various query forms for a given information need, would be:

H3: The ROC graphs for $Q1 = \{t_a, t_b, t_c, t_d, t_e\}$

and $Q1' = \{t_f, t_g, t_h, t_i, t_j\}$

etc.

are the same.

So one ambiguity we are pointing to is that Swets appeared to suggest that H2 was true, but in the experimental analysis reported in his 1969 paper he assumed H1 to be true. Another, completely different, interpretation of the preceding quotation is however possible. This is that in referring to "various forms of query" Swets had adopted a new meaning of query (i.e. observed a different usage of it), namely query as a synonym for information need. (This rather clumsy and certainly misleading usage is embodied in the phrase "relevance to a question".) If that were his usage, in this particular part of his contribution, then he was suggesting that the whole signal detection process should be thought of as a fixed one pertaining to all instances of information need, all similarity measures, and all instances of query (in the former sense: as a set of attributes). Swets's method of data analysis involving the treatment of confounded data (rather than data pertaining to individual combinations of need, weighting function and query as a set of attributes) might appear to be consistent with this. However, in the writer's opinion, the hypothesis is such a sweeping one that it can be seen immediately to be invalid*, and accordingly it seems much more likely that one of

* This is in fact demonstrated in the experimental work described in Section 4 of this thesis.

the earlier interpretations of the passage was intended by Swets, and that the later treatment of confounded data was a consequence of several conceptual errors: (1) that invariance in a set of processes can be demonstrated by defining a (static) composite process, and (2) that modelling a process (in Swets's case using Normal density functions) is equivalent to recording data on it. (The author's later extension of Swets's work departs from both these assumptions.)

The 1963 paper contains discussion on several further points. The difficulty of distinguishing ROC graphs "on quite extreme variance ratios" was commented upon as is the difficulty of obtaining "enough data to reject the normality assumption". He also remarked:

> "The slope of the [ROC] curve at any point will serve as an index of the particular acceptance criterion, and of the breadth of the search query, which yielded that point." (p.249)

The first statement here follows from the definition of the ROC graph, the values of z_c determining unique pairs of probability values that form the co-ordinates of the graph, and conversely, but only if the ROC graph is "proper". (We recall from Section 3.1 that such a ROC graph is determined by a likelihood ratio criterion, or by a receiver-output value criterion in the case where this is monotone with same.) If it is not proper, the ROC graph may have two points with identical slopes, and the statement by Swets is false. The second statement is not clarified or justified by Swets, and creates even more semantic difficulty in that it involves a new usage for the term "query". Previously a "query" was defined as an input to the receiver, i.e. to the z-value generation process. Accordingly

the entire range of z-values is a reflection of the choice of query. Now a second definition is being introduced, in which a query is being identified with a point or interval within the range of z. It is not a casual contradiction, as Swets goes on to repeat the second (implied) definition on three further occasions:

> "It is clear that the difference in the slopes at two points of a steadily rising function is a straightforward measure of the effective change in the breadth of a search query." (p.249)

> "For any given query or form of query, these two probabilities [R and F]can be plotted as a points in the unit square of Fig. 6." (p.250)

> "... the slope of the curve at that point is a measure of the query breadth." (p.250)

The second definition of "query", to which the last four quotations relate, is a vague one since "breadth" is not defined. That two contradictory definitions of "question" should have been introduced may also account (along with the contradictory definition of "pertinence", and the ambiguity in the hypothesis of invariance in the ROC graph) for the scarcity of critical discussion of Swets's work in the years following its appearance, and also for some confusion about the formalism. Farradane, for example, has criticised the formalism on just the ground that it involves a variation in the query. This is to concentrate on the second definition, however, and to ignore the constancy of query required (per the first definition) to assign z values to documents. (Farradane, 1974 and

pers. comm.). We postpone further discussion until a later section (Section 3.3.1.2) since an extension beyond Swets's signal approach is called for.

Before summarising this treatment of Swets's presentation of his formalism, some further comments directed at his 1969 paper are offered. This paper again described information retrieval in signal detection terms. The interpretation of $\underset{\sim}{E}$ as a measure of separation of f_1 and f_2 was further clarified, and a related distribution-free measure, $\underset{\sim}{A}$, was introduced defined by:

$$\underset{\sim}{A} = \int_{z_c=-\infty}^{\infty} R.dF \ .$$

This paper also commented in a little more detail on the range of forms that f_1 and f_2 could take, negative exponential densities being described as well as the Normal equal variance and non-equal variance cases. The signal detection formalism was not however re-examined in any greater detail. "z" was again described as if it were the output of a receiver rather than as a likelihood ratio of same - the possibility of which the 1963 paper had mentioned. "Relevance" was used in place of the 1963 paper's "pertinence", and Recall or Recall ratio was now used for the probability of retrieval conditional on a document being relevant, i.e. the now-conventional usage was observed. "Fallout" was not introduced as a term however, although as before the concept was freely used. Unlike the first paper, the second did comment on the probability that a retrieved document is relevant: the "Precision(ratio)", or "Relevance (ratio)" as it was once called which we will denote by P. Swets did not pursue the relationship between Precision and Recall and/or Fallout however, restricting his contribution here to (1) the observation

that a pair of P and R values does not allow all compartments of the 2X2 table to be reconstructed from it, and (2) a figure showing an "idealised" graph of the Recall-Precision relationship.

Although Swets did not say so directly, he seemed to imply (p.75) that an invariance existing in the ROC graph is not inherited by the Precision vs Recall graph. This implication is however true in the formalism. The relationship $P=GR/(GR+(1-G)F)$, linking P,R,F and G, where G is the 'Generality of the set of relevant documents' (signal to noise ratio), is valid in the formalism.* That is, it is not an empirical relationship but an exact one following from the definitions of these quantities. If therefore in consideration of different information needs, with (in general) varying G values associated with them, it is found that the R vs F graph is invariant for some class of queries, it cannot be the case that the P vs R graph is also invariant. G is explicit in their relationship. Swets may well have understood this intuitively, although the only explicit reference to the influence of signal to noise ratio sensu stricto is through his estimate of same as an input to the determination of z_c, in the 1963 paper. (The a priori odds are equal to, or an estimate of Generality). Figures for signal to noise ratio proper, i.e. with the phrase used as a synonym for the Generality of a relevant set forming a subset of a data base, are in fact given in a discussion of examples towards the end of the 1969 paper (pp. 87-8), and the data quoted on same together with other data allow a Recall vs Precision relationship to be inferred there. But unfortunately further needless semantic confusion is introduced by Swets when he refers to "noise-to-signal"

* We can equally refer to G as the 'Generality of the information need as represented in a set of documents' without changing the concept itself.

ratio as an attribute of the retrieved set. That is, he uses the phrase at that point as a synonym for Precision when expressed as a ratio and inverted. (E.g. when 4 relevant items are retrieved along with 30 non-relevant items, the total number of relevant items being 10 and the size of the data base being 3000 items, the noise-to-signal ratio (per Swets) is 30:4, or 1/Precision, whereas from the point of view of the basic signal-detection formalism, the noise to signal ratio is 2990:10.) In the author's view, the ambiguity attaching to this concept is another reflection of uncertainty, in the original presentations of the theory, as to whether it was signal-detection by humans, or by machines acting for human beings, that was being described. We refer again to the two processes labelled "Signal Detection Process I" and "Signal Detection Process II" given earlier in this section.

Swets does not clarify the basic concepts he uses beyond the stage of the 1963 paper. If anything, the issue of whether fixed or variable relevance in the signal is assumed is confused further, when he writes:

> "It will become clear, by the way, that the decision-theory measure can be applied when judges use several, rather than two, categories of relevance, and that it uses to full advantage the output of a system that ranks or otherwise scales all items in the store according to their degree of relevance to the query at hand." (p.73)

This and later passages show that he uses "degree of relevance" for what he rather more clearly referred to in the 1963 paper as "a fallible index of relevance ", i.e. a z-value that "reflected" a likelihood of (binary) relevance in the way that we earlier described

at length. Again, in the 1969 paper, input to the process is treated as binary in the presentation of the formalism.

The larger part of the 1969 paper is a report of the ROC graphs pertaining to data obtained in three experimental investigations (Cleverdon et al (1966), Salton et al (1966), and Giuliano et al (1966).) Swets worked on "pooled" data in the first two cases, but the Giuliano data related to separated instances of information need and query. The ROC graphs from the experimental data were fitted by straight lines "by eye", in an attempt to demonstrate the suitability of Normal-density forms for f_1 and f_2 (for averaged data). One is inclined, on this graphical evidence, to agree with Swets's view that a straight-line fit is acceptable or "very good" in most cases, though no explicit comment on the sensitivity of such a test is offered. (The inclusion of theoretical ROC graphs based on negative exponential densities for f_1 and f_2, which are "by eye" almost straight lines over the ranges of R and F yielded by the experimental data, is a *prima facie* indication of the insensitivity of the test.) In the case of the Giuliano data, Swets comments:

> "The data points, surprisingly, do not show much greater scatter about a line, but substantially greater variation in the slopes is evident." (p.81)

We do not discuss the analysis in greater detail here for several reasons. First, it was itself an analysis of earlier experimental work that we would need to comment upon in great detail. Secondly, Swets's analysis was presented only in graphical form (as ROC graphs); no numerical analysis was offered. Thirdly, there is a critical and invalidating conceptual weakness in the data analysed,

a point taken up further in Section 3.3.1.2.: The experiments concerned, with the possible exception of Giuliano and Jones's, involved pseudo-sets of relevant documents, since they were defined with reference to a verbal description of information need (or to a linguistic form) rather than to information need as a psychological (and not a priori-verbal) process. It is accordingly very surprising that analysis of such data was undertaken at all, given that Swets's own formalism placed such an emphasis on information need as the primitive entity, evident in the signals transmitted to the retrieval process, and treated the question as a variable articulant in a variety of forms serving to optimise identification of such signals. Later work may indicate that the distributions f_1 and f_2 are insensitive to this feature of the experimental design of the data analysed, but for the moment at least the results are at least of unproven validity, and in the author's view are meaningless.

Swets concluded his second paper with a discussion of examples relating to the number of non-relevant documents retrieved for different Recall values, in effect concentrating more on the Precision/Recall balance, and offered several conjectures as to the kinds of $\underset{\sim}{E}$-value that should be realiseable in the future. In keeping with the developing technology of the time, he also introduced the notion of "on-line" dialogue as a means of improving questions (through feedback), thereby again emphasising the role of questions in information retrieval as variables. He conjectured that $\underset{\sim}{E}$-values of 3.0 or 3.5 may be obtainable for such systems.

One feature of the formalism in its simplest form that may prove to be a significant weakness is in part prompted by the

criticism of the Recall concept by Cooper (1973, 1976). This feature is the supposition that documents in a data base can be marked in a way that denotes their relevance to an information need, i.e. it relates to the postulate of a partitioned data base. Although this seems a simple concept in principle, and although such marking can be implemented in experimental tests of appropriate hypotheses, it nevertheless remains true that in an operational environment relevant documents are not known in advance of the retrieval process being implemented. That is, Recall is an unknown, and relevance judgements are (in practice) made on retrieved sets, not on the whole data base. This weakens the formalism in that it can then be seen as describing a feature (the signal) that is an observable only in principle. It would, accordingly, appear preferable to have a formalism centred on the retrieved set, or on succession of retrieved sets, if one sees the retrieval process (for a fixed information need) as a heuristic one, guided through the data base by a sequence of successively more-accurate questions. (We ignore the complication that knowledge-acquisition itself will be, presumably, heuristic at a deeper level.) Indeed one may conjecture that Swets's usage of the term "noise-to-signal" ratio at the end of his 1969 paper, which he clearly related there to the retrieved set, may have been prompted by this thought. If so, one is again prompted to think that a useful further development of the formalism would be along the lines we have labelled "Signal Detection Process I", or towards a structure embodying both of Processes I and II. Possibly comment in the literature pointing to the symmetrical treatment of the relevant set and the retrieved set (with 'degrees of relevance' serving as the analogue of receiver output values)

will provide such a structure. Hutchinson (1978) has developed a bivariate formalism, said to be based on a comment of Robertson's relating to the generalisation of the 2X2 table (1969: 8). Robertson has noted Fairthorne's concern for symmetry in the 2X2 table (Fairthorne, 1964). The author has offered a brief symmetrical treatment of one measure of retrieval effectiveness based on measures analogous to Fallout and Recall, namely 1-F and R respectively (and equivalent to the earlier Western Reserve University measures of "Specificity" and "Sensitivity", respectively), with "Retrievality" forming the appropriate analogue of "Generality" (Heine, 1973b:33). Possibly there are precedents in the signal detection literature itself. The major task of developing a unifying signal detection formalism is not attempted in this dissertation however, although the argument is carried a little further, by way of the extension of the theory into "heuristics" in Section 3.3.3.3. The author's view is that the theory in its original form should, at least at the present time, be seen as the basic one, upon which this and other generalisations of it can be built. But intuitively there remains a reasonable doubt that a formalism building on an entity which, though observable in principle (and in experiments) is unobservable in practice, is expressed in its optimum form.

Lastly we note that the matter of estimation, in the interpretation of the results of retrieval experiments, is only briefly mentioned by Swets (e.g. Swets, 1969: 74). Considerable credit is due to Robertson (1975) for drawing attention to the need for this. Swets's approach was 'scientific' rather than formal-inferential, in that the degree of scatter of data points (for confounded data) around ROC graphs pertaining to models of processes was portrayed

for what it was, without levels of confidence in the accuracy of the models (as population descriptions) being estimated.

Two summaries of Swets's contribution will now be offered. The first is informative with some additional discussion, the second shorter and indicative.

Informative summary with further discussion.

The information retrieval process can be described as a signal detection process, the latter being an abstract representation of the process of observation. Both a formalism, and descriptive hypotheses expressed within that formalism, are involved. The signal, as an attribute of a document, reflects (or 'is') a notion of 'relevance' to an 'information need'. Both the latter notions are left undefined, i.e. are primitive concepts in the formalism. Unlike the situation described in psychophysical applications of the theory, the randomness in the receiving device, construed as a combination of query and analytic mapping function, is due not exclusively or primarily to a stochastic randomness in the input events, but instead solely to a randomness in the behaviour of the receiving device for signals of constant value. In practical terms,

this randomness is due simply to variety in the sets of attributes attached to documents, as detected by a 2-tuple of query and mapping function. The mapping function measures the similarity of query and attribute-set, i.e. maps each distinct pair (query, document) to a real number on the basis of their similarity. The output of the 'detection device', a set of possible values, is then input to a decision process. The latter is essentially of the form of a threshold value, having the effect that only documents mapping to values greater than or equal to the threshold value are identified as relevant, i.e. retrieved. A question is thus to be seen as a variable entity for a fixed information need. Such notions, like various others, were implicit rather than explicit in Swets's theory and are extricable only when terms such as "pertinence", "query", and "query form", used ambiguously by Swets, are given an interpretation.

Notwithstanding the basic concept of relevance as a binary quality (documents being either relevant or not), Swets did briefly consider the notion of variability in relevance, just as he also considered in passing the usefulness of inputting likelihood values of the similarity-measure values, to the decision process. It is again partly implicit that Swets regarded the information retrieval process as one varying from need to need, and query to query. This is notwithstanding his stating that hypotheses describing invariance of a certain character (in fact, shape of the ROC graph) can usefully be advanced. Such hypotheses were both ambiguous, and not clearly distinguished from the formalism itself, and the two together (hypotheses and formalism) were not primarily advanced in order to provide a joint characterisation of the retrieval process. Instead

they were introduced by Swets primarily as an accessory to definition of novel measures of retrieval effectiveness, and to the applications of these to the evaluation of particular retrieval processes.

Despite the (immediate) origins of the theory in work on psychophysics, Swets's theory is not a theory of the perception of relevance in documents by human beings. This is so since the entities involved are all observables - at least in experimental situations - unlike the entities that make up the theory as it is applied to experiments on auditory etc. perception. As applied to information retrieval, the theory is a description of the behaviour of processes acting as proxies for human behaviour.*

The strength of Swets's theory appears to lie in four areas:

(1) It accounts in a systematic way for the presence of error in the information retrieval process. Error is not viewed as something explainable at the microscopic level, and avoidable; i.e. the theory does not concern itself with the question: 'for a given document, was the correct decision made on it by the decision process'. Instead error is regarded as a macroscopic phenomenon capable of systematic, objective description through probability distributions. This macroscopic view is, as was also stated in Section 3.1, a <u>necessary</u> feature of the signal detection formalism.

(2) It clearly distinguishes between on the one hand the overall discriminating power of an information retrieval process (for a given partitioned data base, retrieval language, query, but not a fixed threshold value) e.g. through <u>E</u> and <u>A</u>; and on the other hand the realised or existing discriminatory power of it when the threshold

---

* The processes are, moreover 'social' rather than individualistic, in that they are designed to respond to the needs of large groups of persons with diverse information needs, over lengthy periods of time.

value is fixed, e.g. through the evaluation of R, F and P. These two attributes are clearly distinguished by the structure of the formalism.

(3) It relates to human judgements (relevance; query form; assigned attributes) in a clear manner, but at the same time presents an area of study capable of objective analysis. All the phenomena within the retrieval process are in principle observables: query, $f_1, f_2, 1(z)$, z_c, R, F, G, P, $\underline{E}$ in a testing situation. The hypotheses of invariance in f_1 and f_2, a major feature of the theory and describing general characteristics that underlie joint variation in human judgements of relevance and queries used to identify relevant documents, are accordingly falsifiable.

(4) The formalism appears to provide, through its conceptual simplicity, a fertile area for further studies of information retrieval. As a consequence of its simplicity, there is a "semantic pressure" leading to clarification of meaning in a field in which terms have perhaps been used all too loosely in the past: (e.g. 'weighting' as both document weighting and term weighting, 'query' as either linguistic statement, Boolean expression or set of document attributes; 'degree of relevance' as both an input-event or an output-event qualifier; 'retrieval system' as both the process as a whole or simply a combination of data base and retrieval software; 'signal to noise ratio' as pertaining to either the partitioned data base as a whole, or just to a retrieved set. Although the tendency of Swets's formalism to lead to clarification of the meaning of such terms now seems apparent, it is a reasonable criticism of Swets's presentations of it that on numerous points these were unclear.

The focus of the theory on the two distributions we have labelled f_1 and f_2 presents a firm foundation for hypothesis formulation and testing. As such the theory encourages a scientific development of information retrieval (rather than any more specialised philosophical or mathematical developments). These distributions can, for example, be hypothesised to be: invariant with respect to query (for given need), or invariant with respect to: query (as a set of terms of fixed size), choice of similarity measure, retrieval language, data base, level of exhaustivity of indexing, etc., for invariance delimited in some way. Some such matters can be examined in the formalism, but prima facie at least they are experimental questions. Swets himself carried out an analysis of experiments by other workers but his results are given graphically rather than numerically, are based on questionable experimental designs, and based (in two cases out of three) on averaged data in which, moreover, signal to noise ratio (G-value) is an uncontrolled variable.

His analyses may also be criticised as having been insufficiently concerned with Precision, but his concentration on the Fallout concept instead (i.e. on what is now referred to as Fallout) is consistent with the main hypotheses he sought (or appeared to seek) to establish. These were that (1) the ROC graphs for confounded data do not vary widely with retrieval method for a given data base; and (2) Normal probability density functions for f_1 and f_2 determine such curves. Swets's concern for Recall and Fallout may appear surprising, given the emphasis on Recall and Precision as the main probabilistic measure of retrieval effectiveness in more recent literature. The reasons for his preferring Recall and Fallout were

possibly as follows: (1) These enable the frequencies in all compartments of the 2X2 table to be reconstructed, which is not the case with Recall and Precision. (The accepted view now on this point appears to be that there is no need for this to be done: effectiveness should be directed at the transmission of relevant documents, not that of non-relevant documents: the fraction of non-relevant documents rejected is thus of no consequence. (See for example, Good, 1967)); (2) The invariance asserted to exist by one of the hypotheses is related to f_1 and f_2 and therefore to R and F); (3) R and F are directly relatable to the errors of statistical decision theory, through $\alpha \equiv F$ (Type I error), and $\beta \equiv 1-R$ (Type II error) respectively; (4) The signal detection formalism is primarily concerned with R and F. The relationship with P is implicit and depends on the value of G; and (5) The overwhelming emphasis in the literature on signal detection theory in psychophysics is on probabilities equivalent to R and F: and this can be accepted as a prototype literature (at least in the early stages) for analogous work in information retrieval.

A last ground for criticism can be seen to be Swets's neglect of the use of logical search expressions in information retrieval. These do not appear anywhere in the formalism as given by him. The usage of Boolean logic as part of the retrieval algorithm in practice is not obviously relatable to the essentially random variable approach that Swets put forward.

Swets's main contribution may in the end prove to be the priority he gave to relevance (signal) as against query (detection device) in portraying the information retrieval process. On the other hand, the weaknesses in his work, in the author's view, are

primarily: his neglect of the logic of retrieval; and his neglect of thedistinction that should in principle be made between formalism and hypothesis. Swets's papers have nonetheless provided an overview of a radically new formalism in information retrieval, rather than a detailed formal presentation de novo, and the perhaps inevitable weaknesses in them should be looked at constructively. Such weaknesses as there were, were basically weaknesses in presentation only; indeed in a sense they demonstrate the potential of the theory for opening up a new field of study.

Indicative summary.

Although Swets's concern was apparently to introduce and justify several novel measures of retrieval effectiveness, in so doing he introduced a major formalism describing the entire retrieval process. The formalism has three main strengths: (1) It accounts in a systematic way for the presence of error in the information retrieval process; (2) It makes a clear distinction between (a) the discriminating power of the process, for a given retrieval language, question, and number-assignment method, evidenced through the separation of two probability distributions, and (b) the bias in that process introduced by implementing a decision threshold at different levels; and (3) It offers a simple structure within which the essentially subjective notion of relevance can be accommodated, as well as objective, controllable entities such as question, similarity-measure and threshold. The most basic feature of the theory is the priority accorded to the partitioning of the data

base (i.e. the relevance judgements), over the question, the latter being introduced (implicitly) as a variable. In its presentations by Swets, the theory had numerous weakness and ambiguities, principally a failure to distinguish formalism from hypothesis, to make completely explicit certain concepts (e.g. "questions"), to state hypotheses clearly, and to incorporate into the formalism the conventional use of Boolean logic in information retrieval. The treatment of the Precision of the set of documents retrieved was inadequate. These and other points are the object of later discussion extending the formalism.

3.3 Relationship of the Formalism with other concepts in Information Retrieval; extensions and applications of the Formalism.

3.3.1 Relationship of the Formalism with other concepts in Information Retrieval.

3.3.1.1 The concept of Retrieval effectiveness.

The problem of defining and measuring the effectiveness of information retrieval is a large one. It was the starting point of Swets's own work, although the signal detection model could alternatively have arisen through analyses of other areas (e.g. of the interaction of questions and relevance judgements). The topic of effectiveness is now covered by a vast literature, representative facets of it being: economics of operation, speed of retrieval, accessibility of documents referenced by the data base, and an assembly of measures concerned with the degree of overlap between what we have termed 'signal' and what is identified by the retrieval process as 'signal' - represented by the sets of relevant and retrieved documents respectively. The balance that should be sought, in any given operational situation, between these different concepts is a complex managerial task, and we refer to King (1971), Lancaster (1968), Salton (1975a), and Vickery (1970) for introductory comment on the problem in its broadest aspects. On the specific problem with which we are concerned, the measurement of set overlap, the literature reviews by Bourne (1966), Keen (1971), Rees (1967b) and Robertson (1969) as well as Swets himself (1963), provide useful historical survey pegs. As mentioned in Section 3.2, Swets's introduction of the 2X2 table allows for convenient representations and comparisons of the various measures. Before reminding the reader of the measures

usually accepted at the present time, three general observations are made:

First, we recall that Cooper has questioned the validity of any concern with documents that are not retrieved (or "unexamined documents" as he terms them) (1973, 1976). This is on the ground that "unexamined documents are without utility (i.e. have zero utility) to the system user"(Cooper, 1973: 371). In the writer's view however this is simply a consequence of a definition of utility that ascribes zero utility to documents that are not retrieved. Instead of ascribing negative utility values to documents not retrieved and which the user would have benefitted from seeing, were they to have been retrieved, negative utility is reserved to designate documents that the user has inspected but rejected as useless. Opportunity cost, in other words, is not recognised in the argument. In recognition of this, the writer does not support Cooper's view, although his argument is clear and almost persuasive. Nonetheless, as mentioned in the conclusion of Section 3.2, Cooper's work is valuable in drawing attention to the retrieved set as the only "reality" of the 2X2 table in operational systems: to recognise opportunity cost in the way we have just mentioned certainly does not overcome the problem that under operational conditions (as distinct from experimental conditions) such utilities are unknowns.

Secondly, we note that a postulate that there are 'kinds of relevance' or 'degrees of relevance' is regarded by some, perhaps reasonably, as pointing to a weakness in the probabilistic measures, in that the latter do not distinguish between relevance of different characters. It seems that one can do little at this stage except say that (1) 'signals' may be represented by different sets of

relevant documents (and that in the case of stipulated 'degrees of relevance' such sets are, by definition, nested); (2) either ordinal or qualitative generalisations of the 2X2 table are suggested by such notions; and (3) redundancy in the 'information' carried by signals may (at the denoter-of-signal's behest) involve a smaller set of documents being defined as relevant than if redundancy were not recognised. This matter carries through into signal detection formalism in a simple way: sets of relevant documents, defined in different ways in respect of the same information need, engender different distributions f_1 and f_2 for a given query and method of assigning z-values.

A third observation is really a different elaboration of the view that the satisfactoriness of a set of retrieved documents, to a person with an information need, cannot simply be measured by the proportions that form the 2X2 table. Instead, it is argued, the cell frequencies as such need to be considered, as well as any ordering of the set of retrieved documents determined by the retrieval process. This view was also put forward by Cooper (1968), summaries of which are offered by Salton (1975a: 247) and van Rijsbergen (1979a: 160). It involves categorising the need for documents (as distinct from the need for information) in various ways. The unconventional categories involve the user specifying a need for one, n, or all relevant documents in the retrieved set. The retrieved set is assumed to be partitioned, with a simple ordering of the component subsets being defined by the process (i.e. a weak ordering of retrieved documents obtains). Cooper then defines "expected search length" (esl) as the expected number of non-relevant documents (say m) that need to be discarded, in a weak ordering of the retrieved

set, before reaching the figure of n relevant documents identified. In effect this entails looking at all permutations (i.e. sequences) of items in the last subset (i.e. the subset in which a tally of n relevant documents is reached), and calculating the expectation of m-m' , where m' is the number of non-relevant items discarded hitherto. The esl is then defined to be m'+E(m-m'), with E(...) operating on a discrete-uniform distribution over the set of sequences. Although a probabilistic concept, expectation, is involved, esl refers to a given instance of retrieval: hypotheses of invariance are not implicit in it. The notion of weak ordering is in fact a possible ingredient of the Swetsian formalism when an extension of it to describe a discrete receiver outcome space is made, but fuller discussion of this point is postponed until later sections (3.3.2.1-3.3.2.2).

To define the usual probabilistic measures of retrieval effectiveness, we first denote the sets of relevant/retrieved documents by A and B respectively, each being a subset of some data base S, and the number of items in a set W (say) by $\|W\|$. Then:

$$\text{Precision}, \quad P = \frac{\|A \cap B\|}{\|B\|}, \quad \|B\| \neq 0,$$

$$\text{Recall}, \quad R = \frac{\|A \cap B\|}{\|A\|}, \quad \|A\| \neq 0.$$

$$\text{Fallout}, \quad F = \frac{\|(S \setminus A) \cap B\|}{\|S \setminus A\|}, \quad \|S \setminus A\| \neq 0.$$

(The set $S \setminus W$ denotes the set complementary to W, in S.) The probability that a document in S is relevant to an information need is denoted by the Generality of the set of relevant documents:

$$\text{Generality}, \quad G = \frac{\|A\|}{\|S\|}$$

G is not a measure of effectiveness as such, but (from a signal-detection viewpoint) the value of the signal-to-noise ratio in the data base as a whole. It follows from the above definition that:

$$P = \frac{GR}{GR + (1-G)F}$$, R and F not both zero, i.e. $\|B\| \neq 0$.

The value of P for $\|B\| = 0$, of R for $\|A\| = 0$, and of F for $\|S \setminus A\| = 0$ can be defined arbitrarily, although the last of these is never met in practice.

A general measure that is a function of both R and P has been advanced by the writer (1973b) on the basis of a metric proposed by Marczewski et al (1958). In the above notation, this metric D, is defined as

$$D(A,B) = \frac{\|A \Delta B\|}{\|A \cup B\|} \quad , A \cup B \neq \emptyset,$$

where $A \Delta B$ denotes $(A \cup B) \setminus (A \cap B)$. D is one measure of the degree of similarity (nearness) of A and B, i.e. of the extent to which the sets of relevant and retrieved documents coincide. As such it is a special case of a more general function discovered by van Rijsbergen (1974) which we label E" hence, namely:

$$E'' = 1 - \left(\frac{1}{\alpha(1/P) + (1-\alpha)(1/R)} \right)$$

where $\alpha \in [0,1]$ in general, and has the value $\frac{1}{2}$ in its D form, and in the forms advanced earlier by Vickery (Cleverdon et al, 1966) and Jardine and van Rijsbergen (Jardine, 1971a). The general function is derived from fundamental considerations of measurement, namely the problem of finding a general function that maps the Cartesian product $[0,1] \times [0,1]$, representing the possible range of the Precision-Recall relationship, onto the "scale" set $[0,1]$, subject to six limiting conditions. The analytical form quoted above is arrived at after one further definition is made relating to the relative importance attached by the user to Precision as against Recall. The general function can in fact be applied to any two variables defined on $[0,1] \times [0,1]$, such as Recall and Fallout, as

remarked by van Rijsbergen, and can be further generalised.

Although there is a degree of arbitrariness in choosing $\alpha=\frac{1}{2}$, this corresponds to a user who values equally an increase in marginal Recall or Precision - a reasonably neutral position. In relating the general evaluation function to Swets's work we choose, also arbitrarily, the D form of the general function which has this α value. In this case we also note that: (1) (like all forms of the general measure) D is normalised so that its value falls in the range [0,1]. It takes the value 0 when A=B, unless $A=\emptyset=B$ when it is defined to have the value 0, and takes the value 1 when A and B are disjoint (when $A \Delta B = A \cup B$.) D needs for its calculation only three of the four cells of the 2X2 table; the tally of "non-relevant documents rejected" not being used. This is consistent with our earlier criticism of Cooper's 'utility-based' approach: whereas Cooper disregarded all non-retrieved documents, we are taking into account those non-retrieved documents that are relevant. (2) Like all forms of the general measure, D, in offering a composite assessment of both Recall and Precision effectiveness, provides a <u>single</u> criterion by which an optimum of a retrieval process can be identified. (Otherwise two separate optima would be identified.) Given that in the Swetsian formalism there is an underlying criterion, namely the threshold value z_c, determining the values of each measure of effectiveness, it would seem to be useful to have a means of identifying a <u>unique</u> optimum z_c value. (3) As a function of the basic Swetsian measures of effectiveness, namely Recall and Fallout, D may be written:

$$D(G,R,F) = \frac{F(1-G)+G(1-R)}{F(1-G)+G}$$

when the signal-to-noise ratio, G appears explicitly. Substituting

the expression quoted earlier for P in terms of G,R, and F, gives D as a function of R and P alone:

$$D(R,P) = \frac{R+P-2RP}{R+P-RP} \quad , \quad R \text{ and } P \text{ not both zero.}$$

Both expressions can be verified by substituting for R,F,G, and P the appropriate functions of the variable names for cell-frequencies in the 2X2 table. To emphasise the dependence of D, R,F and P (but not G) on the threshold criterion, z_c, we could write $D(z_c)$, $R(z_c)$,$F(z_c)$ and $P(z_c)$ for these names in the above identities.

In relating D to R,F,P and G through the Swetsian formalism, it should also be emphasised that the above relations both refer to one instance of retrieval process; i.e. they do not relate to data that have been 'averaged' (i.e. pooled or grouped). In general, the mean values for R and P, for a set of processes, cannot be substituted in D(R,P) to yield the mean value for D. The relations are valid for a specified value of z_c.

It might be asked what functional relationship connects R and P when the relevant sets and retrieved sets are a constant distance (as measured by D) apart. This function is given immediately by the expression for D(R,P):

$$P = \frac{R(1-D)}{R(2-D)+(D-1)} \quad .$$

Graphs of P vs R for various values of D are illustrated in Figure 3.3.1.1-1. The continuity implied by the graphs is not strictly correct, for any data base is finite and accordingly R and F can vary only discretely in practice.

The way in which D varies with z_c in the Swetsian formalism (through D=D (G,R,F)) has already been indicated. We note also

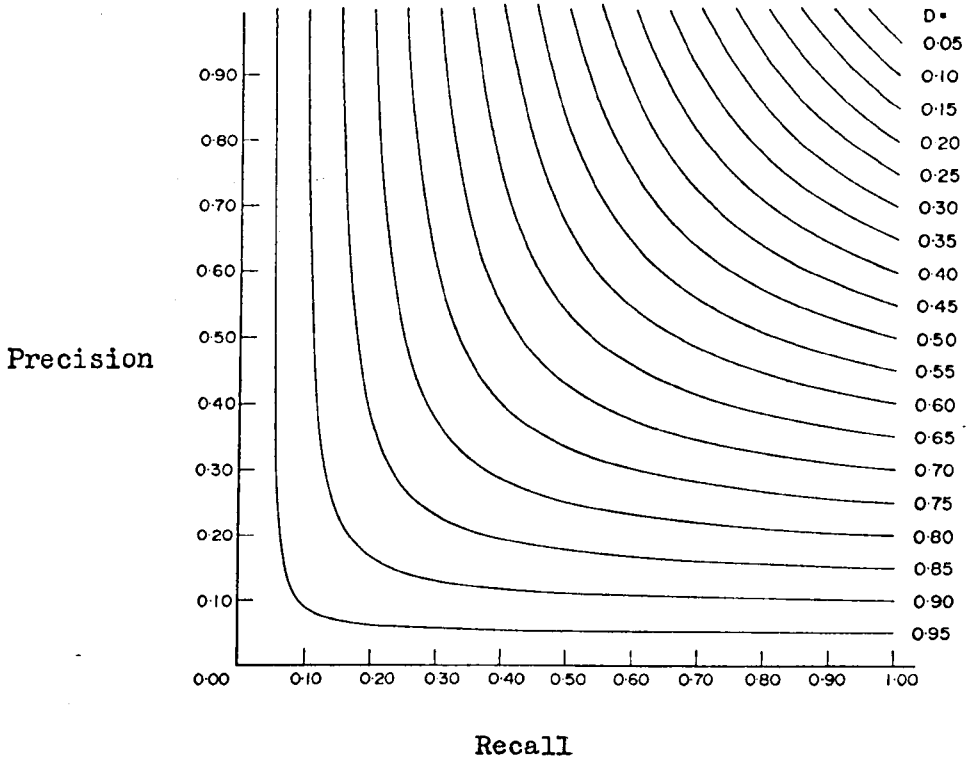


Fig. 3.3.1.1-1. Graphs of Precision vs Recall, for various values of D, a distance measure between the sets of relevant and retrieved documents. When D=1 the 'curve' becomes the axes P=0 and R=0, and when D=0 it becomes the single point (1,1).

that a sample of D values may in practice be defined by variation in (1) z_c, (2) the query chosen to express a given information need, (3) the measure of similarity chosen for a given need, or (4) the data base (for a set of relevant documents common to two or more data bases); for some specified information need. One might also use D as an indicator of retrieval effectiveness when there is variation in the retrieval process brought about by different sets of relevant documents pertaining to different needs, for all other variables held constant. Statistics based on such samples of D values can then be defined. Lower D values will point to instances of more effective retrieval, e.g. to more effective measures of similarity of document and question cet. par..

A fundamental theoretical problem based on the measures of retrieval effectiveness is the following. Given some assumed joint distribution of (1) Precision and Recall, or (2) Generality, Recall and Fallout, what is the distribution of the general measure of effectiveness? The solution to this general problem is relevant to the prediction of values of say D that will be observed for different assumed analytical distributions of P,R,F and G. (For different retrieval processes, G is a random variable, even though it is a constant for any given process. P,R and F are however random variables in any given process owing to their variation with z_c.) This problem was treated by the author (from a probabilistic, rather than a statistical point of view) (1973b: 195) with the following results:

(1) If the joint density function of the bivariate random vector (P,R) is $f(p,r)$ $(p,r \in [0,1])$, then the density function of the random variable D (with values $d \in [0,1]$), which we label $\phi(u)$,

is expressible as:

$$\phi(u) = \int_{1-u}^{1} f\left(\frac{v(1-u)}{u(1-v)+2v-1},\ v\right).\left(\frac{v^2}{(u(1-v)+2v-1)^2}\right)dv \qquad (1)$$

(2) If the joint density of the trivariate random vector (G,R,F) is $h(g,f,r)$, $(g,f,r \in [0,1])$ then the density function of D is $\phi(u)$, where

$$\phi(u) = \int_{w=1}^{1}\int_{v=0}^{1-u} h\left(\frac{w(1-u)}{u(1-w)-(1-v-w)},\ w, v\right).\left|\frac{u-w(v+w)}{(u(1-w)-(1-v-w))^2}\right| dv\ dw \qquad (2)$$

One simple analytical function pertaining to the first case above is that where Precision and Recall, are distributed uniformly and independently: i.e. one in which it might be said that the variation (of whatever experimental form) generating the (p,r) paired values is such that the retrieval process is random to the observer. The situation does not appear to have been treated previously in the signal detection or information retrieval literatures. The expected value of D under these conditions can be found as follows. We assume (P,R) to be uniformly distributed over $[0,1] \times [0,1]$, i.e. that $f(p,r)=1$, $p,r \in [0,1]$. Substitution in (1) gives the density function for D as:

$$\phi(u) = \int_{1-u}^{1} \frac{v^2}{(u(1-v)+2v-1)^2}\ dv$$

which simplifies to:

$$\phi(u) = \frac{2u}{(2-u)^2} - \frac{4(1-u)}{(2-u)^3}\log_e(1-u), \quad 0 \leq u \leq 1.$$

The meaning of $\phi(u)$ is that the probability that D lies in $(a,b]$ is given by:

$$\Pr(a < D \leq b \mid a,b \in [0,1]) = \int_a^b \phi(u)\ du.$$

The expectation of D in this situation is then obtained by evaluating

$$\int_0^1 u\,\phi(u)\,du.$$

An analytical integral does not exist, and numerical methods need to be used. These yield $E(D) \doteq 0.71$. A sketch of the density $\phi(u)$ is shown in Fig. 3.3.1.1-2, reflecting that $\lim_{u \to 0} \phi(u) = 0$ and $\lim_{u \to 1} \phi(u) = 2$. As can be seen, the bulk of the probability is centred on higher values of D. We thus have the intuitive picture that if Precision and Recall are distributed independently (with expected values of 0.50 in each case), the distance between the sets of relevant and retrieved documents will not be 0.50 as might be expected, but a higher value, namely 0.71.

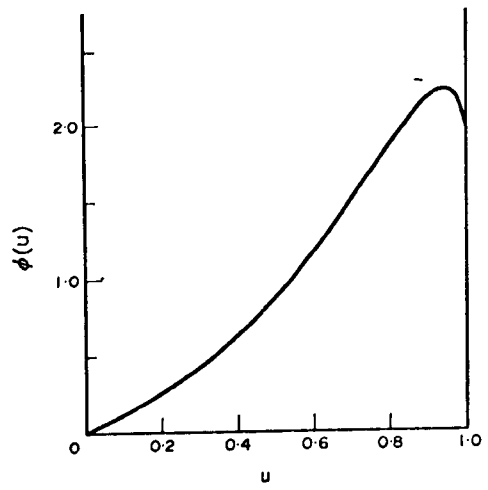


Fig. 3.3.1.1-2

There is scope here for investigating other densities of (P,R). In particular, as remarked by the author (1973b) it would be instructive to examine the consequences of assuming that (P,R) was distributed as a Normal bivariate density. It is however impossible to do other than speculate as to the likely form of the distribution of (P,R) at the present point given the uncertainties as to actual

forms of the distributions f_1 and f_2, and the uncertain degree of independence of the random variables R and F in practice.

A distinction in kind exists between (1) effectiveness measures of the probabilistic type (e.g. Recall, Fallout, Precision, Marczewski-Steinhaus metric), expressible in terms of the cell-frequencies of the 2X2 table and defined by a particular threshold value z_c, and (2) effectiveness measures describing the range of effectiveness in the latter sense, determined in a given retrieval process by variation in z_c. Type '(1)' measures are qualitatively different from type '(2)' measures in that they depend on extra information for their definition, namely the value of z_c. Swets's measures A and E are examples of the second type. A semantic distinction between these two types of measure of effectiveness will be used in the following text, to avoid ambiguity: 'probabilistic measure' will be used for the former type, and 'language measure' for the latter type. Only in the former case is a retrieved set defined. A modification to the language measure E has been suggested by Brookes (1968), namely:

$$S = \frac{\mu_1 - \mu_2}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}}$$

Geometrical interpretations of S and E are given by Brookes in terms of the ROC graph. In effect both measures are interpretable as distances, as shown in Figure 3.3.1.1-3 (based on ordinates used for Figs. 1-2). For distributions f_1 and f_2, both Normal and with variances σ_1^2 and σ_2^2, the value of E is $\sqrt{2}$ multiplied by the distance OI, and the value of S equals the distance ON. Whereas S specifies the straight line AB uniquely, if E is used an accompanying value for the slope of AB must also be given. (An exception is when

the slope is unity, when both E and S completely specify AB.)
Brookes also claims that

> "From a statistical point of view, this normalising factor [i.e. $(\sigma_1^2+\sigma_2^2)^{-\frac{1}{2}}$] is more acceptable than the arithmetic mean of σ_1 and σ_2 because its use simplifies the analysis of sampling variations and the testing of significant differences of the measure of effectiveness since the sampling distribution is known." (p.50)

That two parameters were needed to specify the ROC graph when $\sigma_1 \neq \sigma_2$ was commented upon by Swets in his second paper, (Swets, 1969: 76). (There are other comments on the Swetsian formalism offered by Brookes (e.g. that the measure of similarity between query and document is a continuous random variable), which we treat elsewhere.)

Robertson has proved that Swets's measure A is in fact equivalent to (varies monotonically with) the modified version of E put forward by Brookes (i.e. S), so that if S is accepted as a meaningful measure, A is redundant. (Robertson, 1969). The writer has pointed out that the relationship between A and S established by Robertson may be written:

$$A = \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{-S}{\sqrt{2}}\right)$$

where erfc(x) denotes the definite integral: $\frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2)\,dt$.

A further criticism of A, noted by the author, relates to Swets's assertion that:

> "...the value of A is equal to the percentage of correct choices a system will make when attempting to select from a pair of items, one drawn at random from the irrelevant set and one drawn at random from the relevant set, the item that is relevant." (Swets, 1969: 77)

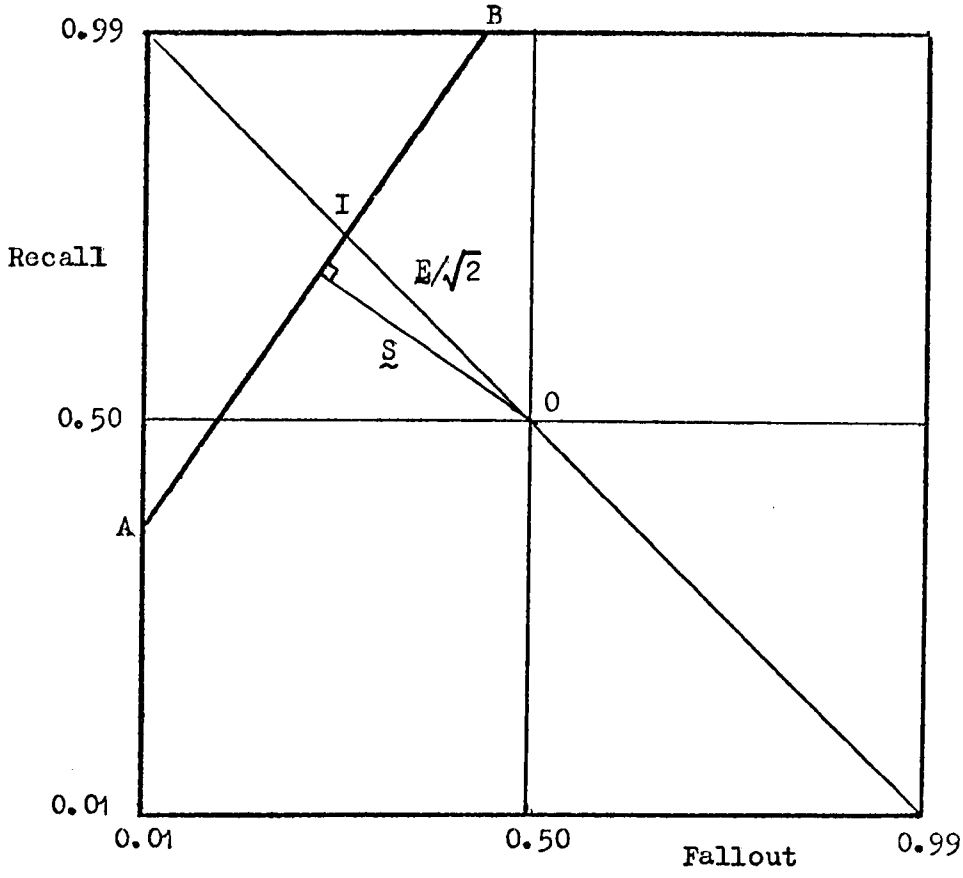


Fig. 3.3.1.1-3 (After Brookes (1968: fig. 7).) AB is a ROC graph, plotted as for Figure 2, in which both f_1 and f_2 are Normal, but where the variances differ. Swets's measure $\underline{E}$ measures the distance OI (apart from a scaling factor of $\sqrt{2}$; i.e. $\underline{E} = \sqrt{2}.OI$), and accordingly to specify AB uniquely the slope of the line needs to be given as well as the value of $\underline{E}$. Brookes's suggestion was that an alternative measure ($\underline{S}$) should be used, defined as indicated in the text and equal to the distance ON, and sufficient to specify AB uniquely.

This assertion is not proved, but appears to refer to a standard result in psychophysics, quoted by Egan as:

> "The area under a proper ROC ... equals the probability of a correct decision in the 'two-interval, forced-choice task'." (Egan, 1975: 46, citing Green, 1966, and Swets, 1964)

But Swets's statement is meaningless since a system "makes no choice at all" (to carry over the anthropomorphism) unless a threshold value on z or l(z) is given. Presumably a careful wording of an exact statement would bring in not only the threshold, or set of thresholds, but also both G and the matter of whether the ROC graph is proper or not. However, in view of (1) the equivalence of A and S noted by Robertson, and (2) the rather artificial notion of an information retrieval system examining all possible pairs of documents, one relevant the other non-relevant, the matter seems hardly worth pursuit.

It has also been pointed out (independently by Harter (1975), and the writer (1975)) that S^2 is identical with the measure G used as the basis of Fisher's "linear discriminant analysis" technique. (Fisher, 1936) There, G provides a measure of separation of the populations of individuals when the individuals are characterised by values of a set of describing variables. This coincidence will prove useful at a later stage in applying Fisher's technique to several information retrieval problems.

The measures E and S certainly do not exhaust the possibilities for measuring the separation of f_1 and f_2. Becker (1968) for example has identified seven different measures, for example $(\mu_1-\mu_2)/(\sigma_1+\sigma_2)$, or

$$\int_{-\infty}^{\infty} (f_1(z)-f_2(z)).\log_e(f_1(z)/f_2(z))\ dz$$

some of which are similar to those described by Mathai et al (1975).

The whole area of retrieval system effectiveness is, like that of document weighting, a fairly active one at present, recent papers that offer significant new departures having been offered by Radecki (1976a) and Guazzo (1977), for example.

It seems plausible however that the fundamental notion of 'a set of relevant documents', the cornerstone of the Swetsian formalism, is a robust and useful one. The probabilistic measures of effectiveness, to which the formalism actually relates, will equally plausibly continue to be used. This, in the author's view, gives the formalism at least a prima facie appropriateness to information retrieval practice. More user-oriented experiments are obviously required however to discern what are the basic properties in documents that users of information systems require, i.e. some further characterisation of, or taxonomy of "signal" needs to be sought. At least as far as 'redundancy' within the set of relevant documents is concerned, there is some evidence (Cleverdon et al, 1976) that this is low, again strengthening the concept.

3.3.1.2 Query and information need.

The main questions we shall discuss here are (1) whether 'information need' is somehow more fundamental than 'question' (already introduced in Section 3.2), and (2) whether Swets was consistent in his description of the 'question'. The discussion is related to published literature, and is based partly on previous comment of the author (1977b).

Swets's basic position, as evidenced in the formalism he advanced, rather than in either the accompanying discussion of it, or his testing of hypotheses expressed in the formalism against experimental data, was that information need is 'prior' to question. The evidence for this is summarised as follows:

(1) The formalism itself acknowledges the partitioned database as the fundamental entity. A question, as a description in language of an information need, is seen as a variable entity for a given need. Not only is it secondary or less fundamental than need on this ground; a question does not actually require to be expressed at all. (A question representing an information need has to be formed only when an individual wishes to communicate his need to a third party: e.g. a computing machine or another person searching on his behalf). For an information retrieval process to be defined (and for a machine to implement that process) a question does require to be defined of course, and as such it forms one essential input to the 'receiving apparatus' that the formalism describes.

(2) It was frequently acknowledged by Swets (e.g. Swets, 1963: 248) that questions will in part determine the effectiveness of the retrieval process, since they in part define the

process. In looser terms, queries can be more-or-less effective. Swets was also clearly aware of the 'heuristic' approach to retrieval (e.g. Swets, 1969: 89), the essence of which is successive improvement in queries for a given target set of relevant documents.

(3) Experimental data was analysed by Swets that was based on sets of relevant documents*.

It appears moreover that despite inconsistent usage what Swets intended by 'question'(or query) was a set of document attributes. Such a specific definition was not given explicitly by Swets, but it is implied by, for example, his interpretation of Salton's cosine measure and Cleverdon's level-of-coordination measure of similarity of question and document record. Such a definition of question as a'set form query' can be usefully abbreviated to SFQ. As emphasised in Section 3.2, the term 'query form' was used ambiguously by Swets, but it appears that (1) this was seen as a separate concept to SFQ, and (2) it was a synonym for a logical, i.e. Boolean expression. (Swets, 1963: 248) We shall at times refer to the latter, i.e. to a question as a set of attributes linked by Boolean operators into a Boolean expression, as a 'Boolean form query' or BFQ.

Since the notion of 'need' being prior to 'query' (in whatever form) represents a fairly radical thesis in information science, we briefly elaborate on it. The argument is also needed to support the claim that Swets's own analyses were invalid. This will be followed by an indication of the type of experimental design that is needed in order to generate or test hypotheses expressed in the Swetsian formalism in a legitimate way.

---

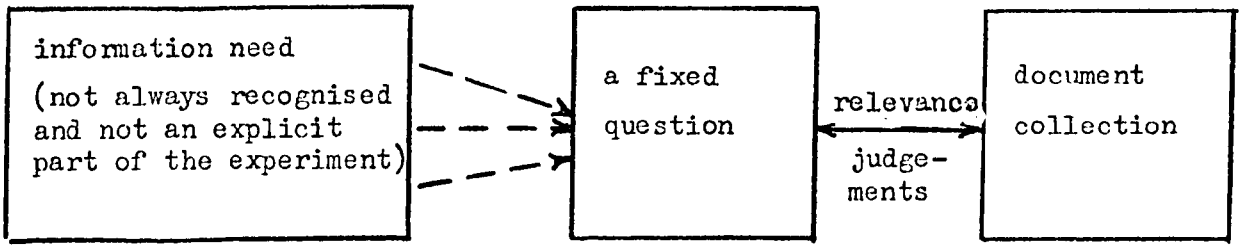
* The author maintains that these sets were inadequately defined.

The notion of 'information need' is the fundamental concept in the study of information retrieval, its raison d'etre. The 'question' cannot be so since for any given need its form and content (i.e. the choice of particular logical operators and document attributes) are both variable. This view, although implicit in Swets's 1963 paper, has apparently been widely adopted in only one area of research in information retrieval: that of the study of the heuristics of retrieval. For synoptic discussion of this research the reader is referred to Salton (1971a: chaps. 10-13; 54-5), and van Rijsbergen (1979a:105), although both writers, as do almost all workers in the information retrieval area, confusingly use the phrase "relevance to a question". The latter phrase involves a contradiction in concepts, from the Swetsian point of view. The phrase 'question Generality', for the ratio $\|A\| / \|S\|$ is also misleading. This should, more appropriately, be referred to as 'Generality of the set of relevant documents' or 'Generality of the need, as evidenced in the data-base', as we anticipated in the last section. In a different context, Taylor (1968) also takes up this viewpoint, as do various writers whose work is reviewed by Rees et al (1967a) and Saracevic (1970b) but it is a reasonable generalisation that a large majority of information retrieval workers, as well as laymen, see the relevance of documents as directed at a verbal artefact, i.e. the question, whether in SFQ, BFQ or simply as a sentence or statement in everyday language. Examples of theoretical papers or monographs in which 'relevance to a question' is introduced as a (pseudo) concept are readily found (e.g. Maron et al (1960), Goffman (1964a, 1964b), Sparck Jones (1971), Paice (1977), Ludwig (1975) and the author's own earlier work (Heine 1974), though not in the later

contribution (1975).) Although Saracevic's major review distinguishes between these notions (i.e. between our own point of view and the notion of 'relevance to a query') and brings out the notion of a query as a variable (p.127), he fails to underline the consequential weakness of the classical experiments.

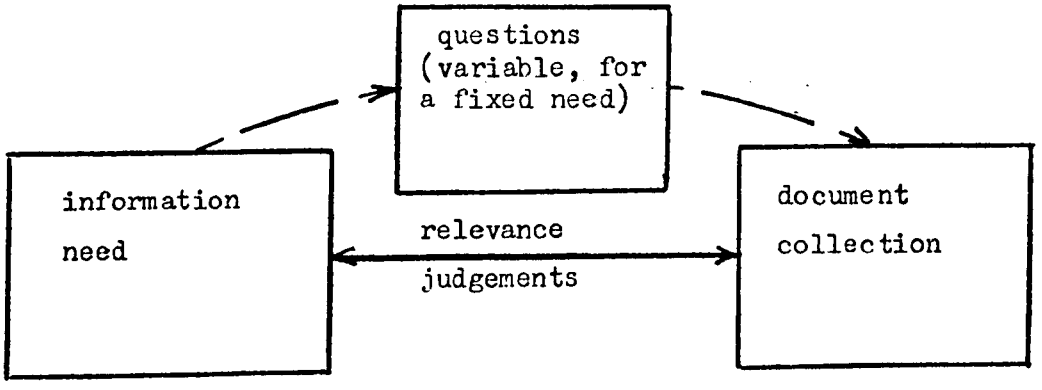
The author emphasises that the point made here is not just a semantic one, although in one or all of the purely theoretical papers it may be the case that the term 'question' can simply be relabelled as 'need' without destroying the particular arguments concerned. The evidence for this is in the literature on experimentation in the area. The now-classic Cranfield and Aberystwyth experiments (see e.g. Cleverdon et al (1966), Cleverdon (1967), Keen et al (1972), Keen (1973)), and numerous others, based partly or wholely on data from them (e.g. Sparck Jones (1971), Robertson (1975), Barhydt (1967), Saracevic (1966) and Ludwig et al (1975), all involved experimental designs in which relevance judgements were made against verbal artefacts describing real or hypothetical information needs, not in reference to subjectively-experienced information needs. As such, the arbiters necessarily needed to assume or imagine what the context of each verbal artefact really was, i.e. what the information need in fact was. In other words the experimental approach that was used involved an artificial situation in which the arbiter of relevance either (1) did not know the need to which the question related, thereby enforcing his giving an interpretation to it, or (2) was aware of both the original need and the question given by him as an expression of that need, introducing ambiguous terms of reference for the relevance judgement. The information needs as such were not

incorporated into the experiments explicitly, or otherwise convincingly, and accordingly the questions had the nature of arbitrary articulations. In diagrammatic form, in the classic laboratory-style experiments we had the following scheme for identifying 'relevant' documents:



The classical experiment

The author suggests a more experimentally-sound evaluative situation would be one in which the relevance judgement, not the question, was treated as the most fundamental entity. The question (as say a statement in English) then appears as an adjunct to the situation. As such, it may be an SFQ or a BFQ and of whatever constitution as may be required, and generated by a variety of methods. This situation is illustrated by the scheme:



The Swetsian experiment

The author's argument is therefore that the correct fundamental entity that one should seek to describe in system evaluation experiments is the relevance judgement, not the question that is variably related to it.

It is emphasised that in criticising previous experimental work in the above way, we are detracting only a little from their very significant contributions. That decisions should have been made to base experiments on the pseudo-concept of 'relevance to a question' is moreover readily understood given that in practice users often approach documents through their attributes (e.g. through a card file, or through a post-coordinate term system) and of course need to formulate a question in order to do so. An experimental situation need not and should not follow this path, however, and must recognise the essential variability in question type and substance.

Having made the above criticism on the basis of Swets's formalism, we are now faced with a surprising fact: that Swets's own attempts at testing hypotheses expressed in his formalism involved experimental data that were incompatible with it: i.e. involved data based on unsatisfactory partitionings of data-bases by relevance decisions. The consequence is that his analyses are of completely unknown validity.

To remedy the weakness in experimental design discussed above is, at least in principle, a simple matter. It is to define sets of relevant documents in assertional terms (i.e. to have users say 'this is relevant', 'this is not relevant', etc.), or to look for such sets through behavioural evidence of some type. Questions directed at retrieving those documents, when dispersed in a data-

base, can then be formed in various ways: algorithmically, or chosen by arbiters. The choice of algorithm is a vast topic, some possibilities being: questions in BFQ based on Boolean minimisation of the sets of attributes attaching to relevant documents (following Quine (1959) and noting Benwell (1974)), questions as SFQ based on the clustering of document attributes within the relevant set, or on the clustering of documents themselves within the relevant set (the most deeply-clustered document(s) yielding the attributes), or simply on the relative frequencies of attributes in the set of relevant documents and the complementary set. Part of the experimental research to be reported in this dissertation follows this rationale.

Lastly, we attempt to clarify the point made by Swets that queries of different "breadth" can affect the retrieval process. What Swets may have meant here is that ROC graphs with different characteristics may be generated by questions as SFQ composed of different attributes but having the same number of attributes, and with the attributes varying in respect of their frequencies of assignment in the data base as a whole. (That ROC graphs so generated are systematically different in shape does not apparently follow from the formalism however: We are merely conjecturing that this usage may have been behind Swets's use of the phrase "question breadth".) Again, he may have simply meant variation in the number of attributes making up a query as an SFQ.

To summarise: (1) The Swetsian formalism gives priority to information need, as evidenced as a set of relevant documents, over any question or set of questions proposed in order to identify that set among a larger set. A question is, in the formalism, a variable

input to the 'receiving apparatus' element of the retrieval process. (2) The data analysed by Swets failed to observe this distinction: all three sets of data were based on an experimental procedure in which the 'relevance' of documents was judged in reference to verbal articulations. Accordingly the hypotheses implied by Swets in his formalism remain untested (pace the lack of explicit form of those hypotheses). (3) Questions can be expressed in set form or Boolean form. (The two are not equivalent although the specification of a threshold value and weak ordering function in addition, will secure equivalence between a pair of queries in these different forms. This point, not considered by Swets, will be clarified in a later section (Section 3.3.3.1).) (4) A satisfactory testing of hypotheses expressed in the Swetsian formalism would entail relevant sets being defined in 'assertional form' in some way, i.e. by an individual marking documents as relevant or not relevant in reference to some information need known to him, but not in reference to an arbitrary description of need in language, i.e. not in reference to 'a question'. The formalism itself does not anticipate either the way in which such assertion could be made in an experiment, or the way in which questions should be chosen in reference to such assertions.

3.3.1.3 The concept of clustering

'Clustering' is a tendency for the members of a set to be associated in groups. The association can be expressed in the form of a partitioning of the set, as a hierarchy of nested subsets, in terms of densities of members in a metric space of similarity values (between set members), or in terms of overlapping groups ('clumps'). (see Cormack (1971), Everitt (1974), or Jardine and Sibson (1971b), for example) The subject is now a large one (see for example the review by Cormack and the general theory by Lance and Williams (1967)) and has been variously applied in information retrieval by, for example, Jardine and van Rijsbergen (1971a), Lunn (1957), Oddy (1974), Salton et al (1975b), Sparck Jones and van Rijsbergen (1973). Review literature in this area of application is cited by van Rijsbergen (1979a: 47). Our concern in this section is solely with the relationship of the Swetsian formalism to the clustering notion although, as remarked in the preceding section, a clustering of the attributes of relevant documents may provide for the algorithmic generation of queries in an experiment to test hypotheses in the formalism.

The essence of clustering is probabilistic dependence between random variables. If a set of individuals, S, is mapped by two random variables of Bernoulli type, X and Y say, to events $\{0\}$ and $\{1\}$, then if X and Y are dependent, $Cov(X,Y) \neq 0$. Each individual will be associated with just one vector of values: $(0,0),(0,1),(1,0)$ or $(1,1)$. Accordingly X and Y together partition S into four subsets. The number of individuals in each subset will depend on the way in which X and Y covary, and in that sense their covariance determines the clustering of this type. (To discuss clustering of

the hierarchical classification type would require definitions of distances of individuals to individuals, individuals to clusters, and clusters to clusters. As these distances do not appear in the Swetsian formalism, although the notion of a partitioned data base does, we do not discuss them further here.)

The link with the Swetsian formalism is, in the author's view, through the receiving apparatus, i.e. the combination of query and mapping function. Consider a question as an SFQ, e.g. $Q = \{t_a, t_b, \ldots, t_n\}$. This will determine a vector of values for any given document record, according to whether or not each attribute is present in the record, e.g. $(0,1,\ldots,0)$. If we denote this vector by V, then the mapping function of the signal detection process will map the pair (Q,V) to some z-value, not necessarily using only the vector of values to do this. (For example, the function may use information on the frequency with which each attribute is used in the data base, as discussed later (Section 3.3.1.4).) In effect then, the mapping function Z partitions the data base according to subsets of S defined by $Z^{-1}(\{z\})$, where $z \in Z(S)$. In this particular sense, clustering is just a synonym for a probability function on S. This probability function is <u>implied</u> by the probability distribution <u>induced</u> by Z on Re. (This matter will be put more formally in Section 3.3.2, which this section partly anticipates.) In view of the mechanism underlying Z, i.e. the action of the receiver (qua inputted query (as SFQ) and similarity measure), this type of clustering is a joint effect of these two inputs. Clustering of the relevant set, A, and its complement, $S \setminus A$, through functions Z_A and $Z_{S\setminus A}$ defined again by the query as SFQ and similarity measure, with these sets as domains,

are likewise effected. The clustering, i.e. partitioning, of these sets is reflected in the induced distributions $f_1(z)$ and $f_2(z)$ of the formalism.

Lastly, we look briefly at a concept known as the "cluster hypothesis" Jardine and van Rijsbergen (1971a), van Rijsbergen and Sibson (1973). This has been stated as "closely associated documents tend to be relevant to the same requests". (We ignore here the objection conveyed by the argument of Section 3.3.1.2, that relevance should be judged vis-a-vis need, not query.) In effect this hypothesis involves for its exact statement the comparison of two probability distributions induced by a measure of similarity between two documents, for (1) all pairs of relevant documents, and (2) all pairs of documents one of which is relevant. (In practice, in obtaining approximations to these distributions, not all possible pairs may be examined.) That these two distributions are separated, rather in the manner in which the distributions f_1 and f_2 are separated in the Swetsian formalism, constitutes the hypothesis. However the effect described differs from the Swetsian position in two fundamental ways. First, it is based on comparisons between document records only: no query is introduced into the discussion as it is with information retrieval. (The distributions are solely a consequence of partitioning the data-base.) Secondly, the distributions, although consequences of a partitioning, do not each relate to one of the subsets of the data-base so defined. A more symmetrically defined hypothesis would describe the distributions induced by pairs of documents taken from each such subset, i.e. pairs of relevant documents, and pairs of non-relevant documents.

The natural development of studies of clustering of document

attributes, when clustering is seen as the partitioning of sets of documents by Bernoulli variables, is the study of dependencies between the random variables mapping documents to attributes. The documents concerned can be in the set of all documents, or be solely relevant or solely non-relevant documents. This matter has been carried a considerable distance by van Rijsbergen (1977), who has examined it from the point of view of optimum document weighting functions. (These are functions that take term dependencies into account, unlike the usual ones based on assumptions of attribute independence, van Rijsbergen's concern being to select the best analytical form for such functions and to estimate the parameters of such functions from sample data.) The matter is also treated in this thesis in a simple way in Section 3.3.3.3, where a novel linear weighting function incorporating information on dependencies is introduced. It is possible that the main contribution of hierarchical clustering notions to information retrieval in the future will be to the question of optimum data base organisation for the manipulation of records, rather than to the logic of retrieval.

3.3.1.4 The concept of document weighting.

As discussed in Section 3.2, the Swetsian formalism postulates the assignment of a numerical value to each document in a data base, prior to a decision on procedure being implemented. This value is determined by a function of (1) the attributes assigned to the document by the indexer, (2) the attributes taken to define the query as SFQ(i.e. in set form), and possibly also (3) attributes of attributes. Examples of the latter are (a) so-called "term (attribute) specificity": the probability that a document has been assigned the attribute, (b) the probability that a relevant document has been assigned an attribute (to be subjectively 'estimated' by the enquirer), or (c) information on the co-occurrence of pairs of attributes, in the data-base or (as an estimate) in the set of relevant documents, as discussed in the last section.

The usual names given to this procedure, outside the Swetsian formalism, are "ranking algorithm", since the assignment of values to documents imposes a partial order on the collection, or "weighting process", since the value to which each document is mapped may be viewed as a "weight" attaching to that document. The literature on systematic ranking/weighting is now extensive, systematic reviews having been contributed by Evans (1973) and Sager et al (1976). The ranking process also features strongly in Salton's work (e.g. Salton, 1968, 1975a).

In the writer's view, the advantage of the Swetsian formalism here is that it focusses clearly on the notion of document weight (Heine, 1973a, 1974), this concept being central to Swets's formalism. The notion of "term weight" (more generally "attribute weight") which enjoyed some popularity in the late 1960s (see, e.g. Matthews

et al, 1967; Sommar et al, 1969) is seen in the formalism as a secondary one. This is so whether the term weight is (1) assigned by the indexer (reflecting an idea of the importance of the term in denoting the subject of the document), (2) assigned by the enquirer to each of the terms making up the query (again to reflect the importance of the term in the enquirer's perception of the subject of interest), or (3) some function of both. The literature on term weights has been effectively reviewed by Salton and Wang (1973), and Sparck Jones (1973). The unsatisfactoriness of the notion as it has been treated in the literature is, the writer asserts, apparent in: (1) The failure to identify, and formalise the description of, a communication channel between indexer and enquirer, evidenced in the separate specifications of 'subject' notions by both indexer and enquirer, a weakness which the Swetsian formalism overcomes through its explicit description of that channel; (2) Its conceptual 'disregard' for the problem of how documents should be weighted for a given set of term weights (almost all authors implying that a simple sum of term weights will define the document weight); (3) Its being (further to the latter point) 'one removed' as a concept from the matter of effectively ranking documents prior to identifying a 'signal' subset of them; (4) The confusion in the literature between (a) the use of term weights to simulate the action of Boolean expressions, and (b) use directed at achieving more effective retrieved sets through the use of document weights (as sums of term weights) as a means of ranking (the latter being clearly evident in for example Matthews (1971)); and (5) Ambiguity as to whether term weights should be assigned purely subjectively, or should be objectively based on

variables describing term specificity etc. The question of a threshold value for retrieving documents weighted by a function (usually sum) of term weights has been largely ignored by writers on term weighting. Matthews and Thomson even make the claim that "a minimum score is used to eliminate irrelevant answers"(!) (Matthews et al, 1967: 51).

The most basic criticism of the early work on weighting is however that it failed to consider the partitionings of data bases by instances of need. The notion that weights might be determined in part by the relevant documents to be retrieved was largely disregarded in favour of a 'subject' oriented thinking. This view, that weighting can usefully be studied in isolation from relevance judgements, is for example implied in the "Shannonian" approaches to optimising the weighting function used (e.g. Zunde et al, 1967; Brookes, 1972).

However, the current work on the number-assignment aspect (rather than on the matter of choosing the most effective terms) of the optimal ranking problem, seems to centre on the incorporation in the document weight of variables reflecting the probabilities of assignment of terms in the various sets of documents (the data-base, the relevant set, and the latter's complement), or estimates of these. Such approaches are more in sympathy with the Swetsian concentration on sets of relevant documents. Ad hoc formulae of this nature, of a variety of types, were apparently first introduced by Barkla (1969). Miller subsequently (and independently) deduced from Shannon theory a document weighting expression based jointly on (1) the specificity of each query term in the data-base, and (2) an estimate by the enquirer of the specificity of each term in

the set of relevant documents, and involving a logarithmic function of the two variables. (Miller, 1971) Sparck Jones, again working independently, subsequently introduced an expression that was in fact equivalent to Miller's but with the latter variable not present (Sparck Jones, 1972). The analysis and comparison of the closely-similar weighting functions defined by these expressions has been discussed in detail by Robertson (1974) and Sparck Jones (1975). Refinement of this work has also been offered by these authors Robertson et al (1976), and as mentioned in the previous section a significant departure in the area has recently been offered taking dependencies between random variables involved (so-called "term-dependence") into account (van Rijsbergen, 1977). Another recent departure has been provided by Salton, Yang, and Yu (Salton et al, 1975c; Yu et al, 1977). Called "term discrimination analysis" this involves assigning to each term in the query (in set form, as usual) a weight equal to the product of (1) a change in the density of documents in the space defined by (a) the attributes of documents, and (b) a measure of the similarity of documents, with (2) a value expressing the frequency of occurrence of the term in the text of the document. However, in that document texts are not usually included in data-bases (although abstracts increasingly are) this method may have limited practical application.

Formal definitions of the function defined by Miller, the cosine function of Salton, and related functions, are as follows*. We denote the set of terms attached to a sample document, d, by T_d, a sample term by t, and frequencies of t in the data base, S, and the set of relevant documents, A, by $u_S(t)$ and $u_A(t)$ respectively.

* We refer to 'functions' here since each analytical expression will, of course, determine a mapping.

The set of terms common to query and document is thus $Q \cap T_d$. Then:

I. The co-ordination level function of Cleverdon is:

$$z(Q \cap T_d) = \| Q \cap T_d \| .$$

II. Salton's cosine value is:

$$z(Q, T_d) = \|Q \cap T_d\| \, / (\| Q \| \cdot \| T_d \|)^{\frac{1}{2}}.$$

III. The logarithmic value of Miller and Sparck Jones is:

$$z(Q \cap T_d, \{ w_i \} , \{ s_i \}) = \begin{cases} \sum_{t_i \in Q \cap T_d} \log_b (w_i / s_i); & Q \cap T_d \neq \emptyset \\ 0, & Q \cap T_d = \emptyset, \end{cases}$$

where i labels the terms in $Q \cap T_d$, $w_i = u_A(t_i) / \| A \|$, $s_i = u_S(t_i) / \| S \|$, and b is any base. Here w_i is the probability that a relevant document will be assigned t_i; and s_i the probability that a document will be assigned t_i, the "specificity" of t_i. In fact Miller's work entails the enquirer subjectively <u>estimating</u> w_i. When w_i = constant (so that each query term appears with equal probability in the set A) we have Sparck Jones's formula as a special case. As expressed by Robertson (1972) this is:

$$z = - \sum_{t_i} \log_b (s_i).$$

IV. An amended form of the logarithmic formula, suggested by Robertson, is:

$$z(Q \cap T_d, \{ w_i \} , \{ v_i \}) = \sum_{t_i \in Q \cap T_d} \log_b (w_i / v_i); \quad Q \cap T_d \neq \emptyset,$$

where $v_i = (u_S(t_i)-u_A(t_i))/(\|S \setminus A\|)$, and which is not in general equal to s_i-w_i. The writer notes that whether this formula will determine z values that are significantly different from those of Miller's function will depend on (1) whether the Generality of the relevant set is small, i.e. whether $\|A\|/\|S\|$ is small, in which case $\|S \setminus A\| \doteq \|S\|$; and (2) whether the term concerned is sufficiently common in the data base that $u_S(t_i) \gg \|A\|$, in which case $u_S(t_i)-u_A(t_i) \doteq u_S(t_i)$. If both the latter conditions are met, $v_i \doteq s_i$ and so the earlier function is (approximately) restored.

If the earlier function and Robertson's function do produce z-values that are approximately the same, then it is likely that the Precision vs Recall graphs determined by each function are precisely identical, since the rank order of the documents may then be unaffected. This intuitive idea is made more rigorous in Section 3.3.2.4).

Lastly, we draw attention again, following Swets (1963) and Bookstein (1974, 1977) to the optimality of the likelihood-ratio weighting function over all other weighting functions. There are two fundamental points here. First, this function is only defined a posteriori. When the data base is partitioned in some way, by a query as SFQ, plus possibly other set-operations, then a likelihood-ratio attaches to each of the subsets, since each of the subsets intersects with a set of relevant documents and its complement. The subsets can accordingly be ordered by these likelihood-ratio values. But this leaves open the problem of identifying an analytical function, with operands restricted to, say $Q \cap T_d$, $\{w_i\}$, $\{s_i\}$ and $\{v_i\}$, which will also give such an ordering. Secondly, the use of the likelihood-ratio function as a weighting function defines

only local optima. This is true in the sense that this function is influenced by both (1) the actual identity, and number of the terms used to define the query, and (2) the further set operations that may be defined on the members of $Q \cap T_d$.

The action of weighting functions will be described more fully when Swets's formalism is formally extended to include the concept of a discrete outcome space (Section 3.3.2.1).

In summary, we observe that the notion of "document weighting" is just one component of the weight. The formalism does not treat weighting as an isolated process but as just one component of a retrieval process, along with query formation on the one hand, and the partitioning of the data base by relevance judgements on the other hand. Moreover the formalism incorporates the notion into a framework of evaluation, and in particular demonstrates the trade-off between R and F that will obtain by varying the threshold, a point almost totally obscured in the traditional literature on weighting.

3.3.2 Extensions of the Formalism.

The last comparable block, Section 3.3.1, related Swets's theory to some recent work in information retrieval at large. In this block we attempt to extend the theory, partly in order to remedy certain weaknesses that have so far come to light, and partly to provide a more robust theoretical framework appropriate to modern retrieval technology. The approach is based in part on previous discussion by the author (1973a, 1974, 1975). Other work of known relevance is cited at the appropriate place in the text.

3.3.2.1 A discrete, ordered receiver outcome space.

Despite the fact that the signal-detection formalism, as introduced by Swets, involves continuous random variables, which we labelled z_1 and z_2, it is clear that the outcome space of the SFQ and similarity measure - the apparatus for attaching values or weights to documents - is not continuous. The number of realisable values of $Q \cap T_d$ will be finite (and of the order of $\|2^Q\|$ for many similarity measures), and even for a similarity measure that produced a different value for every distinct attribute attached to documents, the outcome space will still be finite since the database is finite. This paradox attracted early criticism of the original formalism, and perhaps has been one reason for the slowness of its acceptance. One writer, for example, has written:

> "...The postulated value of z, as some continuous standard of relevance, cannot be matched in practice.... Brookes's suggestion that they might happen to be integer values arising in what is really a continuous variable looks like very special pleading." (Farradane, 1974: 207)

The comment of Brookes referred to was:

> "...this inference [that R and F values lie on a straight line when converted to standard Normal scores] requires the variable z to be continuous. But in the Cranfield tests the mediating variable was the 'level of co-ordination', a discrete variable which takes only the integral values 0,1,2... Can the continuous variable of the gaussian distributions be identified with the discrete 'level of co-ordination'? Swets does not mention this difficulty. However, for the present analysis, it suffices

> to imagine that underlying the discrete variable 'level of co-ordination' there is a continuous variable, z, which conveniently assumes the value 1.00..., 2.00..., 3.00..., and so on, as the level of co-ordination takes the values 1,2,3,... This point can await clarification if the implications of the Swets measure require it." (Brookes, 1968: 46)

In fact Brookes's Figure 6 has two minor errors in it reflecting the paradox: the y-axis is labelled "probability density" instead of "probability", and Normal density functions are shown as <u>envelopes</u> of the discrete distributions of probability on the values of the level of co-ordination measure, which is incorrect. Brookes's Figure 5, on the other hand, shows the role of the 'Normal approximation' clearly. The paradox is however simply resolved by choosing to regard the continuous Normal densities of the original formalism, $f_1(z)$ and $f_2(z)$, as <u>modelling</u> distributions the purpose of which is to yield definite integrals serving as co-ordinates of the ROC graph. The latter, so obtained, is then a continuous line, but it is such that the discrete ROC graph data obtained in practice lies on or near that line. In other words the notion of continuity can be seen as having been introduced simply for ease in computation. This is of course a perfectly reasonable and conventional practice: almost the entire body of classical science and engineering is built on continuous functions which cannot be justified in microscopic (quantum-mechanical) terms. To defend the usage of continuous models in this way is however not to claim that there is any pair of such models that give accurate approximations to the discrete functions $f_i(z)$ in practice.

The discrete outcome space of interest is a mapping by a function Z'' of each document to a real number. (We note that Landry (1971) has attempted to build an indexing formalism on mappings, but the following is not based on Landry's work.) The author suggests Z'' may be seen as a composition of two separate mappings: (1) a mapping from the set of documents into the set of sets of type $Q \cap T_d$, i.e. from S to 2^Q, and (2)a mapping from each number of 2^Q into the real line. If these functions are denoted by Y and W'' respectively, then by definition:

$$Z''(S) = W'' o\ Y\ (S).$$

(A more detailed approach would express Y as a composition of two other functions: one mapping the set of documents to the power set of the set of attributes, and the other mapping the latter power set to the power set of Q. In effect this would distinguish the separate roles of (1) indexing and (2) document term set-query intersection.)

If we attach a subscript to Z'' indicating its domain, so that:

$$Z''_S(S) = W''_S o\ Y_S(S)$$

$$Z''_A(A) = W''_A o\ Y_A(A)$$

$$Z''_{S\backslash A}(S\backslash A) = W''_{S\backslash A} o Y_{S\backslash A}(S\backslash A)$$

- then we have at hand the three functions that should, more literally, feature in the signal-detection formalism. As a matter of terminology, we refer to Z'' as a 'document weighting function'.

To clarify the workings of the functions described above, consider a query (as SFQ) consisting of three terms:

$$Q = \{ t_a, t_b, t_c \}.$$

The power set of Q, 2^Q, is then the set:

$$\{\emptyset, \{t_a\}, \{t_b\}, \{t_c\}, \{t_a, t_b\}, \{t_a, t_c\}, \{t_b, t_c\}, \{t_a, t_b, t_c\}\}$$

with eight (2^3) members. Each document in the data-base will be mapped by Z''_S, and one or other of Z''_A and $Z''_{S\setminus A}$, into 2^Q by a function Y_S and one or other of Y_A and $Y_{S\setminus A}$ respectively. (The choice in each case depends on whether the document is relevant or not.)

From the point of view of Boolean logic, the functions Y_S, Y_A and $Y_{S\setminus A}$ may be viewed as mapping each document in the appropriate set into the set of elementary logical conjuncts defined by $Q = \{t_a, t_b, \ldots, t_n\}$, i.e. into the set of 2^n ($n = \|Q\|$) elementary logical expressions of the form:

$$t_a^{i_a} \wedge t_b^{i_b} \wedge t_c^{i_c} \wedge \cdots \wedge t_n^{i_n} \; ; \quad i_j = 0,1, \quad n = \|Q\|,$$

- where the $t_j^{i_j}$ are logical variables, and where t_j^1 has the value TRUE if and only if the term denoted by t_j has been used to index the document of interest, and where t_j^0 has the value NOT (t_j^1).* This set, which we denote L_Q, can be put in one-one correspondence with the set 2^Q. For example, when n=3 as in the example given above, the correspondence is:

---

* For example, t_a may denote the index form ARTERY, to choose arbitrarily a medical term, in which case the corresponding logical variable t_a^1 records whether it is TRUE or FALSE that a given document has ARTERY assigned to it. The writer is indebted to E.D. Barraclough for pointing out that t_a^1 may in fact be given a broader construction: it may itself denote a logical expression such as $t_r^1 \wedge (t_s^1 \vee t_u^1) \wedge t_o^0$. In particular, it may denote an expression in which all the logical operators are of OR-type, i.e. $t_r^1 \vee t_s^1 \vee t_u^1 \vee \ldots$, reminiscent of the 'exploded term' concept in MEDLINE. One would expect the latter form to be commonly used when the indexing vocabulary is hierarchically organised, since the hierarchy is equivalent to a series of disjunctions of terms.

$$\{\emptyset\} \quad \longleftrightarrow \quad t_a^0 \wedge t_b^0 \wedge t_c^0$$

$$\{t_a\} \quad \longleftrightarrow \quad t_a^1 \wedge t_b^0 \wedge t_c^0$$

$$\{t_b\} \quad \longleftrightarrow \quad t_a^0 \wedge t_b^1 \wedge t_c^0$$

$$\{t_c\} \quad \longleftrightarrow \quad t_a^0 \wedge t_b^0 \wedge t_c^1$$

$$\{t_a, t_b\} \quad \longleftrightarrow \quad t_a^1 \wedge t_b^1 \wedge t_c^0$$

$$\{t_a, t_c\} \quad \longleftrightarrow \quad t_a^1 \wedge t_b^0 \wedge t_c^1$$

$$\{t_b, t_c\} \quad \longleftrightarrow \quad t_a^0 \wedge t_b^1 \wedge t_c^1$$

$$\{t_a, t_b, t_c\} \quad \longleftrightarrow \quad t_a^1 \wedge t_b^1 \wedge t_c^1$$

It is thus simply a matter of notation as to whether the outcome space of the functions Y_i is characterised in terms of subsets of Q or in terms of logical expressions of the type described.

A complication that we note but do not pursue here is that the functions Y_i can 'break up' the elementary logical conjuncts even further, by ANDing them to further propositions that record the entire sets of forms attached to individual documents. For example, if two documents have been indexed by $\{t_a, t_b, t_x, t_y, t_z\}$ and $\{t_a, t_b, t_p, t_q, t_r\}$, the functions Y_i may be such as to distinguish them, notwithstanding that the query (as SFQ) does not contain any of the terms $t_x, t_y, t_z, t_p, t_q, t_r$. Salton's cosine weighting function is of this more complicated type. A second complication is that the attribute attached to the document may not be a term at all but instead a numerical value, for example the present age of the document, x say. In this case the logical variable of interest is $x < x_c$ where x_c is some specified value. Appropriate elementary conjuncts are then of the form:

$$t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n} \wedge (x < x_c).$$

Obviously the query, Q, must now include x_c as well as the search

terms, i.e.

$$Q = \{ t_a, t_b, t_c, \ldots, x_c \}.$$

It is instructive to go back over what we have introduced from a slightly different point of view. The elementary logical conjuncts ℓ_i which are members of L_Q are assigned a particular permutation by a document weighting function that strongly orders them. That is, some document weighting functions will place the ℓ_i in a given order, say:

$$(\ell_3)_1, (\ell_4)_2, \ldots, (\ell_i)_j, \ldots (\ell_{10})_{2^n}$$

where the subscript i is an arbitrary labelling of the 2^n members of L_Q. In general the ℓ_i will be weakly ordered, however, and this can be represented as a partitioning of at least one such permutation, i.e. a 'composition' of the members of L_Q. In the case of document weighting functions of the former type, the value of J must be 2^n, and for document weighting functions of the latter type, the value of J will be less than 2^n.

It may be useful at this stage to portray the various functions we have discussed, and the elementary logical conjuncts, diagrammatically. Fig. 3.3.2.1-1 attempts this.

The function W_S'' will in practice (for information retrieval systems using explicit weighting of documents) attach a real number to each member of L_Q (or of 2^Q). The numbers so generated are not necessarily distinct. The simplest example is:

$$W_S''(\ell) = \begin{cases} 0; & \ell = t_a^o \wedge t_b^o \wedge \ldots \wedge t_n^o \\ 1; & \ell \in L_Q,\ \ell \neq t_a^o \wedge t_b^o \wedge \ldots \wedge t_n^o \end{cases}$$

More complicated functions are as defined previously. For example

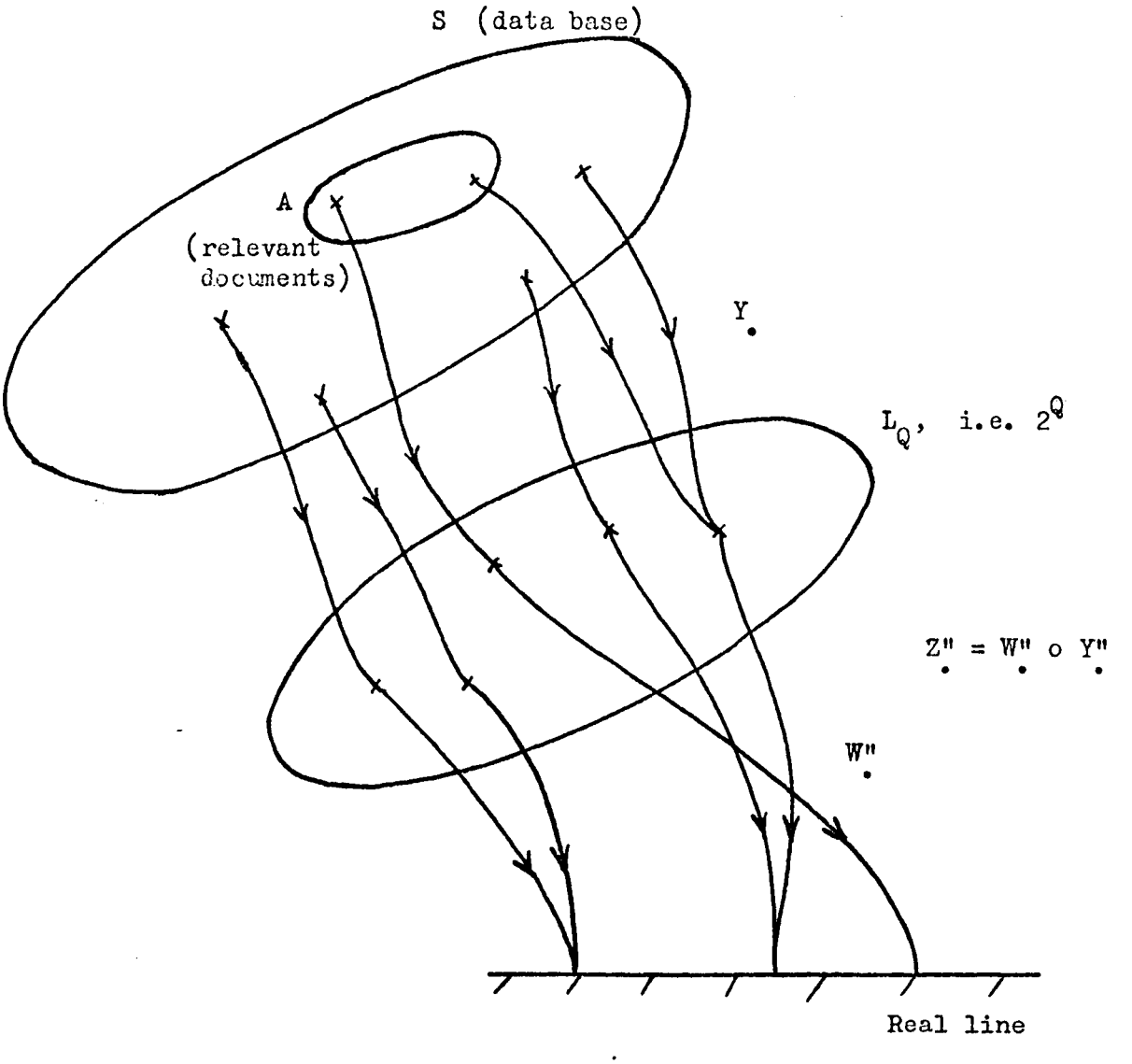


Fig. 3.3.2.1-1 Functions involved in the Swetsian formalism taking discrete z-values into account.

co-ordination level weighting is described by means of:

$$W''_S(\ell) = i_a + i_b + \ldots + i_n; \quad \ell \in L_Q.$$

(The different notation here is just to emphasise the dependence of W''_S on the outcome space defined by Y_S.) In terms of search logic however, we should see W''_S as simply ordering disjunctions of the previously defined elementary conjuncts. For example, $W''_S(\ell)$ as defined for co-ordination level weighting, is in effect simply ORing events together as follows (for $Q = \{ t_a, t_b, t_c \}$):

$$t_a^0 \wedge t_b^0 \wedge t_c^0$$

$$(t_a^1 \wedge t_b^0 \wedge t_c^0) \vee (t_a^0 \wedge t_b^1 \wedge t_c^0) \vee (t_a^0 \wedge t_b^0 \wedge t_c^1)$$

$$(t_a^1 \wedge t_b^1 \wedge t_c^0) \vee (t_a^1 \wedge t_b^0 \wedge t_c^1) \vee (t_a^0 \wedge t_b^1 \wedge t_c^1)$$

$$t_a^1 \wedge t_b^1 \wedge t_c^1$$

– and then placing the new expressions in the order shown*. Although our concern in this section is to redefine the Swetsian outcome space, the reader will see that the discussion raises the question as to whether the numerical values given by an analytical function are of any significance as compared to the rank-order values of logical expressions that the analytical function thereby determines. Since, to answer this question, we need to redefine our concepts of retrieval effectiveness so as to take the discreteness in the outcome space into account, we postpone further discussion on it until the next section. Our remaining concerns here are (1) to redefine the probability distributions induced by Z''_S, Z''_A and $Z''_{S\setminus A}$, and (2) to comment on two constraints on the modelling of these induced distributions.

---

* A different but equivalent viewpoint is that W''_S weakly orders the elementary conjuncts.

A real number, z, corresponding to the event $\{z\}$, which is also a member of $Z''_S(S)$, defines three distinct probability values:

- the probability that a document is assigned that value, given by

$$f(z) = \| Z''^{-1}_S (\{z\}) \| / \| S \|$$

- the probability that a relevant document is assigned that value, given by

$$f_1(z) = \| Z''^{-1}_A (\{z\}) \| / \| A \|$$

- the probability that a non-relevant document is assigned that value,

$$f_2(z) = \| Z''^{-1}_{S\setminus A} (\{z\}) \| / \| S \setminus A \| .$$

The set of such values defines three induced distributions, denoted by f, f_1, and f_2 respectively. If it is the case that a value of z, belonging to $Z''_s(S)$, is not also a value of $Z''_A(A)$ then we refer to z as an 'almost impossible' event for the function Z''_A; similarly for $Z''_{S\setminus A}$. If the mapping of documents is such that one of the 'allowed' numerical values of the analytical function is not mapped to by any document then we refer to that value as a 'compound almost impossible' (CAI) event. (For example, the co-ordination level value of '2' is a feasible one for queries (as SFQ) of size two terms or more. If the query and relevance-assignments are such that no documents at all are assigned this value, then the value is a CAI event.)

When we seek to model the induced distributions (i.e. pursue the same objective that Swets did, but now for discrete distributions rather than continuous), the following two fundamental decisions have first to be made:

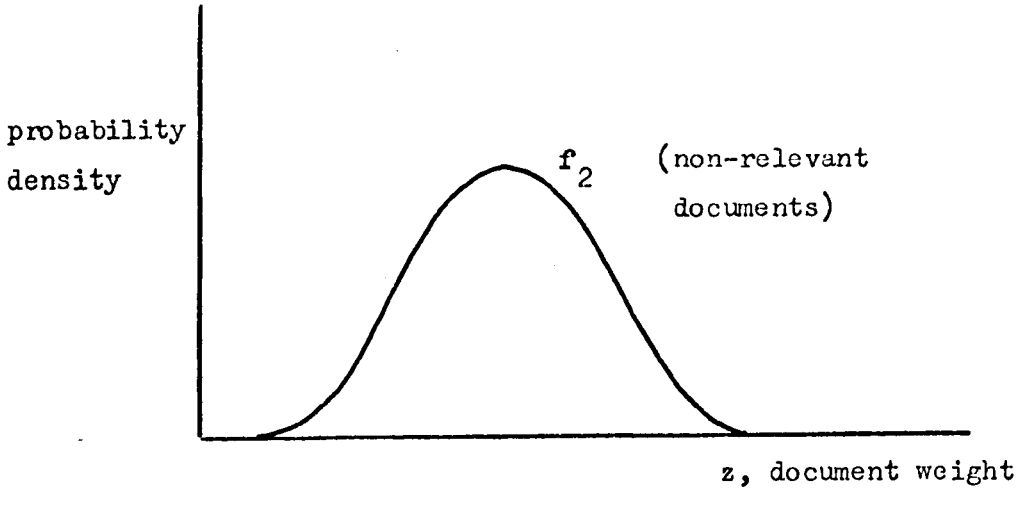
(1) Are the value of $W_i''(\emptyset)$ to be within the scope of the modelling functions?

(2) Are almost impossible, and/or CAI events to be within the scope of the modelling functions?

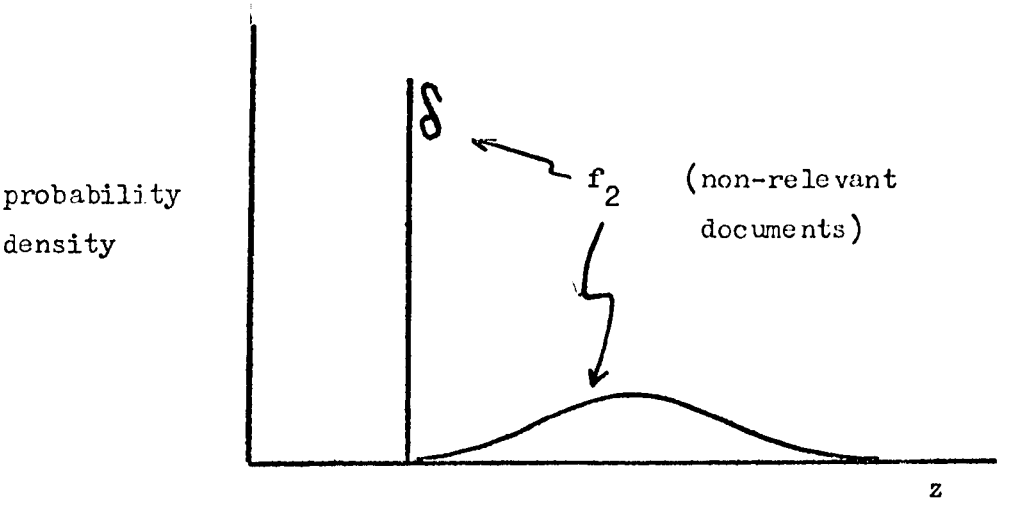
The first of the above questions is critical for modelling $Z''_{S\setminus A}$, since a very large proportion of the non-relevant documents will be mapped to the event $\emptyset$; i.e. for most non-relevant documents (perhaps 99% in practice for $\| Q \| \doteq 5$) the elementary conjunct:

$$t_a^o \wedge t_b^o \wedge t_c^o \wedge \cdots$$

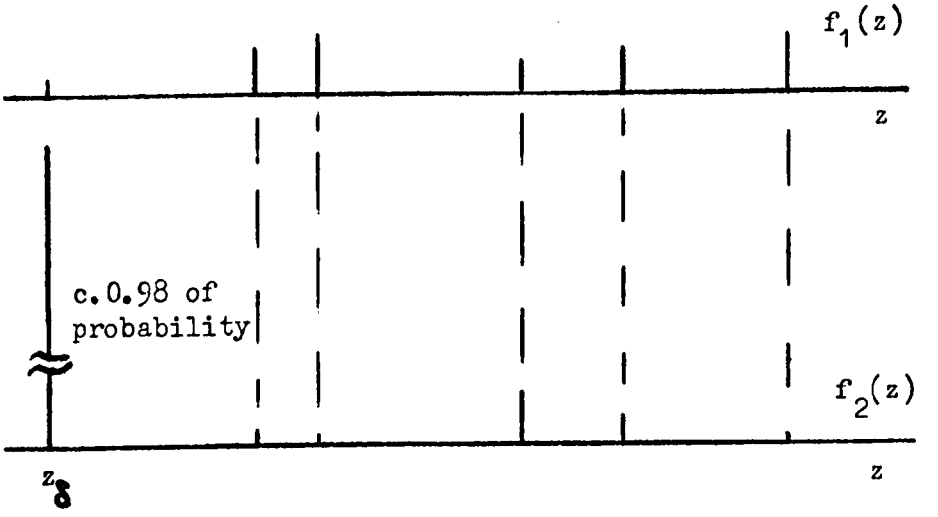
will evaluate to TRUE. This 'spike' of probability was completely ignored by Swets, and to the author's knowledge has also not been commented on by other workers. In effect it makes a nonsense of a modelling distribution in the form of a Normal distribution for non-relevant documents, <u>unless</u> it is understood that non-relevant documents for which $Q \cap T_d = \emptyset$ are disregarded by the model. Diagramatically, what was suggested by Swets as a suitable model:



should, if entertainable at all, be replaced by:



- where δ here denotes a spike of probability density accounting for non-relevant documents for which $Q \cap T_d = \emptyset$, i.e. $z = z_\delta = w''_S(\emptyset)$. Whether a continuous modelling function is worth rescuing by such means is open to question of course. The following diagram illustrates the true nature of $f_1(z)$ and $f_2(z)$, showing how serious it is to ignore the event $w''_S(\emptyset)$ in defining f_1 and f_2.

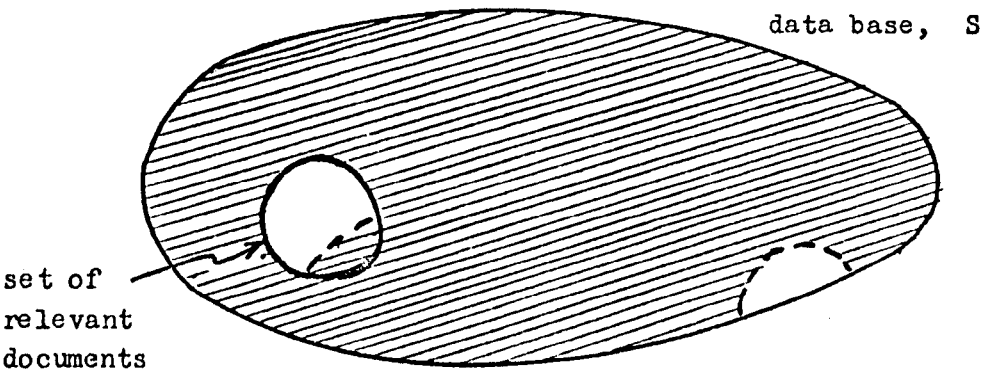


In the author's view, the ignoring of the probabilities attaching to the event $t_a^o \wedge t_b^o \wedge t_c^o \wedge .. \wedge t_n^o$, a consequence of Swets not having defined the outcome space satisfactorily and in

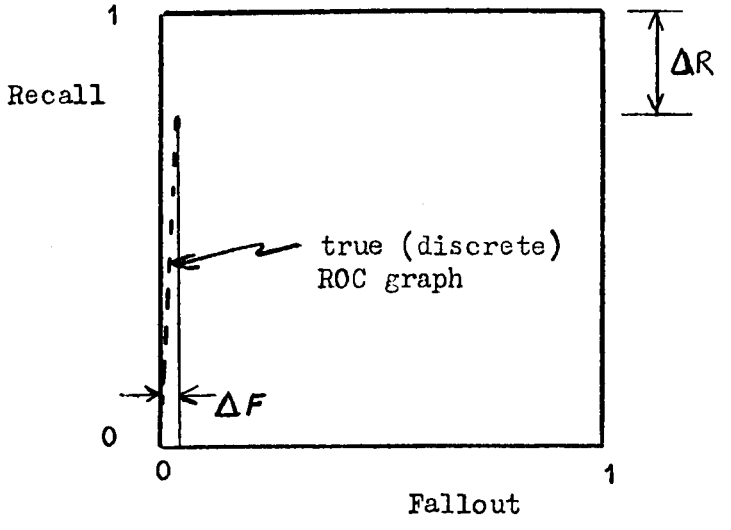
turn a consequence of his not having distinguished between formalism and model, constitutes a severe limitation on the scope of the original (implied) formalism. It also introduces an inconsistency in Swets's treatment of experimental data. To see this more clearly, we first define a document to be 'pertinent' when the query is such that the logical expression $t_a^o \wedge t_b^o \wedge t_c^o \wedge \ldots \wedge t_n^o$ evaluates to FALSE. Pertinence, so defined, will vary from query to query. (As usual, we define a query here as a set of terms.) Since Swets's Normal densities clearly ignored the two probabilities attached to the logical expression described, it must be the case that he was concerned exclusively with pertinent documents, not with the whole data base. That is, Swets was describing only 2% or so of a data base. It also follows that Swets could not have been dealing with Recall and Fallout as he claimed but instead only with the proportions of relevant and pertinent, and non-relevant and pertinent documents retrieved, respectively. That is, Swets was concerned with 'conditional Recall' and 'conditional Fallout' (R' and F' say) defined by

$$R' = k_R R \qquad ; \qquad F' = k_F F$$

where k_R and k_F are usually non-zero parameters that vary with both Q and the partitioning of the data base by the information need. The following diagram illustrates the difference in scope of the documents within the original formalism of Swets and those within the extended formalism:



The white area signifies pertinent documents, for a given query as SFQ. The shaded area signifies non-pertinent documents. Swets treated only the pertinent documents, notwithstanding his claim to be dealing with the whole set of documents. The ROC graphs given by Swets are misleading, since the Recall and Fallout variables require to be relabelled as R' and F'. A true (and discrete) ROC graph is, on the basis of the work reported in this thesis, more likely to be as shown below, with the axes now properly labelled as R and F, and with the intervals ΔR and ΔF shown being of the order of 0.25 and 0.01 respectively for a query of about five terms.



The second of the basic questions we need to consider is whether almost-impossible events and/or CAI events are to be within the scope of the modelling fractions. This is relevant in the following intuitive way. Suppose we map all the events $q \in 2^Q$ to distinct values that are in fact the rank values of $W_s''(q)$. Call the ranking function W_s. Assume, as will often be the case in practice, that for some of these rank values $f_1 = 0 = f_2$, i.e. the events are CAI. Then the variation of retrieval effectiveness defined by varying a threshold value over the rank value is the

same whether or not we include the CAI events. But their presence will alter the number of values that a modelling function has to address. This informal approach is made clearer in the next section.

Modelling functions, for a recognised discrete outcome space, are thus subject to four definitional constraints, which we label as follows:

		$W_s(\emptyset)$ within scope of model?	
		Yes	No
CAI events within scope of model?	Yes	Constraint 1	Constraint 3
	No	Constraint 2	Constraint 4

Table 3.3.2.1-1

In this section we have recognised the outcome events that are such an essential ingredient of the Swetsian formalism, as discrete events. We have also mapped logical search expressions to these discrete events. As mentioned previously by the writer (Heine, 1975) a formal statement of the possibility of linking weights and logical search expressions is due to Angione (1975), with some prior less-general discussion of the matter by Uhlmann (1968), Brandhurst (1966) and Iker (1967). A recent, relevant 'textbook' approach has also been offered (Mott et al, 1972). A classic, basic text in the logic area is Korfhage (1966). None of these works offers discussion in a signal-detection context, however. It is emphasised that the complete self-consistency of Boolean retrieval and retrieval using document weights stated by Angione, has been demonstrated in this section of this thesis only for weighting function with domain $Q \cap T_d$. However it is believed that the consistency between the two approaches is perfectly

general provided a much richer domain is first set up. If the set of all sets of terms assigned to documents is denoted by the set of 'document representatives' $\{dr_i\}$, to use a term introduced by van Rijsbergen, then the richer domain for say Salton's cosine measure is simply the Cartesian product of $\{Q \cap T_d\} \times \{dr_i\}$. Many more elementary logical conjuncts then need to be defined, but the reasoning given in this section would still hold.

To summarise, we have in this section drawn attention to a weakness in what is the most basic feature of Swets's formalism, the outcome space. We have noted that events in this space cannot be continuous as described by Swets, but must be discrete, notwithstanding that the space itself is continuous. We have also noted that this in itself does not invalidate the use of continuous probability density functions, provided these are seen simply as modelling functions. We have also succeeded in linking logical searching, as used in conventional search practice, with the formalism. This was by breaking down the operation of document weighting in to two stages: a mapping from documents to elementary logical conjuncts, the elementary propositions of which are logical variables denoting term absence/presence, where the terms are members of the query (as SFQ); and a mapping from these elementary conjuncts to the real numbers. In Swets's description of the formalism, the particular logical expression $t_a^o \wedge t_b^o \wedge t_c^o \wedge \ldots \wedge t_n^o$ was not mapped to, implying that Swets did not describe all of the documents in the data base. This follows from the exclusion of a 'spike' of probability accompanying the distribution of non-relevant documents, i.e. from his portraying this distribution as a simple Normal distribution. In any attempt to characterise the probability distributions that the functions Z_i induce, definitional constraints

on the modelling functions used are required: four such constraints have been described.

Although the subject of this thesis is Swets's theory, the writer notes briefly here the possibility that the outcome space that is the subject of the Swetsian formalism may be over-specific. The random variables W_i simply record the effect of disjoining logical search expressions in a pre-specified way. A stronger formalism would stop at the probability distributions over the elementary conjuncts, and not limit the way that these are disjoined. Further relevant discussion on this point will be introduced after we have redefined the basic probabilistic measures of retrieval effectiveness (Section 3.3.2.3). This will be in Sections 3.3.3.1 and 3.4.

3.3.2.2 The distributions f_1 and f_2

We have seen that the basic distributions of interest in the extended Swetsian formalism are induced, discrete distributions defined on order-numbers of a set of weakly-ordered elementary logical conjuncts. (The latter in general are weakly ordered, but may be strongly ordered.) We have, in previous sections, defined these two distributions as functions of a real-valued outcome variable z, and denoted them as $f_1(z)$ and $f_2(z)$. But since z serves solely as an ordering device, we will henceforth write these as $f_1(z)_j$ and $f_2(z)_j$, $j \in J$, J an index set of the z-values, when we wish to emphasise the relevance of this ordering.

At this stage it is timely to identify the new random variables of interest, to reflect our abandonment of any primary interest in the actual z values. Our earlier notation in fact anticipates this change. We now define new random variables as follows:

Z_S maps all documents to the index set J

Z_A maps all relevant documents to J

$Z_{S\backslash A}$ maps all non-relevant documents to J.

The functions mapping L_Q to J are labelled W_S, W_A and $W_{S\backslash A}$, where $W_\cdot = W'_\cdot \circ W''_\cdot$, and where the functions $W'_\cdot$ map the z values on to J. Thus

$$Z_S = W_S \circ Y_S(S)$$

$$Z_A = W_A \circ Y_A(A)$$

$$Z_{S\backslash A} = W_{S\backslash A} \circ (S\backslash A)$$

- as shown also in Figure 3.3.2.2-1

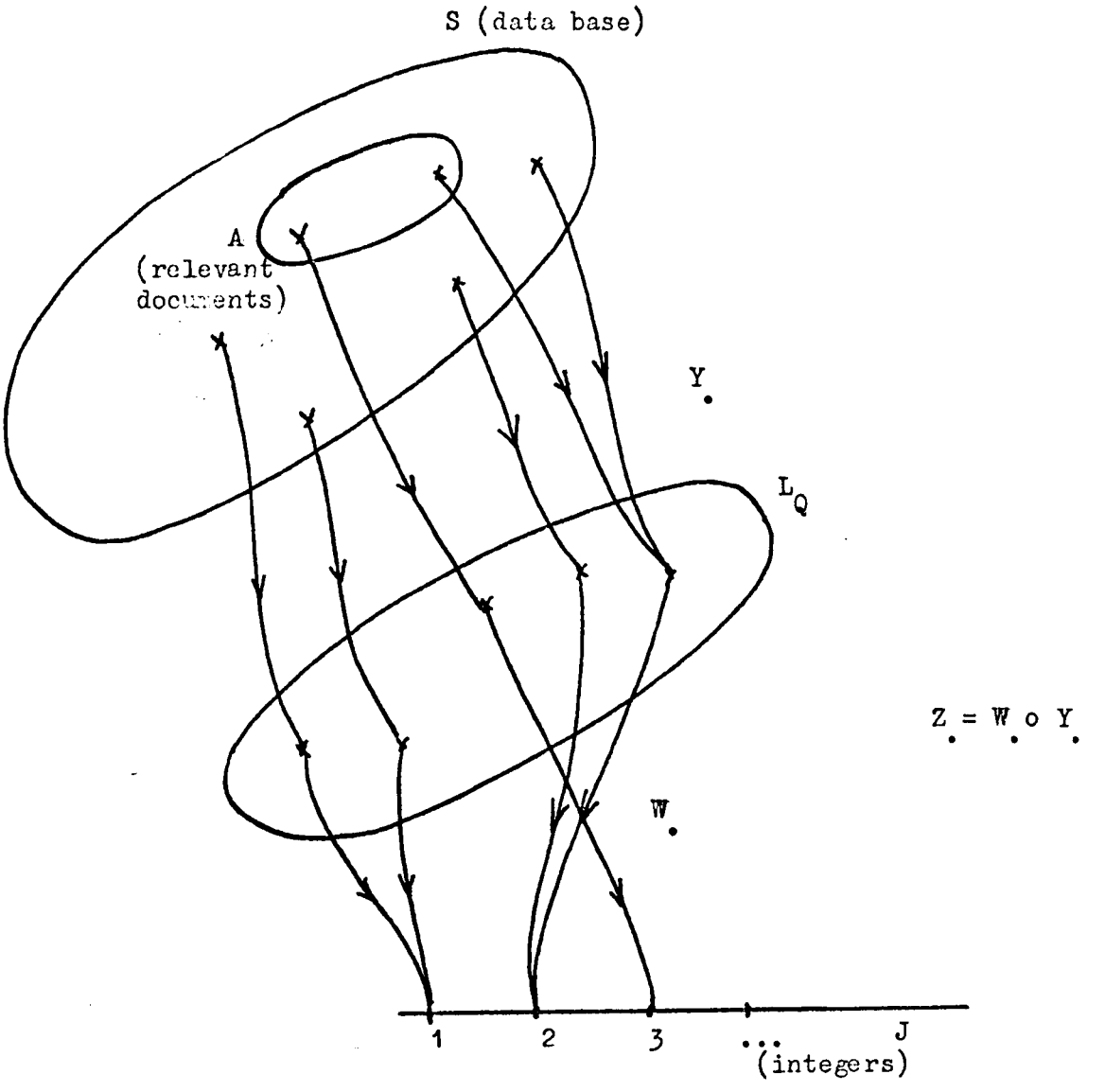


Fig. 3.3.2.2-1 Functions and random variables involved when z values are replaced by their corresponding rank values.

It is of interest to know whether analytical expressions exist that will model the functions $f_1(z)_j$ and $f_2(z)_j$, although exactly how such modelling distributions should be evaluated is a question postponed until after we have discussed the measurement of retrieval effectiveness. The binomial and Poisson distributions over the values of $j \in J$ naturally suggest themselves. But how realistic are they? The value of J is finite, so the Poisson distribution, if considered at all, would need to be truncated with the probability-tail values redistributed in some ad hoc way. This may be unimportant if most of the probability is concentrated in low j values, but this is true only for f_2 in practice. However even for f_2 almost all the probability (perhaps 98% of it) is concentrated on just one value, $j = 1$, suggesting that a step-function modelling approach may be preferable to one based on the Poisson distribution. The binomial distribution as model has some appeal when the particular document weighting expression used is co-ordination level. For then the random variable Z_s'' is of the form of a sum of Bernoulli variables:

$$Z_s'' = X_a + X_b + X_c + \ldots + X_m$$

where X_i has the value 1 if the document has form t_i assigned to it and the value 0 otherwise. In this case Z_s'' can be binomial, but only if two further conditions are true: The parameters of the X_i are identical, and X_i and X_j are independent, i.e. $\mathrm{Cov}(X_i, X_j) = 0$, $i \neq j$. The first of these assumptions is equivalent to assuming that all forms have the same specificity, the second that no clustering occurs. Neither assumption is realistic. When Z_A and $Z_{S\setminus A}$ are defined in a similar way (i.e. using Bernoulli variables defined for the sets A and $S\setminus A$) the

assumptions seem even more vulnerable. In view of the weaknesses in binomial-modelling for this, the simplest, type of document weighting, and similar vulnerability in any attempt to apply the Central Limit Theorem to modelling f_1 and f_2, the question must at present be seen as an open one. The writer emphasises that the absence of any plausible model in no way invalidates the extended Swetsian formalism. It does however argue for a less 'statistical' and more 'scientific' approach to the matter: the distributions f_1 and f_2 should be examined for what they are in practice, without excessive concern being shown for approximating them using analytical expressions.

Two further points relevant to f_1 and f_2 will be discussed briefly here.

Hutchinson (1978) has radically, though perhaps controversially, extended the scope of $f_1(z)$ and $f_2(z)$, by introducing the "degree of relevance" as a parameter. In effect the two induced distributed are postulated as having a conditional character. They are replaced by functions that we can label $f_1(z|A_j)$ and $f_2(z|S\backslash A_j)$, where A_j denotes the set of documents each of which is of a "degree of relevance" greater than j. (As mentioned earlier in Section 2, this approach is influenced by Robertson's work, and by previous literature portraying relevance as a construct capable of quantitative interpretation.) Hutchinson puts forward the hypothesis, on this basis, that f_1 and f_2 are then bivariate-Normal, not necessarily with parameter ρ equal to 0. Implicit in this is the notion that the variable "degrees of relevance" can take on all real values, not just integer values, which seems an unreasonably strong assumption, especially if users are disposed more to think

in terms of types of relevance rather than in quantitative terms. Also implicit in Hutchinson's hypothesis is the assumption that $f_2\,(z|S\backslash A_j)$ relates only to pertinent documents since, as we have seen, any Normal portrayal of f_2 will miss out those documents defined by the logical expression:

$$t_a^o \wedge t_b^o \wedge \ldots \wedge t_n^o.$$

The set concerned is thus not $S\backslash A_j$, as we have provisionally portrayed it, but:

$$(S\backslash A_j) \cap (s:\ t_a^o \wedge t_b^o \wedge \ldots \wedge t_n^o = \text{FALSE}).$$

Hutchinson's contribution could be a very useful one, the writer suggests, if (1) the sets we have written as A_j here are taken to denote sets of documents relevant to a given information need in different ways, i.e. if qualitative criteria are introduced instead of a pseudo-quantitative one, and (2) it is seen as <u>generating</u> hypotheses concerning f_1 and f_2 rather than placing undue and unreasonable weight on a particular <u>a priori</u> hypothesis concerning their joint variation.

We note now a feature of f_1 and f_2 which previous writers on signal detection theory do not seem to have recorded. This is that the moments cannot be independent. For example, consider the first moment about the origin. The means of Z_S, Z_A and $Z_{S\backslash A}$ are constrained by: $E(Z_S) = G\,E(Z_A) + (1-G)\,E(Z_{S\backslash A})$. This follows from:

$$E(Z) = \sum z\, f(z) \qquad \text{(definition)}$$

$$= \sum z \left(\frac{\| Z_S^{-1}(\{z\}) \|}{\| S \|} \right)$$

$$= \sum z \left(\frac{\| Z_S^{-1}(\{z\}) \cap A \| + \| Z_S^{-1}(\{z\}) \cap S\backslash A \|}{\| S \|} \right)$$

since $A \cap (S\backslash A) = \emptyset$

$$= \sum \left(z \frac{\| Z_A^{-1}(\{z\}) \|}{\| S \|} + z \frac{\| Z_{S\setminus A}^{-1}(\{z\}) \|}{\| S\setminus A \|} \right)$$

$$= \sum \left(zG \frac{\| Z_A^{-1}(\{z\}) \|}{\| A \|} + z(1-G) \frac{\| Z_{S\setminus A}^{-1}(\{z\}) \|}{\| S\setminus A \|} \right)$$

$$= G.E(Z_A) + (1-G)\, E(Z_{S\setminus A}).$$

By a similar sequence of steps, the variances of Z_S, Z_A and $Z_{S\setminus A}$, are constrained by*:

$$V(Z_S) = G\, V(Z_A) + (1-G)\, V(Z_{S\setminus A}) + G(1-G)(E(Z_A) - E(Z_{S\setminus A}))^2.$$

Relationships between moments, such as those above, provide structural constraints on hypotheses expressed in the Swetsian formalism. For example, a hypothesis that $E(Z_A) = k\, E(Z_{S\setminus A})$ for a fixed data bases fixed choice of weighting function, and a fixed method of generating queries, for various information needs, is incompatible with the second hypothesis: $E(Z_S) = \text{const.}$, when G also is not held constant.

---

* For the suggestion that this particular relationship should be sought, the writer acknowledges B.C. Brookes (pers.comm.).

3.3.2.3 Re-definition of the probabilistic measures of retrieval effectiveness on a discrete outcome space; the optimisation of the retrieval process; terminological note

The probabilistic measures of effectiveness are easily redefined in the discrete formalism. In order first to arrive at definitions of Recall and Fallout, we first denote the probability distributions over the events $q \in 2^Q$, i.e. over the events $t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n}$, by:

$$\left\{ r_{ab\ldots n}^{i_a i_b \ldots i_n} \right\} \quad \text{and} \quad \left\{ s_{ab\ldots n}^{i_a i_b \ldots i_n} \right\}$$

-for the sets of relevant and non-relevant documents respectively. It is emphasised that each set denotes a probability distribution, not an induced probability distribution*. The events to which the individual probability distributions refer are as yet unordered. Then if we choose a Boolean search expression which is the disjunction of some set of elementary conjuncts, indexed by K say, the Recall and Fallout values will be:

$$R_K = \sum_{k \in K} \left(r_{ab\ldots n}^{i_a i_b \ldots i_n} \right)_k \; ; \quad F_K = \sum_{k \in K} \left(s_{ab\ldots n}^{i_a i_b \ldots i_n} \right)_k$$

$$\left(\text{for } E_K = \bigvee_{k \in K} \left(t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n} \right)_k \right) .$$

The associated Precision value is given by $(GR/[\,GR + (1-G)F\,])_K$, and the associated Marczewski-Steinhaus metric value is given by: $D = ([\,F(1-G) + G(1-R)\,]/[\,F(1-G) + G\,])_K$.

In the extended formalism, our concern is retrieval from the data base using logical search expressions in a certain sequence. The sequence is of course that determined by the numerical value

* An induced probability distribution is one defined by a random variable, i.e. a function mapping a probability space to the real numbers. (Barr, 1971)

to which the elementary conjuncts are mapped, though an optimum sequence is determined by the ratio r/s. Suppose that the order of these conjuncts is recorded by the variable J. This is made clearer by way of an example. If $Q = \{t_a, t_b, t_c\}$, and the document weighting function is co-ordination level, then the values of J, the weights defining the values of $r_{abc}^{i_a i_b i_c}$ and $s_{abc}^{i_a i_b i_c}$, are as follows:

Co-ordination value (weight, z)	Rank value (J)	prob. rel. doc. retrieved at that weight	prob. non-rel. doc. retrieved at that weight
0	1	$r_{abc}^{000} \equiv r_1' = f_1(0)$	$s_{abc}^{000} \equiv s_1' = f_2(0)$
1	2	$r_{abc}^{001} + r_{abc}^{010} + r_{abc}^{100} = r_2' = f_1(1)$	$s_{abc}^{001} + s_{abc}^{010} + s_{abc}^{100} = s_2' = f_2(1)$
2	3	$r_{abc}^{011} + r_{abc}^{101} + r_{abc}^{110} = r_3' = f_1(2)$	$s_{abc}^{011} + s_{abc}^{101} + s_{abc}^{110} = s_3' = f_2(2)$
3	4	$r_{abc}^{111} = r_4' = f_1(3)$	$s_{abc}^{111} = s_4' = f_2(3)$

The variables r_1' and s_1' are defined as shown. We now denote logical search expressions appropriate to each rank value by e_i, e.g.:

$$e_3 = (t_a^0 \wedge t_b^1 \wedge t_c^1) \vee (t_a^1 \wedge t_b^0 \wedge t_c^1) \vee (t_a^1 \wedge t_b^1 \wedge t_c^0).$$

Then for this example the Swetsian formalism is concerned with the effects of using successively more general logical search expressions E_j defined by:

$$E_j = \bigvee_{i=5-j,5} e_i \quad ; \quad j = 1,2,3,4.$$

That is, $E_1 = e_4$; $E_2 = e_4 \vee e_3$; $E_3 = e_4 \vee e_3 \vee e_2$; and $E_4 =$

$e_4 \vee e_3 \vee e_2 \vee e_1$. For the jth search expression, the Recall and and Fallout values are:

$$R_j = \sum_{t=j,4} r'_i \quad ; \quad F_j = \sum_{i=j,4} s'_i$$

with associated D_j and P_j values following as indicated earlier. The expressions e_i can possibly be simplified using 'Boolean minimization' techniques, but this is not pursued here.

The above example is readily generalised. Denote the rank values of the real numbers given by the weighting function to documents in the data base, by 1,2,3, ..., J. Denote by e_i the search expression formed as the disjunction of all those elementary conjuncts mapped to the integer i. Lastly define probabilities attached to each e_i as follows:

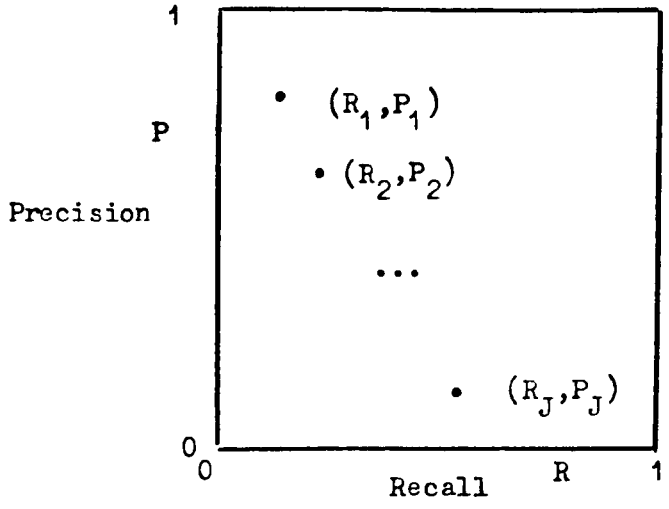
$$r'_i = \sum_{k \in K_i} \left(r^{i_a i_b \cdots i_n}_{a\ b \cdots n}\right)_k \quad ; \quad s'_i = \sum_{k \in K_i} \left(s^{i_a i_b \cdots i_n}_{a\ b \cdots n}\right)_k$$

where K_i is an index set determined by the weighting function; i.e. the summations given here are over the probabilities associated with the elementary conjuncts from which e_i is formed. Then the document weighting function used will determine a sequence of logical search expressions: $E_1, E_2, \ldots, E_J$ defined by:

$$E_j = \sum_{i=J+1-j,J} e_i \quad ; \quad j = 1,2,3, \ldots, J.$$

The optimum such function is of course that in which the e_i are ranked by r'_i/s'_i. Precision and other values follow as indicated earlier. Accordingly a sequence of paired Recall and Precision values is determined by the weighting function: (R_j, P_j), $j = 1,2, \ldots, J$. The <u>set</u> of such values $\{(R(j), P(j) ; j=1,2,\ldots,J\}$

is of course the Recall-Precision graph of interest:



Note that if we retrieve using the search expression E_J, we retrieve all of the data base. Accordingly, in practice search expressions defined by $j < J$ must be used. In consequence of this, a maximum value of Recall exists, and is equal to: $R_{max} = 1 - r^{oo\ \ldots\ o}_{ab\ \ldots\ n}$. Since $S' = \{ s: s \in S \text{ and } t^o_a \wedge t^o_b \wedge \ldots \wedge t^o_n = \text{FALSE} \}$, i.e. the set of pertinent documents for n terms, is a superset of $S'' = \{ s : s \in S \text{ and } t_a \wedge t_b \wedge \ldots \wedge t^o_{n+1} = \text{FALSE} \}$, i.e. the set of pertinent documents for n+1 terms, it follows that including an extra term in Q will monotonically decrease the value of $r^{oo\ldots o}$ and hence monotonically increase the value of R_{max}. (ANDing a new logical variable to an existing search expression must increase (or hold constant) the set of items for which the expression is FALSE.) It is thus apparent that the maximum Recall attainable, in respect of a given data base and given information need, is determined by both the identity of the terms making up the query and the number of such terms.

It is instructive to relate the preceding discussion to our

earlier discussion of permutations and compositions of the members of L_Q. It is apparent that a document weighting function mapping into a permutation of the members of L_Q will define at most 2^n points on the R-P graph, but one mapping into the partitions of a partitioned permutation (i.e. composition) of the members of L_Q will determine a lesser number. It is also apparent that the R-P graphs determined by different compositions of a <u>given</u> permutation will differ <u>only</u> in the number of points defined, i.e. forming a composition of a permutation determines a <u>subset</u> of the points determined by the permutation itself. The following example makes this point clearer:

<u>Example</u>: Suppose $Q = \{t_a, t_b\}$, and $L_Q = \{t_a^o \wedge t_b^o, t_a^o \wedge t_b^1, t_a^1 \wedge t_b^o, t_a^1 \wedge t_b^1\}$. Then the permutation:

$$(t_a^o \wedge t_b^o,\ t_a^1 \wedge t_b^o,\ t_a^o \wedge t_b^1,\ t_a^1 \wedge t_b^1)$$

determines four (R, P) points, defined as follows:

$$R = r^{11}, \quad F = s^{11}, \quad P = G\, r^{11} / (G r^{11} + (1-G)\, s^{11})$$

$$R = r^{11} + r^{01}, \quad F = s^{11} + s^{01}, \quad P = G\,(r^{11} + r^{01})/(G(r^{11} + r^{01}) + (1-G)(s^{11} + s^{01}))$$

$$R = r^{11} + r^{01} + r^{10}, \quad F = s^{11} + s^{01} + s^{10}, \quad P = (\text{etc.})$$

$$R = 1, \quad F = 1, \quad P = G/(G + (1-G)) = G$$

That is four (R,P) points are determined provided r_{ab}^{ij} and s_{ab}^{ij} are not both zero. This will be the case for a document weighting function mapping the members of L_Q into four distinct real numbers ordered as shown.

If now we examine the composition:

$$(t_a^o \wedge t_b^o \mid t_a^1 \wedge t_b^o \,,\; t_a^o \wedge t_b^1 \mid t_a^1 \wedge t_b^1)$$

in which the order of the two middle elementary conjuncts is immaterial, i.e. if we examine a document weighting function mapping into just three real numbers, then it is apparent that at most three (R,P) points will be determined, and that these form a subset of those given above. The points are defined by:

$R = r^{11}$, $F = s^{11}$ $P = (\text{etc.})$

$R = r^{11} + r^{10} + r^{01}$, $F = s^{11} + s^{10} + s^{01}$, $P = (\text{etc.})$

$R = 1$ $F = 1$ $P = G$

- end of example.

Most Recall-Precision graphs determined by partitioning the set of elementary conjuncts, and ordering the subsets so defined, will be 'poor' in the sense that the points (R_i, P_i) will tend to cluster around the origin. (This will be true in particular if the event $\emptyset \in 2^Q$ is given a rank value near the highest rank values.) The aim, from the point of view of maximising retrieval effectiveness, is to identify, for a given query and partitioned data base that composition* of elementary conjuncts which puts this graph a maximum distance from the origin.

In the remainder of this section we address two problems related to the preceding discussion: What does it mean to 'model' the distributions f_1 and f_2? and In what sense(s) can we talk of 'optimization' of the retrieval process?

* See Appendix A for related notes.

Discrete probability distributions $m_1(j)$ and $m_2(j)$ can be defined, using some analytical function of z, so as to approximate or 'model' the observed probability distributions $f_1(z)_j$ and $f_2(z)_j$. In this case, for a direct comparison of m_i with f_i (i=1,2), the set of z values for the m_i must be the same as that for the f_i. If the functions f_i are defined over the integers (rank value of z values) then so must the functions m_i. Other approaches are possible however, and one of these was implied by Swets. If we define modelling distributions $m_i(z)$ over all possible z value (i.e. over the outcome space of a document weighting function, the real numbers) then although direct comparisons with the f_i are no longer possible it is still possible to compare cumulative distribution functions (CDFs) of the m_i and f_i. The observed CDFs are:

$$F_i^*(z_c) = \sum_{z \leq z_c} f_i(z) = Pr_i(z \in (-\infty, z_c]); \; i=1,2$$

where Pr_i denotes the probability that a document weight lies in the interval shown when the document is relevant (i=1) or non-relevant (i=2). These functions are simply related to Recall and Fallout by:

$$R = 1 - F_1^*(z_c) = Pr_1(z \in (z_c, \infty))$$

$$F = 1 - F_2^*(z_c) = Pr_2(z \in (z_c, \infty)).$$

The comparison then is with modelled Recall and Fallout values, denoted R_m and F_m here, defined by:

$$R_m = \int_{z_c}^{\infty} m_1(z)\,dz \quad ; \quad F_m = \int_{z_c}^{\infty} m_2(z)\,dz \quad .$$

In Swets's presentation (in which formalism and model were regrettably confused and, as we have seen, in which only pertinent documents were defined as within the scope of the formalism), m_1

and m_2 were portrayed as Normal distributions:

$$m_i(z) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp(-(z-\mu_i)^2/2\sigma_i^2); \quad i=1,2.$$

In this case R_m and F_m may be rewritten as follows:

$$R_m(z_c) = \int_{z_c}^{\infty} m_1(z)dz$$

$$= \Phi_c((z_c-\mu_1)/\sigma_1)$$

$$= \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{z_c-\mu_1}{\sqrt{2}\,\sigma_1}\right)$$

$$F_m(z_c) = \int_{z_c}^{\infty} m_2(z)dz$$

$$= \Phi_c((z_c-\mu_2)/\sigma_2)$$

$$= \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{z_c-\mu_2}{\sqrt{2}\,\sigma_2}\right)$$

where $\Phi_c(x) = \int_x^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-t^2/2)dt = 1 - \Phi(x)$ (definition)

and $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2).\,dt$ (definition).

It is of course possible to model the variation of Precision (P_m) and other probabilistic measures of effectiveness with z_c, and the co-variation of say R_m and P_m, using these relationships as has been described elsewhere by the author (Heine, 1974).

Still other approaches to modelling the actions of Z_A and $Z_{S\setminus A}$ are possible. We mention two. We can compare, not the CDFs of the m_i and f_i, but the Recall-Precision graphs that they determine (i.e. that m_1 and m_2 determine, and f_1 and f_2 determine). Assuming that the m_i are defined as finite, discrete distributions (e.g. discrete-

uniform, or binomial, or truncated-Poisson), and pertain to the same z values (or their ranks) as do the observed distributions $f_i(z)_j$, we can measure the Euclidean distances between comparable co-ordinate pairs. That is, we can compare:

$$R(J_c) = \sum_{j>J_c} f_1(z)_j \quad ; \quad F(J_c) = \sum_{j>J_c} f_2(z)_j$$

and the associated Precision value, $P(J_c)$, with:

$$R_m(J_c) = \sum_{j>J_c} m_1(j) \quad ; \quad F_m(J_c) = \sum_{j>J_c} m_2(j)$$

and the associated Precision value, $P_m(J_c)$, simply by measuring the Euclidean distance between (R,P) and (R_m, P_m) for each z (or j) value. (J_c is just a varying threshold value. The modelling functions $m_i(j)$ are here defined over the rank values, j.) The adequacy of the functions $m_i(j)$ as models is then indicated by small values of the mean and variance of this Euclidean distance, for all z (or j) values. Again, we can compare, for each value of z (or j) a measure of the distance between the retrieved set of documents and the relevant set of documents, e.g. the distance:

$$\Delta D = D(A,B) - D_m(A_m,B_m)$$

where B is the set of documents actually retrieved (i.e. the set $s \in S$ such that $j_s > J_c$) and D_m is the 'modelled' distance calculated from R_m, F_m and G. In this case we would again calculate the mean value of ΔD, a small value indicating good modelling functions $m_i(j)$. The variance of ΔD should, ideally, be small as well.

Lastly, we discuss what it means to talk of an 'optimum' retrieval process for a given, partitioned data base, i.e. for a given instance of information need. First, suppose a query has

been specified in set form (e.g. $Q = \{t_a, t_b, t_c\}$ where t_a, t_b and t_c denote document attributes). Then a sub-optimal retrieval process can be defined by ordering all elementary logical conjuncts of the members of Q by their likelihood-ratio values. If the logical conjunct is CAI, its likelihood-ratio value (o/o) is indeterminate, but its existence in a search expression is immaterial. If the logical conjunct has no non-relevant documents assigned to it, the value is again indeterminate, but we can usefully assign the conjunct a likelihood-ratio value greater than the maximum value for elementary conjuncts not indeterminate, i.e. prefer it as a search expression. Otherwise, we define likelihood-ratio as:

$$L(q) = \left(\frac{Y_1^{-1}(q)}{\|A\|}\right) \Big/ \left(\frac{Y_2^{-1}(q)}{\|S \setminus A\|}\right) \quad ; \quad q \in 2^Q .$$

If we wish, we can 'fine-tune' the likelihood values attaching to elementary conjuncts to which no non-relevant documents are posted. This is simply by ranking them according to the number of relevant documents posted to them. To talk of likelihood-ratio as a weighting function is a little dangerous perhaps, in that the values that the function attaches to documents are known only <u>a posteriori</u>. This is not the case for more conventional weighting functions which depend on the outcome of $Q \cap T_d$ and term specificity values, say. But likelihood-ratio weighting nevertheless provides a benchmark against which other, <u>a priori</u> functions can be compared. The Recall-Precision graph it implies cannot be bettered <u>for the choice of Q concerned</u>.

The point just emphasised is, in the author's view, a critical one. To pursue the 'ideal' weighting function is to pursue a will-

o-the-wisp if it is not recognised that a joint optimum of query (as a set of attributes) and weighting function (based on observable values) is required. To overcome the sub-optimality of the use of a weighting function that does approach the effectiveness of the likelihood-ratio function, the essential step is the introduction into the retrieval process of a heuristic element, i.e. one in which feedback to the enquirer allows for automatic improvement in the membership of Q, as well as sub-optimisation of the logic of the search expression used. For analytical reviews of the literature of feedback in information retrieval the reader is referred to Salton (1975a) and van Rijsbergen (1979a).

Terminological note:

Although, following conventional practice, the term 'document weighting function' has been used up to this point in the thesis, it can with hindsight now be seen to be an unsatisfactory one. Since it places an unreasonable emphasis on the cardinal values employed to order (usually weakly) the elementary logical components of the terms comprising a query, rather than that ordering operation itself, another term seems preferable. Such a term should capture the essential thought that the ordering (strong or weak) of the elementary conjuncts of query terms serves to weakly order the member records of a data base or a subset of it. The author will henceforth use the term 'document ordering function' for this purpose, abbreviated to DOF. The functions designated Z_S, Z_A and $Z_{S \setminus A}$ are examples of DOFs, the latter two pertaining of course to subsets of the data base. When the emphasis is on the analytical expression used to achieve such mappings, it would seem sensible

to refer to a 'document ordering expression', abbreviated to DOE.

We will in future, restrict the meaning of the term 'retrieval process'. This has been used previously in an intuitive way, but will henceforth be used to denote the triple: partitioned data base, query in set form, and DOE. Each such triple fully determines the functions $f_1(z)_j$ and $f_2(z)_j$. This is not to say that a pair of the latter functions is uniquely determined: several retrieval processes may, in principle, determine the same pair of distributions. but since we will always be referring to a pair of distributions in conjunction with such a triple, we can say that a retrieval process refers to either the triple or the pair of distributions entailed by it.

3.3.2.4 Hypothesised invariance in $f_1(j)$ and $f_2(j)$

Since, as we have argued, the random variables of interest in the extended Swetsian formalism are Z_S, Z_A and $Z_{S \setminus A}$, and not Z''_S, Z''_A and $Z''_{S \setminus A}$, one can abandon interest in the cardinal values that the latter functions determine. We can, accordingly, describe the induced distributions of interest simply as $f_1(j)$ and $f_2(j)$, $j \in J$; not, as previously, as $f_1(z)_j$ and $f_2(z)_j$. The pair of functions $f_1(j)$ and $f_2(j)$, $j \in J$, constitute the natural 'unit of observation' in the extended formalism.

It is naturally of interest to know to what extent these paired functions, determined by each retrieval process, are stable or 'invariant'. In order to put this question more completely, we need to classify the entities that determine $f_1(j)$ and $f_2(j)$. They are as follows:

(1) The data base,

(2) The method used to partition the data base by instances of information need, (implying a value for Generality);

(3) The question (as SFQ):

  (a) The number of terms comprising it,

  (b) the identity of the terms comprising it*,

(4) The choice of DOE.

Such knowledge should provide a basis for strategies aimed at optimising the retrieval process, as well as providing scientific knowledge in its own right. Since our particular interest is in the question: 'How should we determine sequences of logical

* The specificity of a term depends on its identity, as does the extent to which it clusters with other terms in S, A or $S \setminus A$.

search expressions, E_1, E_2, E_3,..., E_J, such that net retrieval effectiveness (as expressed by some criterion measure) is maximised?', hypotheses of very considerable interest are those that determine such sequences. Two examples illustrate this point. We may choose as the criterion measure: mean distance from the origin to the members of the Recall-Precision graph, where the latter is determined by a pair $f_i(j)$. Then a hypothesis that two instances of document ordering expression are ordered by this criterion, i.e. $DOE_i \precsim DOE_j$, is of immediate practical use. One should choose a sequence of logical search expressions determined by DOE_j. Such a hypothesis is, of course, expressible in a strong form - without reservations as to choice of data base, Generality etc. - or in a more qualified form. A second example could be that if the user chooses, for the query, search terms of 'Type I' say, the criterion measure is inferior/superior to that for queries formed from search terms of 'Type II'. The two types of search term are ordered by the hypothesis. In the case of each example, the extended formalism allows the hypothesis to be stated clearly. It provides a logical framework facilitating the statement of particular assertions.

Three further basic points are noted here. First, an experiment can either generate hypotheses or evaluate hypotheses. For example, we can specify factors (1), (3) and (4) above, but vary factor (2). The latter variation will then determine or generate values for, say, one of the moments of $f_1(j)$ or $f_2(j)$, e.g. the mean value of j for $f_1(j)$. The distribution of the latter statistic, in a sample of pairs (f_1, f_2), can be used to infer the nature of the distribution of this statistic in a population

of pairs (f_1, f_2). The latter population is defined by the sample being examined being stated to be a random sample of it. Rather than identify the statistic with a moment of one or other f_i, we could instead identify it with Brookes's measure S, which is determined by both f_1 and f_2, or again with the mean value of D(A,B) say, for $j \in J$, for each pair (f_1, f_2). If, on the other hand, our interest is in evaluating population hypotheses, then these need to be specified a priori, i.e. they are supplied as guesses or fictions, or else on the basis of previous hypothesis-generating experimental work. The problem then is to see whether an interval defined by the population mean and some multiple of the standard error, for some specified statistic and level of confidence, can accommodate the value of the statistic obtained experimentally. The standard error of the statistic of interest may, however, not be known. (This is the case with the mean value of D(A,B), $j \in J$. for each pair (f_1,f_2), and the value of S, for example.) To see the problem of 'attaching meaning to experimental results' in either of these terms, i.e. in terms of formal statistical inference procedures, is however rather artificial, in the writer's view. For no concrete (i.e. non-statistical) meaning can be attached to the populations of (f_1,f_2) pairs to which one refers. The population of partitionings of a data base by instances of information need cannot be specified with certainty. There is no way of 'enumerating' all possible partitionings, even though it could be said to be some subset of 2^S. Accordingly, in the writer's view, there must be some scepticism directed at the use of formal inference procedures in analysing (f_1,f_2) data. To some extent, an intuitive appreciation of the variability within

samples is called for. Certainly, the main need at the present time is for hypothesis generation, not hypothesis evaluation, since no experimental data in the extended formalism have been reported.

A second basic point is that when we refer to 'modelling distributions' we can mean either distributions in a population from which the sample is (contestably) drawn, or we can mean a sample distribution of analytical form serving as an immediate object of comparison with an observed distribution. (Swets's concern was with the latter. The relevance of the former has been pointed out by Robertson (1975).) For example, we can test the assertion that an observed distribution $f_1(j)$ relates to a sample which has been drawn from a population of relevant documents distributed binomially over 1,2,3, ..., J. The binomial 'model' here is of the former type. On the other hand, we can say that the observed distribution $f_1(j)$ is such that another distribution of that same sample of relevant documents is binomial - and then compare 'observed' with 'asserted' without reference to any fictitious population*. A choice between these approaches to modelling needs to be made. Obviously the pure statistical approach allows levels of confidence to be attached to comparisons, but it does so at the price of invoking what is, arguably, a meaningless population. (The documents relevant to the particular information need which has produced f_1 and f_2 are all the documents so relevant.)

The third basic point is that stability or invariance in f_1 and f_2 cannot be demonstrated by confounding the distributions

---

* This philosophy of modelling, with no explicit referencing of the statistical paradigm, is the usual one in say engineering or physics.

f_1 for a set of retrieval processes, doing the same for f_2, and then portraying the variation in measure of effectiveness for these 'confounded' distributions. Swets did just this, in portraying confounded data on a ROC graph, apparently implying that the individual f_i distributions (in his case $f_1(z)$ and $f_2(z)$) were somehow fixed. In view of the ambiguities in Swets's presentation, this may not have been intended, however, although it is the writer's interpretation of Swets's statement:

> "... the assumption that a real retrieval system has a constant effectiveness, independent of the various forms of queries it will handle [i.e. different sets of relevant documents (author)] is open to question. It seems plausible, however, that the sharpness of the retrieval system's query language, and its depth of indexing, and also the heterogeneity of items in store, will determine a level of effectiveness that is relatively invariant over changes in the form of the query." (Swets, 1963: 248)

In view of the above considerations, and preceding arguments, any experimental investigation of data expressed within the Swetsian formalism needs to abide by the following rules:

(1) Hypotheses concerning f_1 and f_2 should clearly state which of these two distributions is involved, or that both are involved, and which moment (or other property) is involved.

(2) If the hypothesis describes the degree of match between an observed property and a modelled property (e.g. the

mean Euclidean distance between (R,P) and (R_m,P_m) points), this must be clearly stated. Such hypotheses are different in kind to those solely describing observed data.

(3) When distributions modelling individual f_i distributions are used, decisions must be made as to whether the event $W(\emptyset)$ is to be within the scope of the modelling function, and likewise whether CAI events are to be within scope (Section 3.3.2.1).

(4) The manner of specifying the partitioning of the data base by information need, the DOE, and the manner of specifying the query (as SFQ), need to be specified.

Neither Swets nor any other worker has put forward such rules, nor put forward hypotheses consistent with the formalism advanced in this thesis.

Lastly, we stress a fundamental difficulty to be encountered in any experimental program investigating $f_1(j)$ and $f_2(j)$ pairs. This is that whereas, by some means or other, the query can be arrived at in an algorithmic way in an experiment, this is not the case with operational retrieval practice. In an experiment the set of relevant documents is known, and the query can devolve from that, but this is not so in practice. The only reasonable stance, in the writer's view, is that experimental findings on f_1 and f_2, based on algorithmically derived queries, should be taken as portraying what retrieval practice is capable of, if so implemented. This is however not to say that algorithmically-generated queries are always superior to intuitively-arrived at queries. Only suitable experimental research can answer this point.

3.3.2.5 The R vs F and R vs P graphs

The treatment of the Precision vs Recall graph by Swets was relatively limited, compared with his treatment of the Recall vs Fallout (ROC) graph. There was no mention of this graph in the 1963 paper. The 1969 paper only (1) defined Precision in terms of the 2X2 table, and (2) provided a diagram of the P vs R relationship that was implied by a given retrieval process (to use our terminology) and a varying acceptance criterion. The diagram concerned, Swets's Figure 2, was however not derived from an individual experimentally-obtained process, the caption referring to the graph being "Idealised example of empirical recall-precision curve, fanned out by varying the acceptance criterion." Swets nowhere notes the standard relationship linking the two effectiveness measures he is most concerned with, namely Recall and Fallout, with Precision, namely $P=GR/[GR+(1-G)F]$.

The latter relationship was introduced into the context of the Swetsian formalism by the writer, in both the original, continuous formalism, and the extended, discrete formalism (Heine, 1973a, 1974). The writer has also introduced the relationship between the Recall-Precision graph, and the extended formalism, as described earlier in Section 3.3.2.3. Both the Recall-Fallout graph and the Recall-Precision graph are describable as sets of ordered pairs $\{(R(j), F(j))\}$, and $\{(R(j), P(j)\}$, $j \in J$, where j serves to number the order of the logical expressions e_i (described in Section 3.3.2.3) determined by a chosen DOE. (It is recalled that a DOE both creates the expressions e_i and orders them.) That these two graphs, and similarly-defined graphs, depend only on the (weak) ordering of these expressions and not on the cardinal

values that imply that ordering is a basic implication of the extended formalism. To the author's knowledge, the action of 'equivalence' between different DOEs in these terms, has not been clearly stated previously. In effect, we are saying that all possible analytical expressions serving to map the members of L_Q to the real numbers, although infinite in number, can produce only a finite number of Recall-Precision graphs (or of other similar types of graph). Further to the discussion of Section 3.3.2.3, and Appendix A, an upper bound to the maximum number of distinct R-P graphs can be found. Any analytical expression mapping the set L_Q to the real numbers can be identified with a composition of the elementary logical conjuncts of the query terms, and hence with one of the distinct R-P graphs.

Of some interest is the question of whether $R(J_c)$ can increase as $P(J_c)$ increases as J_c takes values: J, J-1, J-2, ..., 1. A investigation within the continuous formalism has been reported by the writer (Heine, 1973a), and was extended by Bookstein to a discrete formalism (Bookstein, 1974, 1977). Whether in practice this is possible will of course depend on the specific information need, query (as SFQ), data base, and choice of DOE. In the notation we use here, Bookstein's result was that R and P will both increase if:

$$\frac{R(J_c)}{F(J_c)} < \frac{f_1(J_c)}{f_2(J_c)}$$

(Bookstein defined Fallout and Recall slightly differently, these being the sums of the f_i values for $z \geqslant z_c$, not the sums for $z > z_c$ as we have defined them.) Bookstein proved also that when

f_1 and f_2 are both Poisson, i.e. when

$$f_i(j) = \frac{e^{-\lambda_i} \cdot \lambda_i^j}{j!} \quad ; \quad i = 1,2 \quad , \quad j = 0,1,2, \ \ldots$$

and when $\lambda_1 > \lambda_2$, the inequality described is never satisfied. In fact it can be proved that this is also true when the f_i are each binomial, i.e. when:

$$f_i(j) = \binom{J}{j} p_i^j (1-p)^{J-j} \quad ; \quad i=1,2, \quad j=0,1,2, \ \ldots, \ J$$

and when $p_1 > p_2$. This is perhaps a more meaningful result since J is in practice finite.

3.3.2.6 The bi- and multivariate receiver-value formalism

A way of extending the Swetsian formalism so that more complex signal-receivers are recognised will now be defined. We may imagine a retrieval process in which two queries are compared with each document, and mapped to the real numbers by the same number of functions. For example, a specific information need, represented in the data-base by a set of documents A, is associated with the two queries:

$$Q_a = \{a_1, a_2, a_3, \ldots\}$$
$$Q_b = \{b_1, b_2, b_3, \ldots\}$$

where a and b denote attributes of different character. Then two functions, say Z_a and Z_b, will map each document to two values, say z_a and z_b. Retrieval of documents can then be effected in two distinct ways:

(1) We can retrieve all documents such that $z_a > (z_c)_a$, where $(z_c)_a$ denotes a threshold value in the outcome space determined by $Z_a = W_a \circ Y_a$; and then retrieve a subset of documents such that $z_b > (z_c)_b$, where $(z_c)_b$ is a threshold in the outcome space determined by $Z_b = W_b \circ Y_b$. (These steps could be carried out in reverse sequence, the sets of retrieved documents being necessarily the same in the two cases, since simple set-intersection is involved.)

(2) We can form a further real value $z_{ab} = \psi(z_a, z_b)$ by mapping the pair of values (z_a, z_b) pertaining to each document, to the real line, and retrieve all documents having z_{ab} values greater than some threshold $(z_c)_{ab}$.

The above description can be generalised to cover the case of three

or more output spaces, and reworded to take only the order of the z value into account. There is considerable scope for purely mathematical research here, especially in respect of the (related) questions: What Recall-Precision graphs result from the two methods, for various distributions $f_i(z_a)$ and $f'_i(z_b)$? and: How can the R vs P graphs be optimised through variation in γ for given S, A, Q_a and Q_b? We shall describe one application of the bivariate formalism (using continuous modelling functions) in a later section (Section 3.3.3.2).

Lastly, we draw attention to the structural difference between introducing a bivariate extension to the formalism in the way just described (based on multiple receiver response), and extending the formalism through defining 'signal' in two or more ways as introduced by Hutchinson. As remarked in Section 3.3.2.2, the effect of Hutchinson's extension is that the distributions $f_i(z)$ are replaced by conditional distributions $f_i(z \mid A_j)$, where the A_j denote sets of relevant documents defined subject to different personal criteria of relevance of the document to a given need. Although Hutchinson saw the sets A_j as defined by notions of "degrees of relevance", they can be defined in a less quantitative way as just indicated.

A recognition that extensions of both types are legitimate would open the way for a more general, unifying extension of the Swetsian formalism. This is however not attempted in this thesis.

3.3.3 Applications of the extended formalism

In this section we describe three applications of the extended formalism. These relate to (1) the generation of a probability distribution, for all possible logical search expressions, over the Recall-Precision graph, (2) The incorporation of a quantitative variable, the age of the document, into the DOE, and (3) the generation of a suitable DOE when the retrieval process incorporates feedback from the enquirer as to the relevance or non-relevance of trial-retrieved documents. Of these applications, the first and third are treated on the basis of the extended, discrete formalism as introduced by the writer. The second application is in the language of the original, continuous formalism of Swets, but extended so as to cover bivariate weighting. As such it pertains only to pertinent documents as we have defined them.

Before discussing the three areas in detail, the author reminds the reader that the most basic application of the extended formalism is the 'obvious one'. Namely, that using the formalism it should be possible to identify optimum DOEs, using some convincing experiment. Once an optimum weighting expression has been identified, this is immediately applicable to the problem of generating optimum logical search expressions, i.e. it will generate an optimum sequence of logical expressions (such as were denoted by E_j in Section 3.3.2.3) paired to increasing Recall values.

3.3.3.1 The generation of the probability distribution, for all possible logical expressions, over the Recall-Precision graph

First we remind the reader that the term 'query' (or 'question') has been ambiguously used in the information retrieval literature. In general it has meant either (1) a verbal statement against which an article of relevance is asked to judge the relevance of individual documents (as in the Cranfield or Aberystwyth experiments, or (2) an artefact serving to probe a data base. We have previously argued that the procedure of '(1)' is unsound, i.e. that the phrase 'relevance to a question' is meaningless. Accordingly we restrict our usage of the term query/question to the second usage, and in particular identify a query with a set of terms. Logical search expressions will be referred to here as just that, although this is contrary to some writers' usages. (Salton uses query as a general term to denote both set and logical-expression constructions (e.g. Salton, 1975a: Chap. 4), as does van Rijsbergen (e.g. 1979a: 96 and 106) although context usually makes the meaning quite specific. Heaps (1978) on the other hand uses query/question to denote just logical search expressions. A set of search terms by itself cannot, of course, identify any set of retrieved documents, since all documents in the data base are mappable to its power set.

The question addressed in this section is: 'If, for a given query $Q = \{t_a, t_b, \ldots, t_n\}$, a logical search expression is chosen randomly from the set of all possible such expressions, what is the probability of a given Recall-Precision outcome?'. We follow the notation of Section 3.3.2.3.

It is sufficient to note that any logical search expression,

the elementary propositions of which are the logical variables:

$$t_a^{i_a} \quad , \quad t_b^{i_b} \quad , \quad \ldots \quad , \quad t_n^{i_n} \quad ,$$

can be expressed as a disjunction of a combination of the elementary logical conjuncts: $t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n}$, $i_j = 1$ or 0. It follows that to generate all possible logical search expressions, all we need do is form the expressions that are disjunction of all the combinations of these elementary conjuncts. As remarked in Section 3.3.2.3, an arbitrary logical search expression E_K is associated with Recall and Fallout values obtained by summing individual r and f values, i.e.:

$$\text{for} \quad E_K = \bigvee_{k \in K} \left(t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n} \right)_k$$

we have

$$R_K = \sum_{k \in K} \left(r_{ab \ldots n}^{i_a i_b \ldots i_n} \right)_k \quad ; \quad F_K = \sum_{k \in K} \left(f_{ab \ldots n}^{i_a i_b \ldots i_n} \right)_k$$

A value for P_K follows immediately, once a value for G has been specified. Accordingly, the probability distribution over the Recall-Precision graph, i.e. over the space $[0,1] \times [0,1]$, is given by:

$$Pr(R=r, P=p) = \| \{ E_K \; ; \; R_K = r, \; P_K = p; \; G \} \| / \, 2^{2^n} \quad .$$

Figure 3.3.3.1-1 shows the nature of this surface for the following example of modelled distributions:

Example:

Assume $Q = \{t_a, t_b, t_c, t_d\}$, where the four terms have specificities:

$c_1 = 0.1$
$c_2 = 0.01$
$c_3 = 0.001$
$c_4 = 0.0001$

and assume the information need is such that G = 0.01.

Arrive at values of the $r^{i_a i_b i_c i_d}_{abcd}$ as follows. Assume that the probabilities of co-ordination level values 0,1,2,3,4 are distributed binomially for relevant documents, with mean 2.8. That is:

$$f_1(1) = 0.0081$$
$$f_1(2) = 0.0756$$
$$f_1(3) = 0.2646$$
$$f_1(4) = 0.4116$$
$$f_1(5) = 0.2401 .$$

Assume further that these values of $f_1(j)$ can be assigned to individual variables $r^{i_a i_b i_c i_d}_{abcd}$ by the following rule:

$$r \propto - \sum_{i_u \neq 0} \log_e(c_u) .$$

Thus to 'break up' $f_1(2)$ into its component probabilities we note

$$r^{0001}_{abcd} \propto -\log_e(0.0001) = 9.210$$
$$r^{0010}_{abcd} \propto -\log_e(0.0010) = 6.908$$
$$r^{0100}_{abcd} \propto -\log_e(0.0100) = 4.605$$
$$r^{1000}_{abcd} \propto -\log_e(0.1000) = 2.303$$

Dividing each r value by 23.026 (the sum of the values shown) and multiplying by $f_1(2)$ gives:

$$r^{0001}_{abcd} = 0.0302$$
$$r^{0010}_{abcd} = 0.0227$$
$$r^{0100}_{abcd} = 0.0151$$
$$r^{1000}_{abcd} = 0.0076 .$$

Similarly, we can break up $f_1(3)$ by means of:

$$r^{0011}_{abcd} = -(\log_e(0.0001) + \log(0.0010))$$

(etc.)

and so on.

The values of $s^{i_a i_b i_c i_d}_{abcd}$ on the other hand are assumed to be given by the rule:

$$\begin{cases} s^{i_a i_b i_c i_d}_{abcd} = \left(k \prod_{i_u \neq o} c_u \right)\left(1 - s^{0000}_{abcd} \right) \\ s^{0000}_{abcd} = 0.9800. \end{cases}$$

In other words, we assume that the spike of probability attaching to the event $Q \cap T_d = \emptyset$ has the value 0.98, and the residue of probability is distributed over the elementary conjuncts in a way proportional to the product of the specificities of the query terms but depending only on those query terms for which $t^{i_u}_u$ = TRUE. We also assume that the random variables $(X_2)_u$ and $(X_2)_v$ mapping non-relevant documents to the events TRUE or FALSE for terms t_u and t_v, are independent.

The above rules determine probabilities $r^{i_a i_b i_c i_d}_{abcd}$ and $s^{i_a i_b i_c i_d}_{abcd}$ as shown in Table 3.3.3.1-1. The probability distribution over the Recall-Precision graph, for the set of all possible Boolean expressions based on Q, is then as shown in Fig. 3.3.3.1-1. We note, incidentally, that for a data base of usual size, say 10^6 items, the s-probabilities would need to be rounded to one of the discrete values: $n/(10^6 - \|A\|)$, but this complication is not pursued here. Fig. 3.3.3.1-2 shows the marginal distributions for Recall and Precision.

	$r^{i_a i_b i_c i_d}_{abcd}$	$s^{i_a i_b i_c i_d}_{abcd}$
$t_a^o \wedge t_b^o \wedge t_c^o \wedge t_d^o$	0.0081	0.9800
$t_a^o \wedge t_b^o \wedge t_c^o \wedge t_d^1$	0.0302	1.768×10^{-5}
$t_a^o \wedge t_b^o \wedge t_c^1 \wedge t_d^o$	0.0227	1.768×10^{-4}
$t_a^o \wedge t_b^o \wedge t_c^1 \wedge t_d^1$	0.0617	1.768×10^{-7}
$t_a^o \wedge t_b^1 \wedge t_c^o \wedge t_d^o$	0.0151	1.768×10^{-3}
$t_a^o \wedge t_b^1 \wedge t_c^o \wedge t_d^1$	0.0529	1.768×10^{-7}
$t_a^o \wedge t_b^1 \wedge t_c^1 \wedge t_d^o$	0.0441	1.768×10^{-6}
$t_a^o \wedge t_b^1 \wedge t_c^1 \wedge t_d^1$	0.1235	1.768×10^{-10}
$t_a^1 \wedge t_b^o \wedge t_c^o \wedge t_d^o$	0.0076	1.768×10^{-2}
$t_a^1 \wedge t_b^o \wedge t_c^o \wedge t_d^1$	0.0441	1.768×10^{-6}
$t_a^1 \wedge t_b^o \wedge t_c^1 \wedge t_d^o$	0.0353	1.768×10^{-4}
$t_a^1 \wedge t_b^o \wedge t_c^1 \wedge t_d^1$	0.1098	1.768×10^{-9}
$t_a^1 \wedge t_b^1 \wedge t_c^o \wedge t_d^o$	0.0265	1.768×10^{-4}
$t_a^1 \wedge t_b^1 \wedge t_c^o \wedge t_d^1$	0.0960	1.768×10^{-8}
$t_a^1 \wedge t_b^1 \wedge t_c^1 \wedge t_d^o$	0.0823	1.768×10^{-7}
$t_a^1 \wedge t_b^1 \wedge t_c^1 \wedge t_d^1$	0.2401	1.768×10^{-11}

Table 3.3.3.1-1 Modelled probability distributions over the sets of relevant ($r^{i_a i_b i_c i_d}_{abcd}$) and non-relevant ($s^{i_a i_b i_c i_d}_{abcd}$) documents, with the events defined by TRUE values of the elementary logical conjuncts shown.

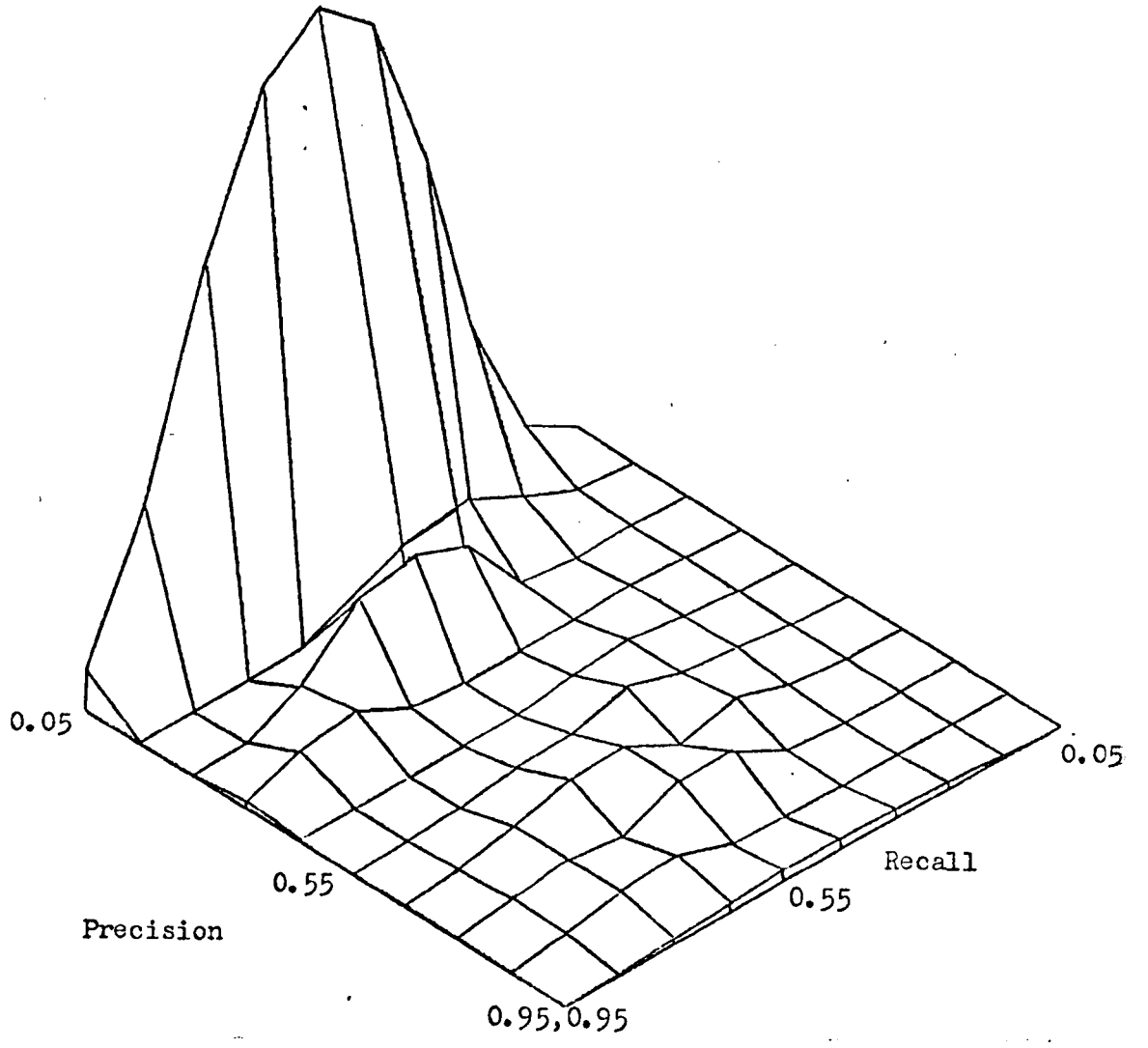


Fig. 3.3.3.1-1. The distribution of all possible Boolean search expressions over the Recall-Precision graph, for the data shown in Table 3.3.3.1-1.

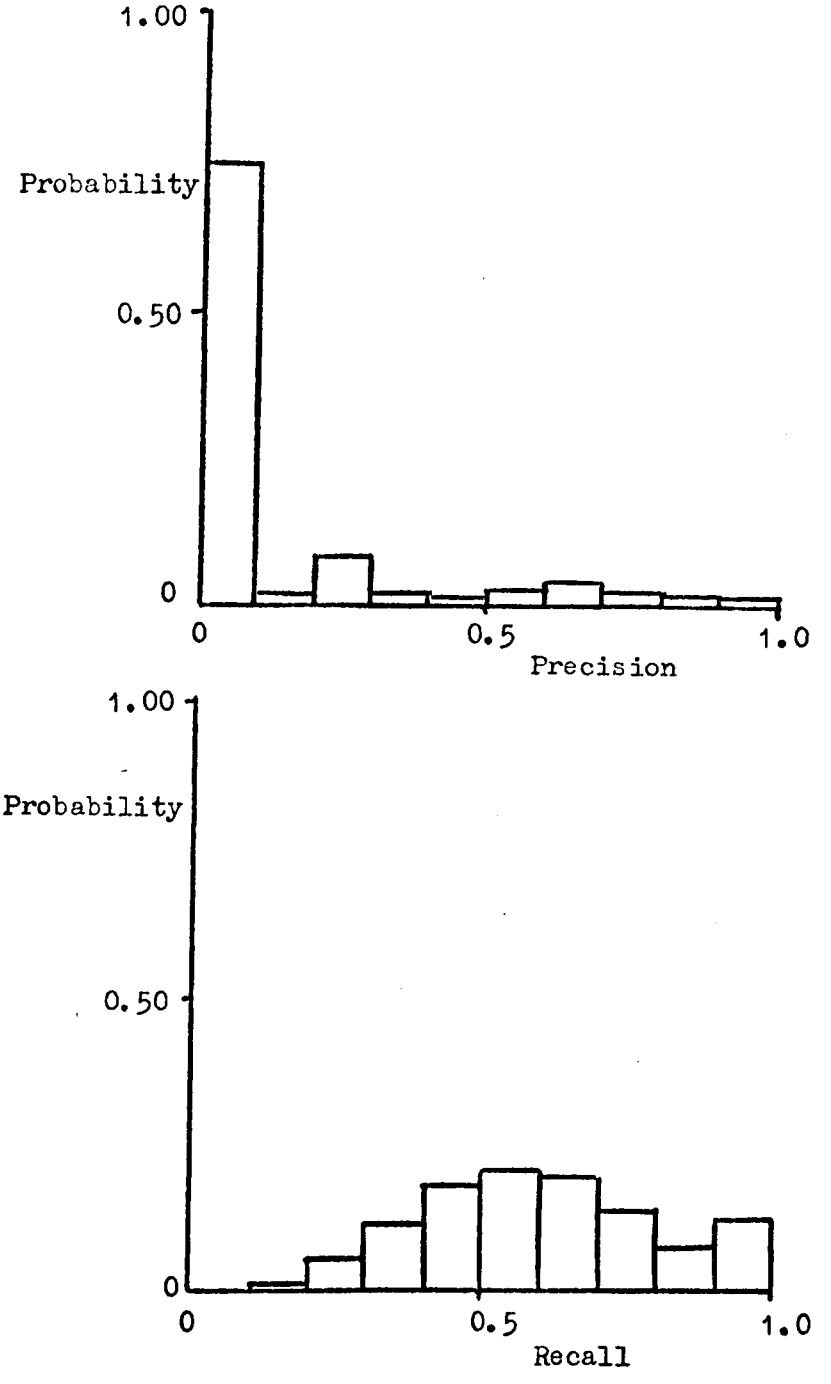


Fig. 3.3.3.1-2 The marginal distributions over the Recall and Precision intervals (0,1), for the surface shown in Fig. 3.3.3.1-1.

The relevance of the extended Swetsian formalism to the probability distribution just described and exemplified is threefold. First, it allows the distribution easily to be defined, i.e. to be defined by the probabilities that are central to the extended formalism. Secondly, it suggests how the probabilities can be modelled, by means of a probability tree based on one or two analytic induced distributions. (In the example, the analytic induced distribution was for relevant documents only, and was chosen to be the binomial distribution.) Thirdly, the operation of weakly ordering the elementary logical conjuncts, whether in an ad hoc way or by means of a prior mapping of them to document weights, is seen to have a clear purpose: one seeks to identify by such means a sequence of logical search
expressions that will give optimum search performance at the various levels of Recall (or Precision) that might be sought. The weak ordering, in effect, should be such as to give low ranks to inefficient search expressions and give high rank values to the more efficient search expressions by one or other such criterion, with the intention of choosing higher rank values first, of course.

3.3.3.2 The joint treatment of document semantics and document age

In Section 3.3.2.6 we extended Swets's basic univariate model of the retrieval process by introducing the concept of a bi- or multivariate formalism. This section applies that extension to a specific instance of the bivariate formalism. This is that of portraying document semantics (the subject of the document as represented by language, in particular word-attributes) and document age, as two variables that are capable, separately and jointly, of defining retrieval processes. We relate the age variable to the widely-researched phenomenon of document obsolescence, which is redefined for this purpose as a perceived, signed attribute of a partitioned data-base. In developing this application of the Swetsian formalism, we will keep to the original, continuous representation, in order to both simplify the mathematics, and to enable calculus methods to be employed. Two further remarks are needed to place our discussion in the context of formalism we have developed. These are: (1) Our approach will be a modelling approach, in the sense that specific analytical forms will be assumed to characterise the four main random variables involved. The point of this is to allow analytical methods to predict the general form of the R-P graph when certain retrieval strategies are followed. (2) For simplicity, the analytical forms chosen will ignore the spike of probability attaching to the event $W(\phi)$ for non-relevant documents and for a weighting function based on the linguistic similarity of document and query (as SFQ). We will thus be concerned only with pertinent documents as we have defined them. The effect of this is that the measures of retrieval effectiveness identified

of Section 3.3.2.1
in the discussion are of a 'conditional' character, as discussed in that section. We note also that the use of multivariate weights in information retrieval has been previously discussed by Di Fondi (1969) and Williams (1965). However, neither author discusses the use of variables of distinct characters: only language based variables are considered.

We start with the observation that the term 'literature obsolescence' has been ambiguously used to describe either or both of the following notions:

A. The information given in documents becomes less accurate or relevant, in some absolute sense, with increasing document age.

B. Document users behave as if 'A' were true.

For recent analytical discussion of related theory and experiment see Brookes (1970) and Meadows (1974: Chap. 5), and for a comprehensive critical review see Line et al (1973). The latter authors state, in agreement with the distinction made above:

> "It is most important... to be quite clear whether changes in library use... or in the value and interest of knowledge are being considered..." (p.318)

In the following, it is assumed that 'A' is meaningless in an objective sense, however meaningful it may be in subjective terms, and an amended form of 'B' is taken as the basis for the objective description of the obsolescence involved in information retrieval:

B'. Document users, through their behaviour, define a probability distribution on the ages of a set of documents the members of which they perceive to be relevant to their needs. (The behaviour is "asserting

such relevance".) Obsolescence exists when the mean of such a distribution differs significantly from the mean of the distribution on the ages of a set of documents from which the relevant documents are drawn.

It is also noted that (1) the interest in absolescence from an information retrieval standpoint is restricted to user behaviour of the 'relevance-assessment type', and not to behaviour in the form of (say) requests for documents, or citations in documents; (2) as such, the interest is restricted to obsolescence of the 'synchronous' type, to use Line's terminology; (3) the obsolescence of interest is <u>signed</u>: positive [negative] obsolescence being defined when the mean age of relevant documents is less [more] than the mean age of all documents; (4) in order for such obsolescence to be defined, a set of documents from which relevant documents are drawn is required to be defined at the time the enquiry is made.

Thus the viewpoint adopted here is that obsolescence in information retrieval is a dynamic observable entity, formally defined through probability distributions, and varying from information need to information need and user to user. Given behaviour in the form of an assertion as to the membership of a relevant set, i.e. a partitioned data-base, the existence and sign of obsolescence would be fully determined. In an operational information retrieval situation however, the relevant set is only perceived by the enquirer. In that case the existence and sign of the obsolescence could form part of the user's search profile, just as the user's perception of the terms assigned to relevant documents forms the main part of the profile. It is just this

perception that is of concern.

A pivotal assumption in the model to be described is that the user is competent to estimate the mean age of relevant documents. This assumption is justified, in the author's view, by (1) the consideration that there is no difference in principle between users' perceptions of this type of attribute and perceptions of other types; and (2) the observation that many information services implicitly recognise the usefulness of document age as an indicator of relevance. (This is evident in the practice of presenting retrieval output in a form in which references are ranked by age, or limiting SDI to recent material.) Presumably the human origins of such practices (which incidentally usually assume the existence of positive obsolescence in the enquirer's relevance judgements) lie in the tendency of newer documents to synthesise the information in older documents, the acceptability of the usefulness of the citation system for gaining access to older literature, and the 'entropic' tendency for older documents to have been seen previously by the enquirer, as well as the usefulness of date-of-publication as a convenient way of coping with natural redundancy in the literature.

Further complications that might be considered in building a complete theory based on index terms and document age jointly are: user/data base heuristics, the validity of the distributions assumed, and the consequences of users incorrectly estimating the true mean age of relevant documents. Obviously exploration of the last of these points would need to be accompanied by exploration of how accurately users choose words as attributes of the relevant documents that they seek.

We assume that to each document in a collection, two numbers are assigned by the 'detection device' (retrieval algorithm). One is a conventional linguistic or subject weight, x say, based on the degree of agreement of the set of words or word codes attached to the document, with the set of words taken to represent the user's enquiry, e.g. the cosine correlation measure of Salton. The other number is the age of the document at the time of the enquiry. This 'age weight' is denoted by t. As usual, the subscript i attached to a quantity or variable relates that item to the set of relevant documents when i=1, and to the non-relevant set when i=2. The assumptions about the distributions are as follows. First, the random variables x_i associated with x-values have probability mass functions which can be approximated* by Normal probability densities:

$$f_i(x) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp(-(x-\mu_i)^2/2\sigma_i^2), \quad x\in(-\infty,\infty). \qquad (1)$$

Secondly, it is assumed that the random variables T_i associated with the set of t-values for the entire set of documents in the discipline have mass functions that can be approximated by negative exponential densities:

$$g'_i(t) = \eta_i \exp(-\eta_i t), \qquad t\in[0,\infty). \qquad (2')$$

(Since it is only the broad effects of the two methods which are of interest, the 'double-exponential' model favoured by some authors is ignored.) In view of the fact that some operational retrieval systems do not allow users immediate access to all items in the data base, but only access to those with an age less than

---

* As previously described, the approximation is actually between the CDFs of $f_i(x)$ where x is continuous, and the CDFs of the discrete probability functions which are in fact observed. As usual, expressions (1) relate to single instances of need and enquiries, not to averaged data, as examined by Swets.

some specified data of incorporation in the data base, (2') is rewritten so that this fact is reflected in the assumed densities for T_1 and T_2. If the ages of documents that can be retrieved, for the retrieval system of interest, fall in the interval $[0,A]$, then expressions (2') become, after normalising:

$$g_i(t) = \eta_i/(1-\exp(-\eta_i A))\ \exp(-\eta_i t), \quad A>0,\ t\in[0,A]$$
$$= 0, \qquad t>A. \tag{2}$$

The mean values and variances of the X_i are:

$$E(X_i) = \mu_i \quad ; \quad V(X_i) = \sigma_i^2 \tag{3a}$$

and of the T_i are:

$$E(T_i) = \frac{\exp(-\eta_i A)(A + (1/\eta_i)) - (1/\eta_i)}{\exp(-\eta_i A) - 1}$$

$$V(T_i) = \frac{A^2 \exp(-\eta_i A}{\exp(-\eta_i A)-1} + E(T_i)\left(\frac{2}{\eta_i} - E(T_i)\right). \tag{3b}$$

As a third and final assumption it is accepted that X_1 and T_1, and X_2 and T_2, are independent; i.e. that there is no systematic tendency for older documents to have less or more index terms in common with a question profile than newer documents. Since document age and document subject (expressed by index terms) are such unlike concepts this seems plausible, notwithstanding that indexing terminology changes slowly with the development of knowledge.

Representing the joint densities of X_i and T_i by $h_i(x,t)$, we have, in view of the last assumption:

$$h_i(x,t) = f_i(x)\ g_i(t), \quad (i=1,2).$$

The effect of ignoring the small amount of covariance between X_i

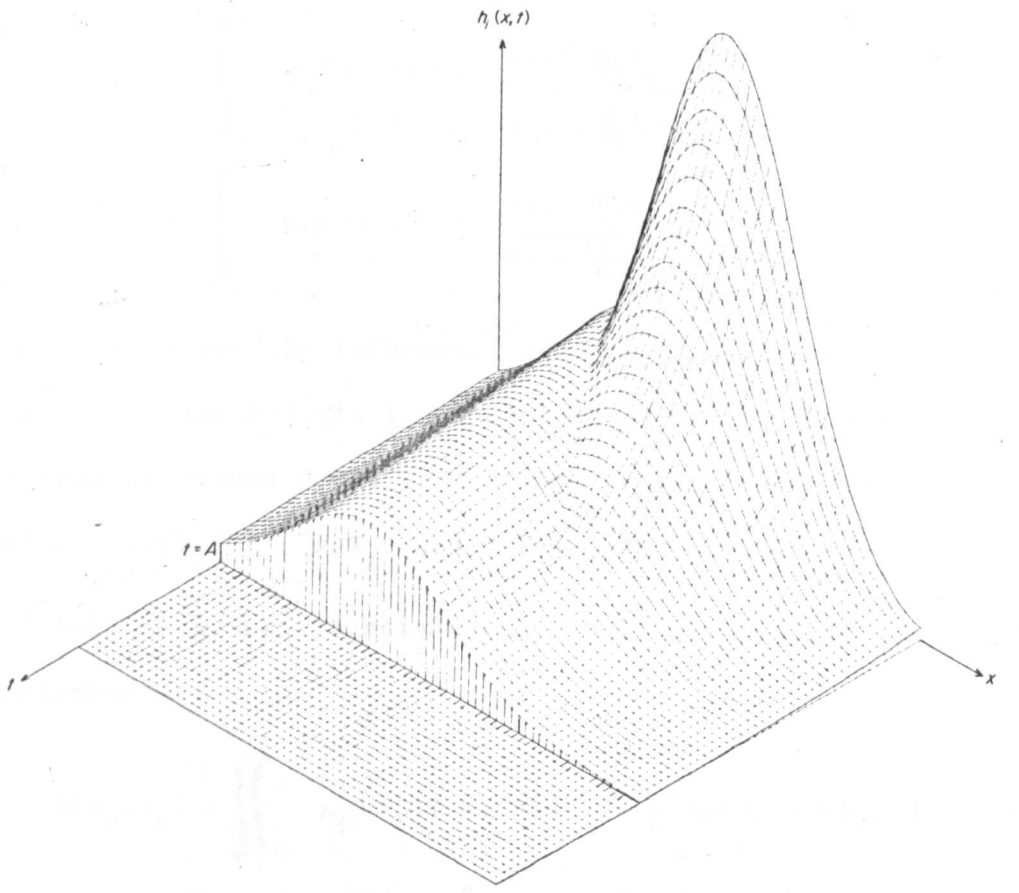


Fig. 3.3.3.2-1. The joint distributions of the modelling distribution pairs $h_i(x,t)$ described in the text.

and T_i which is likely to exist in practice (due to 'language following knowledge' will be that any theory building on this will describe an upper bound for retrieval effectiveness.

The functions $h_i(x,t)$ are illustrated in Figure 3.3.3.2-1. The marginals are the expression-pairs (1) and (2). For the 'age-of-document' variable t we have, for positive obsolescence:

$$R(t_c) = \int_0^{t_c} g_1(t)\,dt = \frac{\exp(-\eta_1 t_c)-1}{\exp(-\eta_1 A)-1}$$
$$F(t_c) = \int_0^{t_c} g_2(t)\,dt = \frac{\exp(-\eta_2 t_c)-1}{\exp(-\eta_2 A)-1} \qquad (6)$$

with P again readily inferred. For negative obsolescence one would retrieve documents with t values between t_c and A, when the Recall and Fallout values would be obtained from expressions (6) by writing $1-R(t_c)$ for $R(t_c)$ and $1-F(t_c)$ for $F(t_c)$.

For the subset technique the Recall will be, for positive obsolescence:

$$R(x_o,t_c) = \iint_C h_1(x,t)\,dx\,dt, \quad C=\{(x,t);\ x>x_c,\ t<t_c\}$$
$$= \int_{x_c}^{\infty} f_1(x)\,dx \int_0^{t_c} g_1(t)\,dt$$
$$= R(x_c)\ .\ R(t_c)$$
$$= \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{x_c-\mu_1}{\sqrt{2}\,\sigma_1}\right) \frac{\exp(-\eta_1 t_c)-1}{\exp(-\eta_1 A)-1} \qquad (7a)$$

using the independence assumption, and (5) and (6). Similarly:

$$F(x_c, x_t) = \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{x_c-\mu_2}{\sqrt{2}\sigma_2}\right)\cdot\left(\frac{\exp(-\eta_2 t_c)-1}{\exp(-\eta_2 A)-1}\right) . \qquad (7b)$$

Again, when negative obsolescence obtains, we write $1-R(t_c)$ for $R(t_c)$ in (7a), and $1-F(t_c)$ for $F(t_c)$ in (7b). We note that the same formulae for Recall and Fallout are given when the steps are reversed, i.e. when we first select documents by their t values, and then select a subset by their x values; a result that would not be true if X_i and T_i were not independent.

Turning to the bivariate weight technique, it is seen that before the Recall and Fallout curves predicted by the distributions h_i can be calculated, the forms of the probability density functions that describe the variation of random variables: $Z_i = \lambda_1 X_i + \lambda_2 T_i$, $(i=1,2)$ are required. The means and variances of the variates Z_i can be calculated by substituting (3a-b) in:

$$\begin{aligned} E(Z_i) &= \lambda_1 E(X_i) + \lambda_2 E(T_i) \\ V(Z_i) &= \lambda_1^2 V(X_i) + \lambda_2^2 V(T_i) + 0 . \end{aligned} \qquad (3c)$$

Call the density functions of the Z_i, $r_i(z)$. Then:

<u>either</u> $$r_i(z) = \int_{\frac{z-\lambda_2 A}{\lambda_1}}^{z/\lambda_1} h_i(x, s[x,s])\left|\frac{\partial s}{\partial z}\right| dx, \quad \text{for } \lambda_1,\ \lambda_2 > 0$$

<u>or</u> $$r_i(z) = \int_{z/\lambda_1}^{\frac{z-\lambda_2 A}{\lambda_1}} h_i(x, s[x,z])\left|\frac{\partial s}{\partial z}\right| dx, \quad \text{for } \lambda_1 > 0 > \lambda_2,$$

where $s(x,z)=t=(z-\lambda_1 x)/\lambda_2$; $\left|\partial s/\partial z\right| = 1/\lambda_2$; and $h_i(x, s[x,z])$ $= f_i(x)\ g_i([z-\lambda_1 x])/\lambda_2)$. The limits of integration correspond to the range of x-values when z=constant and t varies over its range (i.e. $t\in[0,A]$.) Note that $r_i(z)$ first case $\neq -\ r_i(z)$ second case . Such density functions must of course be positive. To obtain more tractable integrals, the variable of integration can be changed from x to $u=x-(z/\lambda_1)$, giving:

for $\lambda_1, \lambda_2 > 0$:

$$r_i(z) = \int_{-\lambda_2 A/\lambda_1}^{0} f_i\left(u + \frac{z}{\lambda_1}\right) g_i\left(\frac{-\lambda_1 u}{\lambda_2}\right)\left|\frac{1}{\lambda_2}\right| du$$

$$= k_i \int_{-\lambda_2 A/\lambda_1}^{0} \exp\left(-\left(\frac{\left[u + (z/\lambda_1) - \mu_1\right]^2}{2\sigma_i^2}\right)\right) \exp\left(\frac{\eta_i \lambda_1 u}{\lambda_2}\right) du$$

where $k_i \equiv \dfrac{\eta_i}{(2\pi\sigma_i^2)^{\frac{1}{2}}\left|\lambda_2\right|\ (1-\exp(-\eta_i A))}$;

or

$$r_i(z) = k_i \int_{-\lambda_2 A/\lambda_1}^{0} \exp\left[-(a_i u^2 + b_i u + c_i)\right] du, \quad \lambda_1, \lambda_2\ 0 \qquad (8a)$$

where $a_i \equiv 1/2\sigma_i^2$

$$b_i \equiv (\lambda_2 z - \mu_i \lambda_1 \lambda_2 - \sigma_i^2 \eta_i \lambda_1^2)/\lambda_1 \lambda_2 \sigma_i^2$$

$$c_i \equiv (z^2 - 2\mu_i \lambda_1 z + \mu_i^2 \lambda_1^2)/2\lambda_1^2 \sigma_i^2.$$

For the other case, where the age of the document is to be multiplied

by a negative quantity, the densities are:

$$r_i(z) = k_i \int_0^{-\lambda_2 A/\lambda_1} \exp\left[-(a_i u^2 + b_i u + c_i)\right]. \quad \lambda_1 > 0 > \lambda_2. \qquad (8b)$$

with a_i, b_i, c_i, k_i as above. When obsolescence is positive, the coefficients λ_1 and λ_2 should be chosen so that $\lambda_1 > 0 > \lambda_2$. In that event the effectiveness of retrieval using a weight $z = \lambda_1 x + \lambda_2 t$ is predicted using (8b). Conversely, when the user perceives negative obsolescence, both coefficients should be positive, and (8a) would be used. It is the ratio of λ_1 to λ_2 that determines retrieval effectiveness, so λ_1 can be conveniently put equal to 1. Since the values of a_i, b_i and c_i are in general not all positive, the integrals must be evaluated numerically. Once the variation of $r_i(z)$ with z has been so obtained, further numerical integration yields the Recall and Fallout values according to:

$$R(z_c) = \int_{z_c}^{\infty} r_1(z)\, dz; \qquad F(z_c) = \int_{z_c}^{\infty} r_2(z)\, dz \qquad (9)$$

for any specified threshold value z_c.

To evaluate language measures defined by the moments of the distributions involved, such as $\underset{\sim}{E}$ and $\underset{\sim}{S}$, (3a-c) are substituted in the definition. For $\underset{\sim}{S}$ we then have:

$$\underset{\sim}{S}_X = \frac{E(X_1)-E(X_2)}{(V(X_1)+V(X_2))^{\frac{1}{2}}} \; ; \quad \underset{\sim}{S}_T = \frac{E(T_1)-E(T_2)}{(V(T_1)+V(T_2))^{\frac{1}{2}}} \; ; \quad \underset{\sim}{S}_Z = \frac{E(Z_1)-E(Z_2)}{(V(Z_1)+V(Z_2))^{\frac{1}{2}}} \; . \qquad (10)$$

Brookes's measure, like Swets's measure, is a signed measure. For the usual linguistic weighting functions, both $\underset{\sim}{S}_X$ and $\underset{\sim}{S}_Z$ will be

positive. However $\underline{S}_T < 0$ [> 0] when positive [negative] obsolescence is perceived by the enquirer*.

Further developments of the work described here, including a simulation study, are reported by Heine (1977a), but are not repeated here in view of the concentration in this thesis on the discrete formalism.

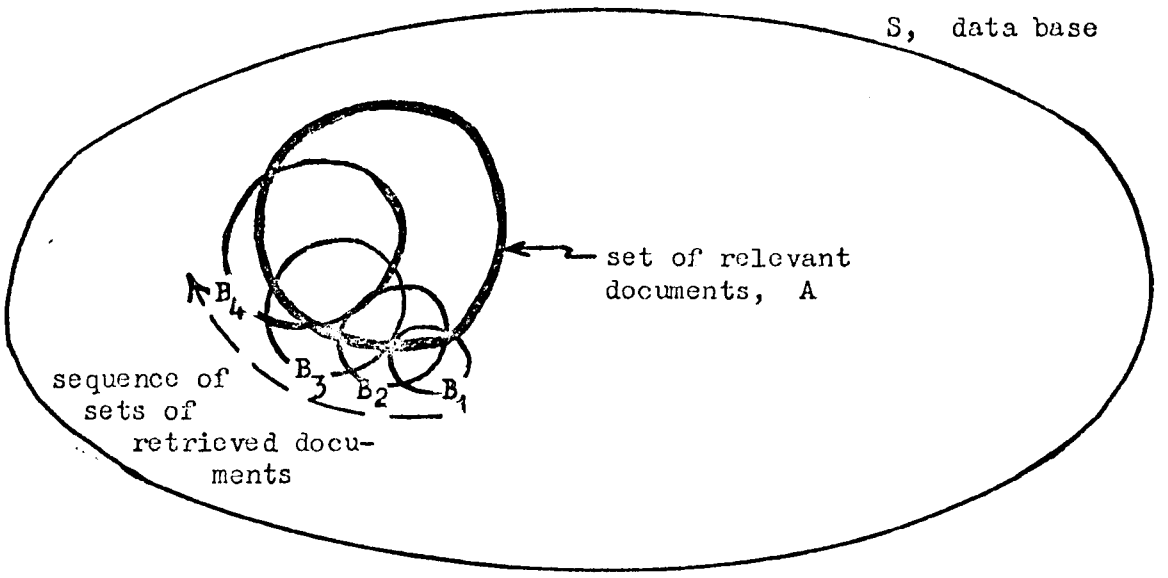
* It would be perfectly feasible to alter the convention, so that $\underline{S}_T$ is negative when obsolescence is negative, but this would go against our intuitive view that the usual form of obsolescence (embodied in the usage of the word) should be regarded as the positive one.

3.3.3.3 The heuristic retrieval process

One of the essential components of the Swetsian formalism is the question, as SFQ. We have repeatedly mentioned that this component is a 'variable' in the formalism, for a given data-base, and information need. Given a partitioned data base and a DOE, a query can be more or less effective, as measured by, say, the value of $\underline{S}$ determined by it being large, or by the R-P graph being displaced bodily away from the origin. We can imagine a process in which a sequence of queries, each as SFQ, is put to a data-base, and in response to each query an R-P graph is implied. Under operational conditions the Recall-values implied by a given query and a varying threshold value will not be known, although they may be known in a laboratory-like situation. Instead they are estimated values based on (1) the known number of relevant documents retrieved, and (2) an estimate, k, of the fraction of relevant documents retrieved. Calling this value, R', the Recall value perceived in any search will be:

$$R' = k \| B \cap A \| . \qquad \text{(k fixed for any query).}$$

If the sequence is appropriately chosen, there will be a steady improvement in the measure of effectiveness chosen. This could be said to describe a 'heuristic search' in relation to the information need prompting the search. This concept is illustrated below.



To avoid confusion, we note that a heuristic search procedure, as described above, is different in character from a procedure involving heuristics of relevance. In the latter case, one recognises that a user adapts his concept of what is relevant, i.e. adapts his information need, according to information received from those members of the body of documents actually retrieved and read. (In a variant of this, the adaptation is based on information received from attributes of documents, such titles, abstracts, authors' names etc., which may appear perhaps during an on-line search on a terminal's screen.)

The above two interpretations of 'heuristic search' are represented differently in the Swetsian formalism. Although the second interpretation can be represented simply by an analysis of the variation of successive retrieval processes, it in effect represents a development of the first interpretation. Accordingly,

we limit our discussion to the first interpretation, while noting in passing that the second defines an active research area, recently developed by Oddy (1974, 1977).

The literature on heuristic searching of the first kind is numerous. Salton (1975a: 472-83) offers a theory of "query space modification" that builds on the supposition that the differences between the mean values of the similarity-measures for relevant documents and non-relevant documents should be maximised, i.e. $E(Z_1)-E(Z_2)$ in our notation should be maximised. This in effect involves maximising the numerator of Swets's language measure E or of Brookes's language measure S, or (equivalently) Fisher's measure of the separation of two populations, $G = S^2$ (Fisher, 1936). In the author's view, this criterion is unsound in that the variances of neither Z_1 nor Z_2 are taken into account. Both the location and spread of the distributions induced by these random variables need to be considered, since both will affect the relative values of R and F, and hence will affect the R-P graph. Early work carried out at Cornell University on the problem by Ivie, Ide, Rocchio, Kerchner, Salton and others is reviewed by Salton. In Britain, work at UKCIS on "automatic profile construction" by Barker et al (1972) has attempted to solve the same problem but without accompanying theory, i.e. their approach was essentially empirical. More recently, Vernimb (1977) working for the Commission of the European Communities, has attempted to build a working system incorporating heuristic searching of the first kind, again on an *ad hoc* basis. Vernimb's approach apparently involved ranking terms by the ratio: frequency in set of relevant and retrieved documents/frequency in data-base, and

constructing what Vernimb calls "partial queries" and "loosened queries" as the refined queries; but the description of the methodology, like that given by Barker et al, is unclear and ambiguous.

Only the theoretical side will concern us further here. We have noted the work of Salton above. Van Rijsbergen (1979a: 106) has drawn attention to a valuable theorem of Nilsson (1965), the essential point of which is that an optimum query can be identified by a finite number of iterations of a heuristic process, irrespective of the starting point. Accordingly, the objective of heuristic searching is a realistic one, and the theoretical problems are (1) describing the heuristic search, and (2) optimising the search, so that (say) the number of steps taken to reach the optimum query is a minimum. We shall limit our discussion here to (1) a summary of the process from the point of view of the continuous Swetsian formalism, as offered by Yu and others; and (2) a description of a novel heuristic algorithm suggested by the discrete Swetsian formalism.

Yu et al's (1976) approach is based on the modification to a query given as follows, and was suggested by the work of Rocchio and Salton (Rocchio, 1965, 1966):

$$Q^* = Q + \alpha \sum_{d_i \in A \cap B} d_i - \beta \sum_{d_i \in (S \setminus A) \cap B} d_i .$$

Their notation has been altered slightly. Here Q is a query expressed as a vector of 0s and 1s defined over the (ordered) set of terms; i.e. it is a query in set form. B is the set of documents retrieved using a given measure of similarity (and, by

implication, a threshold value), and α and β are positive-valued parameters. d_i is a vector-representation of document d_i, again as a vector of 0s and 1s over the ordered set of terms. The measure of similarity chosen by Yu is in fact the inner (or dot) product of the two vectors. The objective of the approach is then stated as being:

> "to compare quantitatively the performance of Q and Q* and to investigate how the values of α and β ...affect the retrieval performance of Q*." (p.274)

The main results of their investigations, in the notation previously used in this work, were as follows. We first add the notation that Z_i and Z_i^* (i=1,2) are the random variables corresponding to Q and Q*, and that z_c^* is the threshold used when query Q* is used‡. z_c^* is chosen to be such that the same number of documents is retrieved when the new query, Q*, is used, i.e.

$$\|B(z_c^*)\| = \|A\| \sum_{z \geq z_c^*} h_1^*(z) + \|S \setminus A\| \sum_{z \geq z_c} h_2^*(z)$$

$$= \|A\| \sum_{z \geq z_c} h_1(z) + \|S \setminus A\| \sum_{z \geq z_c} h_2(z)$$

$$= \|B(z_c)\| .$$

Then:

(1) The Precision for Q*, for threshold z_c^*, is greater than or equal to the Precision for Q, for threshold z_c, if and only if:

$$(z_c - E(Z_1)) \sqrt{V(Z_1^*)/V(Z_1)} + E(Z_1^*) \geq (z_c - E(Z_2)) \sqrt{V(Z_2^*)/V(Z_2)} + E(Z_2^*).$$ (Theorem 3.1, p.277)

‡ Thus Z_1 and Z_2 correspond to our earlier notation Z'_A and $Z'_{S \setminus A}$.

(2) If α and β are sufficiently small, then the Precision of Q* is greater than the Precision of Q. (Theorem 3.4, p.279)

(3) The optimum values of α and β, which determine the best new query Q*, lie on a finite portion of a hyperbola. (Theorem 4.4, p.280).

In deriving these theorems, it is assumed that the distributions induced by Z_i and Z_i^* are (1) continuous, and (2) Normal. The limitations of this approach have been dealt with at length in Section 3.3.2.2, and we simply repeat here the essential point that the large value of probability attaching to the event $W(\phi)$, for documents in sets $S \setminus A$ and S, is disregarded in this approach. This in effect imposes a "conditional" character on the conclusions arrived at in Yu et al's approach. Their results would need modification if it could be established experimentally that $Pr(\{Z_2=z_0\})$, where $z_0=W(\phi)$, varied strongly from Q to Q*, for a given data-base.

We now describe a simple, novel approach to heuristic retrieval of the first kind. The approach represents an application of the discrete Swetsian formalism. It explicitly recognises term-dependence. Indeed the existence of pairwise dependence is the basis of the heuristic procedure, such dependencies being examined in both the set of relevant documents retrieved and the set of non-relevant documents retrieved. Our approach will be essentially that of applying linear discriminant analysis to Bernoulli random variables defined for each term in each partitioning of the retrieved set, the dependencies just defined appearing as covariances between these variables. First, we denote the set of retrieved documents by

W_3, choosing a different symbol from the customary one to emphasise that the retrieved set is a provisional one. We denote the set of relevant documents retrieved by W_1, and the set of non-relevant documents retrieved by W_2, so that $W_1 \cup W_2 = W_3$. Denote the set of attributes making up the query, Q, by $\{t_p\}$. Now denote the random variables associated with an attribute a_p and sets W_i $(i=1,2,3)$, by X_{pi}. Each of these random variables is a Bernoulli variable, since it maps each member of the set concerned to one of two values, commonly chosen to be 0 and 1. There are $3 \| Q \|$ Bernoulli random variables so defined. Thus:

$$X_{pi} = \begin{cases} 1 & \text{if } d \in W_i \text{ and } d \text{ is assigned attribute } t_p \\ 0 & \text{if } d \in W_i \text{ and } d \text{ is not assigned attribute } t_p. \end{cases}$$

The probability of the event $X_{pi}=1$ is accordingly just the fraction of the documents in W_i that are assigned attribute t_p. (X_{p3} is thus an estimate, based on the sample W_3 of the entire data-base, of the specificity of t_p, but the estimate may be a biased one.)

The problem we now set ourselves is this: What linear function of the random variables X_{p3} ($\| Q \|$ in number), itself a random variable, will yield a retrieved set more effective than the set W_3? To solve this problem define:

$$Z_i = \sum_{p=1}^{\| Q \|} \lambda_p X_{pi} \qquad (i=1,2,3),$$

the third of these random variables, Z_3, being the function intended to meet the objective just stated. In essence, the problem is to obtain coefficients λ_p that are 'most effective'. (We note in passing that when $\lambda_p=1$ (all p) the Z_i measure simple

level-of-coordination. Also that when Z_{pi} and X_{qi} are independent and identically distributed for all p and q, the variable Z_i is Binomial.) As the measure of effectiveness denoting maximum effectiveness, we choose Brookes's (language) measure $\underline{S}$, in its squared form: $\underline{S}^2$; i.e. our problem is taken to be finding values for the λ_p so that $\underline{S}^2$ is a maximum. In order to apply Fisher's discriminant analysis technique to this problem, we also define:

(1) h_{pi} is the parameter of X_{pi} (so that $E(X_{pi})=h_{pi}$, $V(X_{pi})=h_{pi}(1-h_{pi})$.)

(2) for two attributes t_p and t_q:

$$S_{pq} = \sum_{i=1}^{2} \| W_i \| \; E\left[(X_{pi}-E(X_{pi}))(X_{qi}-E(X_{qi}))\right] \qquad \text{(definition)}$$

$$= \sum_{i=1}^{2} \| W_i \| \; \mathrm{Cov}_i(X_{pi}, X_{qi}).$$

That is, S_{pq} is defined as a weighted sum of the covariances of random variables defined for the two attributes, and for the sets W_1 and W_2, the weights reflecting the relative sizes of the two subsets of W_3. (The subscript to Cov denotes: the set of documents (W_1 or W_2) that is the common domain of the two random variables concerned.) In particular we note:

$$S_{pp} = \sum_{i=1}^{2} \| W_i \| \cdot V(X_{pi})$$

$$= \sum_{i=1}^{2} \| W_i \| \cdot h_{pi}(1-h_{pi}).$$

(3) the numerical difference in the means of the two dis-

tributions induced by X_{p1} and X_{p2} is denoted by D_p, i.e.

$$D_p = \left| E(X_{p1}) - E(X_{p2}) \right| .$$

and the vector of such values for the $\| Q \|$ attributes is denoted:

$$\underline{D} = (D_1, D_2, \ldots, D_{\|Q\|}).$$

The linear discriminant analysis algorithm then entails defining the $\|Q\|$ X $\|Q\|$ matrix $(\lambda_p S_{pq})$, and solving the $\|Q\|$ component equations of:

$$(\lambda_p S_{pq}) = \underline{D}^T$$

for the values of λ_p.

To illustrate the application of the technique by an example, consider a retrieved set partitioned as follows. There are just four attributes involved. The 14 documents retrieved are arbitrarily labelled A-E (relevant documents retrieved) and F-N (non-relevant documents retrieved):

Relevant documents retrieved:

	A	B	C	D	E
X_{11}	0	1	0	1	0
X_{21}	1	0	1	1	0
X_{31}	0	1	1	0	1
X_{41}	0	0	1	0	1

$\| W_1 \| = 5$

Non-relevant documents retrieved:

	F	G	H	I	J	K	L	M	N
X_{12}	1	0	0	0	1	0	1	0	0
X_{22}	0	1	1	0	0	0	0	1	0
X_{32}	0	0	0	0	0	0	1	0	1
X_{42}	0	0	0	0	0	1	0	0	0

$\|W_2\| = 9$

There are eight Bernoulli random variables involved, four for each set. These correspond to the four attributes labelled 1,2,3,4 in the subscripts. For each random variable we can calculate an expectation, and for each pair of random variables in one of the two sets we can calculate a covariance. For example,

$$E(X_{21}) = 3/5 , \quad E(X_{41}) = 2/5$$

$$\begin{aligned} Cov_1(X_{21},X_{41}) &= E[(X_{21}-E(X_{21}))(X_{41}-E(X_{41}))] \\ &= E[(X_{21}-3/5)(X_{41}-2/5)] \\ &= 1/5[(0 - 3/5)(0 - 2/5) + (0 - 3/5)(1 - 2/5) + \\ &\qquad 2(1 - 3/5)(0 - 2/5) + (1 - 3/5)(1 - 2/5)] \\ &= -1/25. \end{aligned}$$

The calculation of covariances is made clearer if we represent frequencies of co-occurrences of events in a 2X2 table, for example (for the two random variables given above):

		X_{21} = 0	X_{21} = 1	
X_{41}	0	1	2	3
	1	1	1	2
		2	3	5

Similarly, $Cov_2(X_{22},X_{42}) = -1/27$. Accordingly:

$$S_{24} = S_{42} = 5(-1/25) + 9(-1/27) = -8/15.$$

The weighted sums of variances are similarly obtained. For example, since:

$$V(X_{21}) = 3/5(1 - 3/5) = 6/25, \quad V(X_{22}) = 3/9(1 - 3/9) = 2/9,$$

- we have:

$$S_{22} = 5(6/25) + 9(2/9) = 16/5.$$

Also, the D_p values are readily found:

$$e.g. \quad D_2 = \left| E(X_{21}) - E(X_{22}) \right| = \left| 3/5 - 3/9 \right| = 4/15.$$

In full, the constants required are as follows:

$S_{23}=S_{32}=-22/15$, $S_{14}=S_{41}=-17/15$, $S_{24}=S_{42}=-8/15$, $S_{34}=S_{43}=26/45$, $S_{12}=S_{21}=-6/5$, $S_{13}=S_{31}=2/15$, $S_{11}=16/5$, $S_{22}=16/5$, $S_{33}=124/45$, $S_{44}=94/45$. Also, $D_1=1/15$, $D_2=4/15$, $D_3=17/45$, $D_4=13/45$.

These are to be included in the equations:

$$\lambda_1 S_{11} + \lambda_2 S_{12} + \lambda_3 S_{13} + \lambda_4 S_{14} = D_1$$

$$\lambda_1 S_{21} + \lambda_2 S_{22} + \lambda_3 S_{23} + \lambda_4 S_{24} = D_2$$

$$\lambda_1 S_{31} + \lambda_2 S_{32} + \lambda_3 S_{33} + \lambda_4 S_{34} = D_3$$

$$\lambda_1 S_{41} + \lambda_2 S_{42} + \lambda_3 S_{43} + \lambda_4 S_{44} = D_4$$

yielding:

$$\begin{aligned} 48\,\lambda_1 - 18\,\lambda_2 + 2\,\lambda_3 - 17\,\lambda_4 &= 1 \\ -18\,\lambda_1 + 48\,\lambda_2 - 22\,\lambda_3 - 8\,\lambda_4 &= 4 \\ 6\,\lambda_1 - 66\,\lambda_2 + 124\,\lambda_3 + 26\,\lambda_4 &= 17 \\ -51\,\lambda_1 - 24\,\lambda_2 + 26\,\lambda_3 + 94\,\lambda_4 &= 13. \end{aligned}$$

Solving for the four unknowns gives: $\lambda_1=0.234$, $\lambda_2=0.329$, $\lambda_3=0.242$, $\lambda_4=0.283$. Assuming these values to be the best estimates of the comparable coefficients that would serve to discriminate the set of all relevant documents from the set of all non-relevant documents, the optimum weighting function attach-

ing to random events of the type $(X_{13}, X_{23}, X_{33}, X_{43})$, where each event is a random vector, will be:

$$Z_3 = 0.234X_{13} + 0.329X_{23} + 0.242X_{33} + 0.283X_{43}.$$

This random variable, Z_3, based on the question (as SFQ) $\{t_1, t_2, t_3, t_4\}$ may accordingly be regarded as an optimum one for this data-base and information used, and for the information obtained from the trial retrieved set. A second retrieved set would in general yield different values of the coefficients λ_i. Indeed, depending on how the attributes are identified as elements of the new question, the attributes themselves could be quite different. A possible way of identifying the four (say) best attributes would be to rank all attributes appearing in the retrieved set by $E(X_{i1})/E(X_{i2})$, and select the top four. Or information on their dependencies could also be taken into account, i.e. clustering information in selecting the question attributes. Document age might also be a useful discriminating variable. Only an experimental approach can be of positive use here.

Finally, we note that the above example suggests that Boolean search expressions should be generated from the query $Q = \{t_1, t_2, t_3, t_4\}$ as follows:

(1) Order the elementary logical conjuncts $t_1^{i_1} \wedge t_2^{i_2} \wedge t_3^{i_3} \wedge t_4^{i_4}$, $i_j = 0,1$, by means of the DOE:

$$Z = 0.234 \| Q \cap T_d \cap \{t_1\} \| + 0.329 \| Q \cap T_d \cap \{t_2\} \| +$$

$$+ \quad 0.242 \| Q \cap T_d \cap \{t_3\} \| + 0.283 \| Q \cap T_d \cap \{t_4\} \|.$$

(The expressions denoted $\| \ldots \|$ evaluate to 0 or 1.)

(2) Define a logical search expression by ORing together elementary conjuncts, from highest ordering elementary conjunct to some lower-ordered elementary conjunct, terminating the ordering according to the level of Recall required.

In practice, a more ad hoc searching style may be advantageous. For example, prompted by the findings of Reisig (1972:205), we could build into the search expression a requirement that the author should be one in a specified list of authors whose works it is known tend to be highly relevant.

3.3.4 Further, more marginally-related work

In this section we briefly discuss work of relevance to the three main sections above (3.3.1, 3.3.2, 3.3.3) but not already discussed in those sections.

The Swetsian formalism can be seen as just one formalism within which information retrieval can be described. A recent, extensive review of alternative formalisms has recently been offered by Robertson (1977a) and there is some reviewing content in van Rijsbergen et al (1980). In view of the scope of this dissertation, we only indicate where the main points of difference between the Swetsian and other formalisms appear to be.

The work of Salton (1968: Chapters 4 and 6.3), Zunde (1971) and Turski (1971) is primarily concerned with the relations obtaining in the Cartesian products of (1) the set of documents and the set of attributes (terms), (2) the set of attributes with itself, and (3) the set of documents with itself. Salton's work in this connection is largely concerned with applying matrix theory, coupled with the notion of document-document similarity, to the problem of improving queries (as SFQ), through "Linear associative retrieval". He claims (1968: 133) that the technique is not so satisfactory as the development of the query through the use of a thesaurus. Zunde's work, like Turski's, introduces relatively sophisticated novel concepts (e.g. "reference relation", "incitence relation", "co-incitence relation", "associated weighted graph", "intensity of co-incitence", "inward and outward degree" or "generalisation relation", but the complexity of the work appears to be to little effect in that no

applications of the theory of a non-trivial nature are suggested. This is due, in the opinion of this writer, to the failure of both authors to consider a partitioning of a set of documents by a notion of relevance to information need, as in the Swetsian formalism. The paper by Hillman (1964), although entitled "Two models for retrieval system design" does not in fact introduce two models as we have defined 'model', but instead compares two interpretations of Boolean algebra used in operational retrieval systems. We have already discussed at length the role of Boolean expressions in the Swetsian formalism, and little value is added to the Swetsian formalism by Hillman's discussion.

The formalisms of Gebhardt (1975) and Ludwig et al (1975), are also only marginally relevant. Gebhardt's contribution was to define a notion of continuous variability of relevance attaching to documents, this being represented as a random variable X_i, for document d_i, in his notation. (The mapping was of "jurors" (i.e. arbiters of relevance) to Re, for each document.) This approach has recognisable origins in earlier work (see for example Saracevic 1970b), and is similar in spirit to the work of the fuzzy-set theorists (e.g. Radecki 1976b, 1977) or Tahani (1976) and the later work of Hutchinson discussed earlier in Section 3.3.2.2. Unfortunately the presentation by Gebhardt is at times obscure, and his work tends to be immune from criticism as a result. He defines random variables Y_i which appear to be equivalent to variables Z_A and $Z_{S\setminus A}$ in our notation, the variability being created by the choice of question for a given need. This is a most useful idea, if correctly interpreted here, and legitimately extends the Swetsian formalism in a way reminiscent of Robertson's function

$\phi(d_i)$, but with the variation restricted to that caused by question-variation for a fixed need. Gebhardt defines a quantity: $G = \sum X_i Y_i \Big/ \sqrt{\sum X_i^2 Y_i^2}$, the summation being over all documents, and expressing the correlation between the random variables X_i and Y_i (his notation) for a given juror. This offers a measure of the effectiveness of the question for the need concerned - an example of a language measure in our terminology. But it is not established or apparent that relevance can be expressed as a cardinal-valued variable, and if this is not conceded the values of X_i, and hence the value of Gebhardt's measure G, are meaningless. It is also intuitively difficult to reconcile his approach with the Recall-Precision graph approach to effectiveness (to which the Swetsian formalism naturally leads) but this alone is of course not a criticism of his approach.

Ludwig and Glockman, on the other hand, attempt to model the (presumed) ordered set of documents in a data-base, for a given query (as SFQ) and need. (The term 'need' is however not used by them, and query may have been used by them as a synonym for need: we simply offer an interpretation of what they have written.) This ordered set may be viewed as a vector: (1,0,0,1,1,1,...) where '1' ['0'] denotes the relevance [non-relevance] of a document to the need. Their concern is to model the probability distribution appropriate to this, i.e. (for a vector of m items) to model the distribution: (1/m,0,0,1/m,1/m,1/m,...). To do this, indicative reference is made to standard results in combinatorial theory and the calculus of variations, and a particular analytical form is suggested. Their approach is reminiscent of W.S. Cooper's earlier work on "expected search length" (discussed here in

Section 3.3.1.1) which makes the weaker (i.e. superior) assumption that the data-base is only partially ordered. The Swetsian formalism arises implicitly in their work by their considering the distributions defined on the set of relevant documents and its complement, though neither Swets's work (nor Cooper's) is cited by them. Since the assumptions behind the probabilistic model they refer to are not stated, their approach appears to offer nothing to Swetsian theory at the moment.

Robertson has built on the work of Maron et al (1960) and Cooper (1977) in promoting the acceptance of what is termed the "probability ranking principle". This has been stated by Robertson as:

> "A reference retrieval system should rank the references in the collection in order of their probability of relevance to the request, or of usefulness to the user, or of satisfying the user." (Robertson, 1977b:294)

or by Cooper as:

> "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possibly on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data." (quoted by Robertson: 295)

The probability concerned, although not defined in either of the above two statements, is later denoted as "$\phi(d_i)$" and defined as:

"for any given document d_i ... $\phi(d_i)$=P(document relevant | document is d_i)=P(d_i is relevant)."

If the above definition is accepted, the formalism advanced by Robertson is unclear since a further variable γ defined by:

"γ =P(document relevant)"

is then identical to $\phi(d_i)$. Possibly d_i denotes a retrieved document (since Robertson states that ϕ corresponds closely to Precision) in which case $\phi(d_i)$ then denotes the probability that a document is relevant conditional on it being retrieved. The argument leads to the conclusion that documents in the collection should be ranked on the basis of the $\phi(d_i)$ value attached to each document. The present writer's objections to the principle are threefold:

(1) The principle is unclearly put. It is not clear whether $\phi(d_i)$ is (a) an estimate of the Precision of a particular set of documents that includes d_i - which appears to be implied by Robertson's comment (p.298) that "the proof relates only to a single request", or (b) an expected value of Precision for a set of sets of relevant documents, with queries related to each of them. The latter interpretation would seem to be implied, but Robertson does not clearly say that a random sample of sets of relevant documents, with queries attaching to each, is necessary in order to define $\phi(d_i)$, nor what form the queries take (e.g. if in set form, what the threshold values are), nor the algorithmic basis of query formulation for each set. The retrieved sets are not defined.

(2) The variable nature of 'question' for a given information need, is obscured.

(3) It is not clearly asserted whether, if $\phi(d_i)$ is an expected value for the Precision of a set of sets of relevant documents, each member set containing d_i, $\phi(d_i)$ is a biased estimator of the equivalent population value.

The 'principle' is, despite the above formal criticisms, potentially an intriguing hypothesis, and one that the Swetsian formalism can be reconciled with. However, variability in the content (i.e. document attributes) and form (set form, or logical search expression) of queries, as well as in the DOE chosen, and the threshold value chosen, creates very considerable ambiguity in its statement. It is falsifiable, but only when particular constructions are placed upon it.

Bookstein and Cooper have offered a "general mathematical model" which is in essence a less-general expression of the Swetsian formalism as we have described it. (Bookstein et al, 1976) A set of document records is assumed to be partitioned and weakly ordered by means of a function matching a request with each record. The matching is not necessarily into the real numbers but into "retrieval status values", equivalent to a mapping into the integers, as discussed earlier by the writer (Heine, 1975). Although Swets's work is not cited, Bookstein and Cooper mention it: it is said to "extend" the model to "include patron evaluations of the documents". The concern of these writers here is only with the partitioning of the data base as a whole by a retrieval mechanism, not, as in the Swetsian formalisms, with the partitionings of a set of relevant documents and its complement. This weakens the conceptual power of their model (i.e. formalism, as we have defined the term) very considerably and it is difficult to see why it is

claimed to be "general". Their not introducing partitionings of the data base by relevance judgements is consistent with their not incorporating retrieval effectiveness measurement in their formalism.

The work on term-dependence discussed earlier (Section 3.3.1.3) by, for example van Rijsbergen, Salton and others, is clearly related to the Swetsian formalism. If two documents are similar to each other, carrying say four terms in common, then they will also be similar to a question (as SFQ) if the question includes some or all of these four terms. Accordingly, clustering of documents is reflected in the probability values at the centre of the formalism. The essential difference from clustering theory appears to be that the Swetsian formalism portrays structural relations (the relations between documents and terms) not as nestings of sets of documents/terms but as distributions of partitioned sets on to an integer-numbered outcome space, i.e. as probability distributions, contingent both on a partitioning of the data-base and on an informative probe of the relations, namely the question as SFQ.

The work on signal-detection theory in psychophysics has been touched on in Section 3.1. We note again here that the overwhelming emphasis in this area is on the ROC-graph, usually described in terms of continuous variables. The Swetsian formalism in information retrieval, as we have described it, is concerned largely with the R-P graph on the other hand, and is expressed in terms of discrete variables - although continuous variables may be introduced to model these, as in Section 3.3.3.2. There appears to be considerable scope for the transfer of findings between the two areas, especially in regard to (1) the estimation of the R-P graph, on a

similar basis to the estimation of the ROC-graph, developing and continuing the work of Grey et al (1972) and Tague et al (1978) for example, and (2) in respect of the identification of individual processes analogous to the individual retrieval process that we have identified.

Other areas of thought are also relevant to the formalism and its application. The matter of question-formation, for example, is receiving increasing attention, for example in Belnap and Steel's monograph (Belnap, 1976) and in theoretical and experimental work by Kochen (1967b, 1974b). From the point of view of the Swetsian formalism, training users to "perceive" the optimum question is fundamental to the effective transfer of information by formal (system) means, and investigation of errors (or "illusions") in this perception is at least as important as the matter of finding optimum weighting functions; indeed, both variables should be jointly optimised. This matter has been largely ignored in the past simply because of the semantic confusion created by the early (and continued) use of the misleading phrase "relevance to a question", which we have criticised in Section 3.3.1.2. The theory of optimum medical diagnosis (e.g. Good (1971) and Victor (1974)) is obviously closely relevant (by analogy) to information retrieval theory, and it is hoped that the link with Good's work, already established by Robertson and van Rijsbergen, will further develop. The main benefit to Swetsian theory may possibly be in respect of the estimation of errors in the distributions f_1 and f_2, e.g. estimates of the standard errors in their moments for populations of retrieval processes. Again, there may be a reciprocal benefit: diagnostic accuracy being reinterpreted in terms of the extended formalism that we have introduced.

3.4 Summary, and final evaluation of published criticisms of Swets's theory

3.4.1 Summary of extended formalism

The discussion in preceding section has led us to the following viewpoint. The continuous formalism put forward by Swets can be, and for accuracy should be, replaced by a discrete formalism. The new formalism is predicated on the use, in operational information retrieval systems, of explicit or implicit logical search statements. The essence of the new, or extended, formalism, is as follows:

(1) a data base is regarded as partitioned, by an instance of an information need, into a set of documents relevant to that need, and a complementary set.

(2) a query $Q = \{ t_a, t_b, \ldots t_n \}$ defines a set of elementary logical conjuncts $L_Q = \{ t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n} \}$; $i_j = 0,1$; t_a^1 has value TRUE when t_a is assigned to a document . Each of these elementary conjuncts is associated with a probability pair (r,s), where r is the probability that a relevant document evaluates the conjunct to TRUE, and s the probability that a non-relevant document evaluates the conjunct to TRUE. An optimum ordering of the elementary conjuncts is by the ratio r/s, for a fixed query, data-base and information need.

(3) The members of L_Q are weakly ordered by a document ordering expression (DOE). The members of each subset of L_Q so defined are disjoined (i.e. ORed) to form a sequence of

composite logical search expressions $e_1, e_2, e_3, \ldots e_J$; $J \leq 2^n$. A DOE conventionally acts by mapping the members of L_Q to the real numbers.

(4) To each logical search expression e_i a probability pair (r_i', s_i') is attached, r_i' measuring the probability that a relevant document evaluates e_i to TRUE, and s_i' the probability that a non-relevant document evaluates e_i to TRUE.

(5) If the DOE is fixed, then J is fixed, and so is the sequence of logical search expressions that it determines:

$$E_j = \bigvee_{i=j}^{J} e_i \quad , \; j = 2,3,4, \ldots J,$$

where Recall, $R_j = \sum_{i=j}^{J} r_i'$ and Fallout, $F_j = \sum_{i=j}^{J} s_i'$,

for each E_j. Here R_j and F_j each increase monotonically, but P_j may not, as j takes the successive values:

J, J-1, J-2, ..., 1.

For a given information need, query Q, and DOE, an optimum ordering of the expressions e_i is determined by the ratio r_i / s_i. A DOE in practice thus serves to identify a suboptimum sequence of logical search expressions E_j. Even if the ordering of the E_j is 'optimum' in the sense that the e_i are ordered exactly as if ordered by r_i / s_i, the DOE will still be suboptimum, since the identity of the terms comprising Q can be varied.

(6) The set of ordered variables $\langle e_i, r_i, s_i \rangle$ defines what we have referred to as a 'retrieval process'. In effect this is a pair of induced probability distributions $f_1(j)$

and $f_2(j)$, with outcome space $j=1,2,3,\ldots,J$, and with each of the j-values labelling one of the composite logical expressions e_j. The properties of a retrieval process, for a given data base, information need and query, are not predicted by the formalism itself but must be investigated experimentally.

The motivation for undertaking experimental investigation of the properties of retrieval processes is the identification of means that lead to a joint optimum of Q and the DOE, i.e. of a means of identifying 'good search terms' and 'good means of generating logical search statements'. However in a sense the extended formalism, although accurate, is over-specific in that the composite expressions e_i do not require to be defined. A characterisation of information retrieval based simply on the distributions $f_1(j)$ and $f_2(j)$ over the <u>permutations</u> of the members of L_Q would suffice, i.e. for the case $J=2^n$. (This is so since the graph of the pairs of cumulative probabilities that we have denoted R and F for a strongly ordered set L_Q will contain (as a subset) all of the probability pairs (R,F) arising from a weakly ordered set L_Q.) But the more specific characterisation of information retrieval that we have offered is more general than this. It is also in furtherance of our basic objective of reconciling the description of information retrieval using explicit document weighting (the original "signal detection model") with the description of information retrieval using search logic.

With the continued development of retrieval software, the

pattern in the future will probably be as follows. An enquirer will specify a query as a set of terms (not as a logical expression) and in addition specify a desired level of Recall. The software (either within the data base management program, or else within an intelligent terminal) will then perform the tasks of generating the elementary logical conjuncts of the query, ordering them (i.e. defining a permutation of them), and inputting a disjunction of higher-order elementary conjuncts as a search expression. (Sophisticated software could form a Boolean minimization of this disjunction prior to its inputting as a search expression. The rationale of organization of the data base itself is another factor that could determine the form of the expression.) One's interest then would be in the induced distributions $f_i(j)$, $i=1,2$ for each permutation of the members of L_Q, i.e. for $J=2^n$, and of course in the sensitivity of the Recall-Precision graph to choice of permutation. If an algorithm was used to order the members of L_Q which only weakly ordered them, then this would be of interest to an enquirer only if the Recall-levels were insufficiently refined, i.e. if the 'jumps' in Recall from partition to partition of the set L_Q were such as not to conform with what was required. Improved software should, of course, also include a heuristic feature whereby feedback on the relevance/non-relevance of trial retrieved documents was used to amend (1) the identity and number of the terms comprising Q, and (2) the permutation of the members of L_Q (for the re-defined Q) used to generate the next logical search expression, as described in Section 3.3.3.3.

We have discussed briefly the use of functions that approximate the behaviour of $f_1(j)$ and $f_2(j)$, or their cumulative

sums, i.e. modelling functions in the strict sense. One position is that the introduction of such functions is optional. It is not essential or even desirable in the expression of the formalism.

A comparison, in tabular form, of the main features of the extended formalism with those of the original, continuous formalism of Swets, is as follows:

Original description of information retrieval by Swets	Amended description of information retrieval as given in thesis
1. Ambiguous in respect of being addressed to individual or grouped processes, but experimental results were based on compounded data.	1. Basic description addresses one individual information retrieval process.
2. Distributions in the formalism are continuous.	2. Distributions in the formalism are discrete.
3. Observed and modelling distributions tend to be confused in the formalism, though ROC graphs from observed data are discrete.	3. Observed and modelling functions are treated separately. Modelling is not seen as a necessary part of the theory, although of use in validating other 'deeper' theories.
4. Logical searching is not explicitly described.	4. Logical searching is the basis of the formalism, which is seen as basically addressing sequences of search expressions. Document weighting serves solely to order search expressions.
5. Variability in the information retrieval process for a fixed information need is not recognised.	5. Variability in the information retrieval process for a fixed information need is clearly related to a question set (Q) and to a compound function (WoY).
6. Not all records in the data base are brought within the scope of the formalism.	6. All the records in the data base are brought within the scope of the formalism.

3.4.2 Final Evaluation of published criticisms of Swets's theory

With an extension of Swets's formalism established, we are in a position to evaluate the published criticism of Swets's theory. (This could not have been offered in earlier reviewing discussion since the extension was not then established, although wherever possible we have referred earlier to published material relevant to specific aspects of the formalism e.g. the contributions of Bookstein (1974, 1977).). Since there appear to be only two analyses directed at the formalism as a whole, this can be succinctly done. The analyses are by van Rijsbergen (1979a: 158) and Robertson (1977a: 131). (Neither author offers a distinction between formalism and model qua approximating fraction introduced within the formalism, so that some interpretation of their remarks is still necessary.)

Van Rijsbergen questions the appropriateness of the "Swets model" on four grounds. He first challenges the statement that the linearity of the ROC graph implies that $f_1(z)$ and $f_2(z)$ are each Normally distributed. This is correct, but could be put in a stronger form. The exclusion of the 'spike' of probability attaching to the all-negated elementary conjunct for non-relevant documents by itself implies that $f_2(z)$ cannot be Normally distributed. (Van Rijsbergen does not point out that the event to which $t_a^{i_a} \wedge t_b^{i_b} \wedge \ldots \wedge t_n^{i_n}, \bigvee_{i_j=0 \text{ all } j}$ is mapped is excluded from the original formalism.) Secondly van Rijsbergen observes that the values of the DOE are not necessarily continuous, as assumed or at least implied by Swets. This is true, but it does not by itself argue against the use of continuous modelling functions, i.e. the use of

density functions with CDFs that approximate the CDFs of the observed discrete random variables, i.e. the CDFs of $f_i(j)$, i=1,2. Thirdly, he observes that it is not true that the distributions $f_i(z)$ are similarly distributed (let alone continuously, let alone Normally). This again is true - we have observed the 'spike' of probability for non-relevant documents attached to the all-negated elementary conjunct which alone renders the two induced distributions distinct from each other. Fourthly, van Rijsbergen criticises the concentration by Swets on R-F variation rather than R-P variation. The present thesis remedies this.

Robertson's review, although a valuable and full survey, and analytical in character, is not a formal, mathematical review - although the literature addressed is usually mathematical. Accordingly some of the concepts he uses are not entirely clear. The term 'request", used in such phrases as "request definition", "request independence" and "requests of different generality" is ambiguously used, but the last-given usage implies that he identifies a request with a data base partitioning, perhaps a verbal description of same. However if so he does not recognise that for a given (fixed) "request" (i.e. a fixed partitioning of a data base) <u>various</u> logical search expressions made up from <u>various</u> sets of query terms can be constructed. Robertson refers to Swets's theory as "highly developed" and comments, as we have done, on the similarity of Bookstein and Cooper's formalism and Swets's formalism. He also claims that the threshold value central to Swets's formalism was regarded by Swets as given prior to retrieval being effected, in agreement with the interpretation of the present writer. (This is of course not to say that Swets's formalism does

not describe the effects of allowing that threshold to vary - that is one of the major features of the formalism, of course). Robertson is also in agreement with the fundamental point (made by the present writer previously (Heine, 1973a, 1974)) that Swets's description of retrieval can be individuated down to the level of the specific partitioning, specific query (as set of terms), and specific DOE, i.e. to what we have referred to in this thesis as a 'retrieval process': The averaging (confounding) of data is not a necessary step in the statement of the signal detection formalism.

Robertson observes also that the need to have a DOE which produces output events ranked by the likelihood ratio (Bookstein, 1974) conflicts with "the common sense basis of most match functions." The present writer agrees with this: certainly likelihood ratio weighting is a posteriori in character, i.e. of no practical application except in providing an ideal DOE by which analytical DOEs may be judged. A further objection by Robertson is that the "parameters" of the distributions that we have labelled here $f_i(j)$, $i=1,2$; $j=1,2,\ldots,J$, are affected by tied ranks. We interpret this to mean that the moments of these distributions will vary with the value of J that the choice of DOE determines. (Otherwise, to refer to "parameters" is to confuse observed data with modelling data.) This, if a correct interpretation, seems certainly to be a valid observation, and it perhaps suggests that either (1) the values of the moments of $f_i(j)$ should be normalised in respect of J (e.g. by multiplying the moments by J_{max}/J, i.e. by $2^n/J$); or (2) our interest should be restricted to the 'standardised' DOEs which map documents into 2^n discrete events,

i.e. restricted to $f_i(j)$ defined over the permutations of the members of L_Q; or (3) our interest should be restricted to the Recall-Precision graphs that DOEs determine, i.e. that the distributions $f_i(j)$ should be seen as an 'intermediate mechanism' for producing R-P graphs and not treated as objects of study. We could of course also abandon all operational or experimental interest in the distributions $f_i(j)$ and the R-P graphs that they determine, and instead study the distribution of all possible logical search expressions over the Recall-Precision graph, as described in Section 3.3.3.1 here. But this would be to abandon interest in the procedure whereby (sub)-optimum logical expressions can be identified.

Robertson correctly points out an earlier error of the present writer - that the Central Limit Theorem determines that when data for different retrieval processes are confounded the distributions $f_i(z)_j$ can be approximated by Normal density functions. Lastly, the present writer points out that Robertson's discussion does not make clear, or even identify at all, the reality that the distributions $f_i(z)_j$ (or the distributions $f_i(j)$) will vary from query to query (with query as 'a set of terms') as well as weighting function to weighting function, for any given instance of information need. The discussion in this thesis, and especially the experimental results reported in the next section, both emphasise that this is so and attempt to identify the extent of the variability that is likely to obtain in practice.

In summary, the published, analytical criticism of Swets's contribution has been perceptive but has not identified what we claim to be the most fundamental weakness in the original formulation of it: the exclusion of logical searching.

4. AN EXPERIMENT TO GENERATE HYPOTHESES IN THE SWETSIAN FORMALISM

The preceding three main parts of this thesis have attempted to extend the formalism put forward by Swets, and to suggest areas of application of it. The applicability of such theory depends, however, on the extent to which valid hypotheses describing retrieval processes in such terms can be put forward. Such hypotheses would serve to describe the extent to which 'invariance' in retrieval processes exists, in relation to one or other variable characterising retrieval processes in the formalism. In this part we describe an experiment* on information retrieval the main aim of which was to identify such hypotheses (i.e. generate them) and assess their significance. Secondary aims of the experiment were (1) to develop an experimental methodology that other investigations might follow (and which could in part contest hypotheses generated here), and (2) to investigate the extent to which the observed distributions could be modelled using simple analytical expressions to generate distributions comparable with those observed.

The experiment sought primarily to describe the properties of a variety of retrieval processes, defined by allowing limited variation in (1) the set of relevant documents, (2) the DOE and (3) the algorithm for generating the query (in set form). On the other hand both the data base and the procedure whereby relevant documents were identified were fixed. The Generality of the processes varied over a wide range. In describing the retrieval processes involved, emphasis was given to the evaluation of a wide

---

* The term 'investigation' might be preferred to that of 'experiment', but the latter has been chosen, despite the exploratory nature of the work, in view of the definite constraints imposed.

variety of descriptive statistics, the variation of each of which (for various categories of retrieval process) defined the main hypotheses being generated. Correlations between these descriptive statistics were also examined selectively, with particular regard to their variation with Generality, these generating further subsidiary hypotheses. In this way objective descriptions of the retrieval processes were arrived at in the usual scientific manner. Analytical expressions serving to model the observed distributions were also introduced and evaluated using further descriptive statistics. Formal statistical 'inference' based on the variation in the descriptive statistics was fairly limited, in view of the sample design being 'exploratory' rather than 'random'. (The sample was deliberately designed to pick up more variation in the descriptive statistics than a random sample would and also, necessarily, had to be of a 'quota' character.) Moreover, the sampling distribution of some of the statistics used (especially in comparing observed and modelled distributions) are not known, limiting the scope of inference by formal means. It was felt also that evaluation of modelling distributions should be restricted to those modelling <u>individual</u> processes, at the present stage. The extent of the success of such individualised modelling, as well as the identification of areas of sensitivity in the basic descriptive statistics should, it was felt, (1) provide a basis for further experimental work building on the present work, and (2) suggest modelling distributions for populations from which <u>all</u> the sample distributions could be considered to have been drawn.

4.1 Experimental constraints

4.1.1 Definition of the data base to be examined

The experiment was undertaken on the MEDLARS data base, i.e. on the MEDLINE file and its related BACKFILEs, prepared by the U.S. National Library of Medicine. This data base dates from 1964, and is entirely machine-accessible although only the MEDLINE file of more recently included records is immediately accessible on-line.

4.1.2 Definition of the sets of relevant documents

The objections to defining sets of relevant documents on the basis of intuitive assessments of the linguistic similarity between a description of need (a "question" in the terminology of some earlier experimenters) and a document (or description of a document) have already been discussed (Section 3.3.1.2.) The main objection, to recapitulate, is that relevance is judged in an artificial situation, since the need that prompted the description of need is either (1) not known, and hence ambiguously surmised, or (2) directly known and hence providing an alternative criterion of relevance, again making the situation an ambiguous one. Accordingly the sets of documents so defined contain an unnecessary source of error, the extent of which is unknown. In place of this faulty method, it is asserted that relevance judgements must be sought via (and as) behavioural evidence of some appropriate type. The particular evidence chosen in the experiment being described here was that characterised by the sets of documents cited by review papers, in particular medical review papers. The 'literature review' is, by definition, a review of information in some particular area of knowledge. The material identified as significant by the reviewer (and cited by him) may accordingly be taken as defining an information need. The paper implies to its reader that the documents referred to are those that he would need were he to extend his reading (and his knowledge) by one more step. (Alternative behavioural evidence of relevance would be provided by identifying material citing the review paper though this seems a less convincing criterion.) It seems reasonable to assume, in the particular case where papers are review papers, that their authors are

sufficiently au fait with their subject, and its literature, that material not cited by them has been rejected positively by them, not rejected by default. Miller's "extension ratio' should be a minimum for such papers.

Four sampling frames of medical review papers were initially considered: Ulrich's Irregular Serials and Annuals, the British Library Lending Division's (formerly National Lending Library's) KWIC Index to Some of the Review Serials in the English Language Held at the N.L.L., Unesco's List of Annual Reviews of Progress in Science and Technology, and the sections headed "Bibliography of Medical Reviews" included in Index Medicus and Cumulated Index Medicus. The last item was in fact selected for various reasons, chiefly that it was a by-product of the most comprehensive secondary information service in existence in the medical field. The scope of the Bibliography of Medical Reviews is stated by the National Library of Medicine to be "articles which are well documented surveys of recent biomedical literature". The bibliography appearing in the February 1976 issue of Index Medicus was chosen as the sampling frame to be used. Of the 482 items included, 7 were rejected, 6 because insufficient data were quoted and one because it was a non-serial entry. The sample of review papers eventually chosen, prior to the identification of items cited by them as the sets of relevant documents of interest, was defined partly as a quota sample and partly subject to ad hoc criteria. (As mentioned earlier, it was felt that given the total absence of previous comparable data in the literature, a fairly flexible set of criteria should be used to identify areas of sensitivity and homogeneity.) The criteria used were as follows:

- That the review paper cited between 20 and 40 items, or more than 170 items, or was a member of a set of review papers published in one 1975 issue of a particular journal (Nephron). (The latter condition was provided in order to see if homogeneity in the information retrieval processes obtained when the information needs pertaining to them were those recognised by a group of one workers in the same field.)
- that the review paper itself could be obtained through the local library system by 1 May 1976.
- that the author of the review paper was able to reply to the author's correspondence with him by 25 July 1976.

Some review papers, while meeting the first criterion, were excluded because the author's address was not obtainable or the information sought in correspondence with authors was not supplied.

Each of the 31 review papers defined by using the above criteria had attached to it a list of cited literature. These lists were reduced so that only items in them that had been identified by Index Medicus after 1963 (and hence were in the MEDLARS data base) were included. This was done by manually checking Index Medicus. The authors of the papers were asked to supply two dates delimiting the scope of their review: the dates (of publication) after which and before which no document would have been eligible for mention by virtue of its date of publication alone. This step was necessary in order to measure the sizes of the sets of documents from which the sets of relevant documents were to be retrieved. (When the earlier delimiting date preceded the date of commencement of MEDLARS, the latter date prevailed of course.) The numbers of references included in the 31 sets of

relevant documents, together with the numbers of references in the review papers from which they were selected, the total number of references of the data base of eligible date, and the Generality of the implied information need, are given in Appendix B. Intuitively, it is seen that the retrieval problem of concern is that of identifying about 27 documents in a parent set of about 2'6 documents. The partitionings‡ labelled 3,9,15,17,20 and 27 relate to the six review papers that were included in one journal issue and were on a common medical subject as defined by the common attendance of the authors at a conference. These particular sets of relevant documents were rather larger than the others, the mean G value for them being 28.9'-6 (standard deviation (n-1):14.8'-6) as against a mean G of 9.2'-6 (standard deviation: 8.5'-6) for the others†.

The experimental viewpoint adopted in regard to the definition of relevant documents was thus, in summary, that a review paper provides, through the citations made by it, a set of documents that reference a particular information need, i.e. are in a sense coherent or 'relevant'. (For the author of a review paper, or another person, to claim that relevant items were deliberately excluded from it, would be to imply the existence of a new criterion of relevance.) This is not to say that alternative operationally-defined definitions of relevance cannot be stated, indeed we have already mentioned an alternative definition, just that the work described here was concerned to emphasise the importance of operational definition in this regard, and was

---

‡ The data base partitioning of interest, in all cases, was that of the part of MEDLARS cut off by the two dates mentioned above.

† The significance of the difference, using a t-test with population standard deviations not assumed equal, is 0.001 (one-tailed).

chosen to be constrained by the choice of operational definition that we have described. The sets of relevant documents examined are thus homogeneous or 'controlled' in regard to this variable.

4.1.3 Definition of the DOEs to be used

Four DOEs were chosen, denoted W1, W2, W3 and W4. These will be defined using the notation of Section 3 as expressed for weights applied to elementary logical conjuncts of the form:

$$t_a^{\delta_a} \wedge t_b^{\delta_b} \wedge t_c^{\delta_c} \wedge \ldots \wedge t_n^{\delta_n} \in L_Q$$

for a query in set form of size $n = \| Q \|$. The superscript δ_i has the value 1 when the corresponding term (t_i) is attached to the document being "weighted". For example, $\delta_b = 1$ if $t_b \in \{Q \cap T_d\}$, else $\delta_b = 0$. Our interest is in the ranking of these elementary conjuncts by the real values to which they are mapped by the weighting function. We will treat the subscripts, a,b,c,... as if integer-valued.

W1: Co-ordination level

$$\text{weight} = \sum_{i=1}^{n} \delta_i .$$

W2: Miller DOE

Define: ψ_{i1} = probability of the event $\{t_i \in d\}$ for relevant documents

ψ_{i2} = probability of the event $\{t_i \in d\}$ for non-relevant documents

(ψ_{i2} is approximately equal to the probability of the event $\{t_i \in d\}$ for all documents, i.e. to the "specificity" of t_i.)

Then:

$$\text{weight} = \sum_{i=1}^{n} \delta_i \log(\psi_{i1} / \psi_{i2}) .$$

W3: Likelihood ratio DOE

Define $\chi_j(\delta_a, \delta_b, \ldots, \delta_n)$ $(j=1,2)$ as the probability of the event $t_a^{\delta_1} \wedge t_b^{\delta_2} \wedge \ldots \wedge t_n^{\delta_n}$ =TRUE for relevant documents $(j=1)$, and non-relevant documents $(j=2)$. Then for $\chi_1 \neq 0 \neq \chi_2$:

$$\text{weight} = \chi_1/(\chi_1 + \chi_2), \quad \chi_1 \neq 0 \neq \chi_2.$$

For elementary conjuncts that are CAI (i.e. $\chi_1 = 0 = \chi_2$) we define:

$$\text{weight} = 0, \quad \chi_1 = 0 = \chi_2.$$

Lastly, for elementary conjuncts to which no relevant documents are posted but to which some non-relevant documents are posted we define:

$$\text{weight} = -a, \quad a > 0, \quad \chi_1 = 0, \quad \chi_2 > 0.$$

This convention determines the elementary conjuncts attracting only non-relevant documents are given a common weight less than that for any conjunct attracting one or more relevant documents. Arbitrariness is necessary here even though we are in effect fixing the value of J in so doing.

W4: Modified Robertson/Sparck Jones DOE

Define ψ_{i1} and ψ_{i2} as for W2. Then define:

$$w_i = \log [\psi_{i1}(1-\psi_{i2})/\psi_{i2}(1-\psi_{i1})].$$

Then:

$$\text{weight} = \sum_{\delta_i=1} w_i + \sum_{\delta_i=0} -w_i.$$

The existence of the second summation constitutes a modification to Robertson and Sparck Jones's formula. It serves to lower the document weight according to both the

number of query terms not present in the document and the 'influence' of such terms as reflected in their w_i values.

It is emphasised that of the above DOEs only W1, W2 and W4 are applicable in practice, since W3 is a function of data that is not accessible except in laboratory-like environments (i.e. when the relevant set is known.) W3 will of course yield the best possible Recall-Precision graph for the particular query (in set form) chosen. The aim in using analytical expressions such as the others, in information retrieval, is of course to anticipate the ranking of elementary conjuncts that is effected by W3. Unfortunately the optimality of W3 may be obscured when particular measures of retrieval effectiveness are chosen, by virtue of the rather arbitrary weights assigned to documents when the likelihood ratio is indeterminate. In comparing W1, W2 and W4 we note that W2 and W4 are different in character from W1 in that each uses information on the probabilities of assignment of query terms to relevant documents, and W1 does not. This may be seen either as a 'heuristic feature' of them (in that an earlier retrieval attempt may have given partial information on these probabilities), or as entailing in practice the enquirer 'estimating' such probabilities intuitively. One usefulness in the experiment to be described lies in its establishing levels of plausibility as to the superiority or otherwise of such weighting functions. (If the user, even with perfect foreknowledge of these probabilities, could not retrieve documents more effectively using such functions than by using, say, W1 then there would appear to be little merit in seeking to use them in operational systems: the additional user-effort and host-system software required to implement them would be pointless.)

4.1.4 Definitions of the query in set form

Two forms of query in set form were defined for each set of relevant documents. Each query was chosen to be of exactly five terms, since it had been found from experience that this figure represented a reasonable compromise between (1) a query containing so many terms that the majority of x_1 and x_2 values would be zero (i.e. where the refinement to the retrieval process offered by increasing the number of elementary conjuncts that could be ranked had negligible effect), and (2) a query containing so few terms that the elementary logical conjuncts generated by it could not effectively 'separate' or distinguish the two types of document. (In common language, these extremes could be said to be equivalent to semantic distortion of the information need through over-refined or insufficient description, respectively.) It was initially considered that using five terms might provide a little too refined an approach (since 32 elementary conjuncts are then defined and ranked), and that a query of size four (defining only 16 elementary logical conjuncts) would be adequate. However, in that each query of size K can yield K less-specific sub-queries of size K-1 (through Boolean disjunctions of each of the terms, one at a time), it was concluded that the more taxing problem of manipulating data from a query of size five should be solved.

Two queries in set form were in fact chosen for each set of relevant documents, which we label as QFORM1 and QFORM2. The algorithms adopted to define them, for each instance of partitioning of the data base, were as follows:

QFORM1: 1. Identify all terms in the set of relevant documents that occur with a frequency in excess of

a given threshold value. (The latter was set to 1 for all sets of relevant documents except the larger ones where it was set to 2.)

2. Evaluate ψ_{il}/ψ_{is} for each such term, where ψ_{is} is the probability of occurrence of the term in a large random sample of the data base. (The sample was in fact the MEDLINE file in the final run of the experiment, although in earlier trials the local MEDUSA file was used - approximately 1/10 the size and a less random sample in that documents in certain languages were excluded from it.)

3. After ranking terms by the ratio mentioned, i.e. by their likelihood ratio, select the top-ranking five terms.

QFORM2: 1. As for QFORM1, with a threshold value of 2.

2. Cluster all such terms according to the sets of documents to which each term is posted, using single-linkage (Euclidean distance) nearest-neighbour clustering.

3. Choose the five deepest-clustered terms. If a choice of terms is so implied, for some or all terms, choose those terms having the highest value of ψ_{il}/ψ_{is}.

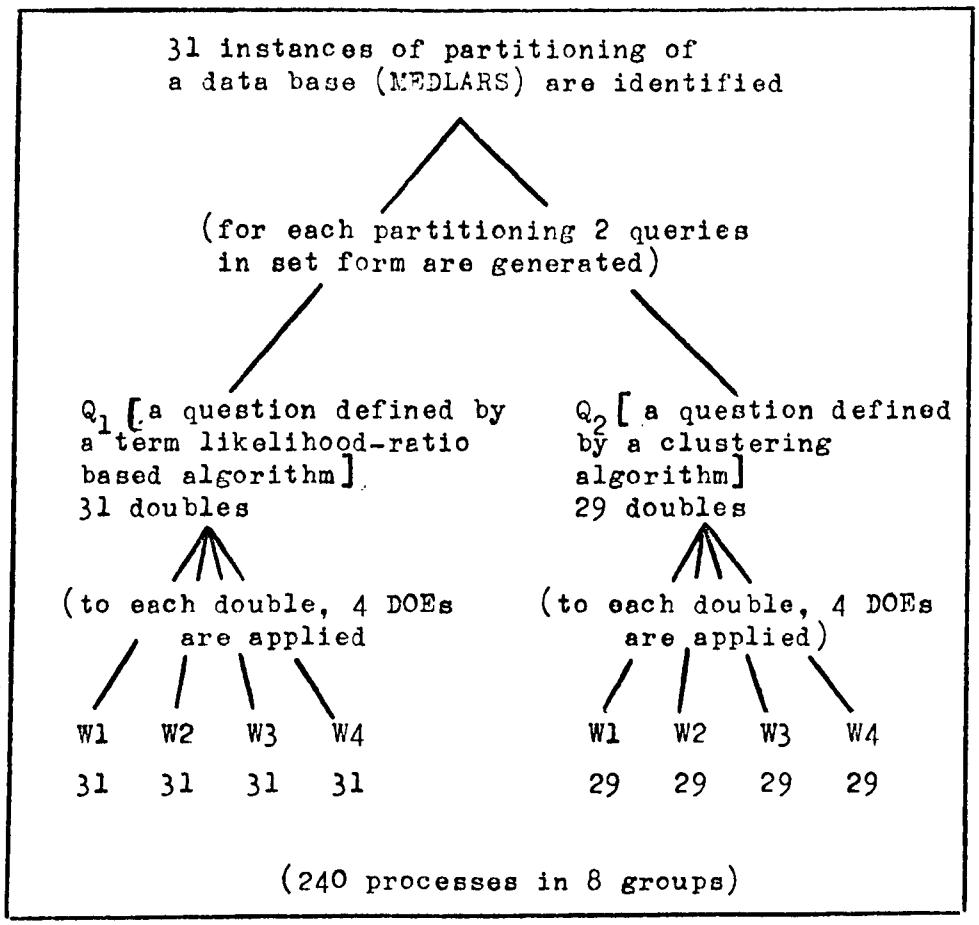
It was considered, though without rigorous proof, that dependence between the terms assigned to relevant documents was reflected in forming queries of type QFORM2, while totally disregarded in the procedure for forming QFORM1 queries. The method leading to QFORM2

queries does not involve comparison between depth of clustering in the set of relevant documents and depth of clustering in the set of non-relevant documents, so that intuitively it does not have the 'discriminating' character of the method for QFORM1 queries. Nevertheless QFORM2 queries provided plausible, alternative queries with which the QFORM1 queries could be compared.

4.1.5 Definition of the retrieval processes examined

The pair of observed distributions $f_i(j)$, i=1,2, which constitute a retrieval process, are determined by the triple: <partitioned data base, query in set form, DOE>. The variation we have so far described accordingly allowed 31 (partitionings) x 2 (query types) x 4 (DOEs) = 248 retrieval processes to be defined in principle. The actual figure for the number of processes examined was in fact less than this, at 240, since in the case of two of the partitionings accurate data on MEDLINE frequencies could not be obtained for all of the elementary conjuncts, in the case of QFORM2 queries. Figure 4.1.5-1 shows, schematically, the generation of the retrieval processes from the initial partitionings.

The method used to generate the 240 information retrieval processes examined was as follows. (Terminology: a 'double' denotes a combination of a partitioning of a data base and a question.)



Each process so identified was analysed to identify a wide range of (1) descriptive statistics, and (2) statistics comparing one of two modelling distributions (for that process) with the observed distributions. The variation of these statistics within each of the 8 groups was taken to define (i.e. generate) hypotheses in the Swetsian formalism. The statistics also enabled the four DOEs used to be ordered (for each question type) by the values of one of three effectiveness criterion variables; and enabled the question-generation algorithms to be ordered for each DOE.

Fig. 4.1.5-1 Scheme adopted for generating 'retrieval processes' within the experiment.

4.1.6 Definition of the modelling distributions

The plan of the experiment included an attempt to model, for each retrieval process, the two observed distributions of documents over the rank values of the events e_i identified in Section 3.3.2.3, the ranking being achieved by means of one or other of the DOEs described in Section 4.1.3. The distributions were chosen to be defined according to descriptive constraint '1' of Section 3.3.2.1; that is, rank values were attached to CAI events and the event $z=Z(\emptyset)$ was included in the distribution. Modelling functions for the distributions were accordingly also so constrained. Obviously there is a wide variety of analytical forms from which modelling functions could be selected, e.g. binomial, Poisson, uniform-discrete, hypergeometric, etc, and in view of the lack of background theory here it was decided to compare the observed distributions with the two simplest analytical forms: the binomial and the uniform-discrete. (The Poisson distribution was rejected, in favour of the binomial, on the ground that J is finite and often fairly small: e.g. of value n+1 for the DOE we have labelled W1.) An important qualification needs to be added however, in the case of the distributions for non-relevant documents. This is that the analytical form used in such cases was matched only against the tail of the distribution - defined by the distribution over all events except the first (lowest-ranking) event. It specifically excluded the very large 'spike' of probability that invariably attaches to the elementary conjunct:

$$t_a^o \wedge t_b^o \wedge t_c^o \wedge t_d^o \wedge t_e^o$$

for non-relevant documents. In general (i.e. for most queries as SFQ) no such spike exists for the distribution of relevant

documents.

A set of elementary conjuncts having been ranked by means of a DOE, there are $J \leq 2^{\|Q\|}$ rank values upon which the distributions $f_i(j)$ are defined. The modelling distributions were chosen to be:

For the sets of relevant documents:

(1) The binomial distribution b(J-1,h),
where h=(mean observed rank value)/(J-1).

(2) The uniform-discrete distribution u({1,2,3,...,J}), the probability at each rank value being 1/J.

For the sets of non-relevant documents:

(1) A 'spike + binomial' distribution:
for r=1, probability=(no. of non-relevant documents mapped to r=1)/(no. of non-relevant documents),
for 1< r< J probability is distributed as b(J-2,h) (1-P ({1})), where h=(mean observed value of (r-1)/(J-2)).

(2) A 'spike + uniform-discrete' distribution:
for r=1, probability=(no. of non-relevant documents mapped to r=1)/(no. of non-relevant documents),
for 1< r< J, probability is distributed as u({2,3,4,...,J}), where individual probability values equal ((no. of non-relevant documents)-(no. of non-relevant documents mapped to r=1))/(J-1).

The above modelling distributions were used in matching pairs. That is, either the pair: binomial and 'spike + binomial'; or the pair: uniform-discrete and 'spike + uniform-discrete', were used to model each retrieval process. The value of the spike of probability was taken to be exactly that of the observed proba-

bility for r=1, so that a degree of freedom in the model is thereby lost. As emphasised earlier in this thesis, the probabilities attached to the event $t_a^o \wedge t_b^o \wedge t_c^o \wedge \ldots$ have not been treated previously in the literature.

4.1.7 Measurement of the observed probabilities of individual elementary conjuncts

The two observed probabilities attaching to each of the 32 elementary conjuncts defined by each query type for each partitioning of the data base, were obtained as follows. In the case of the set of relevant documents the probabilities were found by direct measurement of the document frequencies for each elementary conjunct. In the case of the set of non-relevant documents they were found by estimation using the MEDLINE file. (In earlier investigations the Newcastle MEDUSA file had been used for the purpose but this proved less satisfactory than MEDLINE in that (1) it was a much smaller file - a critical factor given that many of the frequencies of elementary conjuncts for non-relevant documents were zero, or close to zero, for an efficient query of size five; and (2) the frequencies concerned were almost always given explicitly by the host operating system of MEDLINE whereas with MEDUSA the operating system gave only a standardised verbal estimate of the frequency (e.g. "Expected return - small") based on an assumed independence in the assignments of the terms.) The experiment estimated the frequencies of non-relevant documents, for each elementary conjunct, as the product of the frequency pertaining to the MEDLINE file as a whole with the ratio: (size of MEDLARS data base within time scope of review paper)/(size of MEDLINE file at date of searching). This estimated frequency was rounded to the nearest integer for each elementary conjunct, and the total number of non-relevant documents was, in subsequent analyses, assumed to be the sum of these rounded frequencies. The data obtained thus contained the implicit assumptions that

(1) the MEDLINE file represented a random sample (of size approximately 500,000 items at the time of searching) of that section of the data base from which it was assumed the relevant documents were to be retrieved, and (2) the elementary conjunct frequencies for all documents provided a satisfactory estimate of the frequencies for non-relevant documents.

4.2 Data acquisition and data flow

Machine-readable copy of records representing relevant documents were obtained from MEDLARS files by (1) inputting suitable queries to a "STORESEARCH" search-description, (2) having the results of the search routed to a magnetic tape instead of the usual line-printer*, (3) transferring the magnetic tape contents to a local disc file, and (4) editing the latter file so as to exclude superfluous records.

The sets of records of relevant documents, now represented in machine-readable form, were processed by a sequence of programs in order to obtain the information from which the query form QFORM2 could be generated for that set, and to obtain some of the information - the raw frequencies of assignment of terms - prior to determining query QFORM1 for that set. In the latter case, inputting the more frequently assigned terms to the MEDLINE file gave the additional frequency information needed to identify the QFORM1 query. The frequencies of assignment of the 32 elementary conjuncts for each query type were also obtained (for the set of relevant document only) from tables generated by these programs. For non-relevant documents, the equivalent frequencies were obtained by searching the MEDLINE file (a second time, in the case of QFORM1 queries), and normalising the frequencies using the method described in section 4.1.2.

The array of 32 pairs of frequencies, for each combination of partitioned data base and query, was input to a further program which defined 4 retrieval processes for each array. These corresponded to various rankings of the 32 rows of the array according to the working of the DOEs W1, W2, W3 and W4. The program

---

* The co-operation of the National Library of Medicine staff in allowing this non-routine step is gratefully acknowledged.

merged events where a common weight was entailed. It then derived, for each retrieval process so defined, the data described in Section 4.3. The latter was output both to paper using a line-printer and in summary form to a disc file which was later to serve as the file of input data of a standard statistical analysis program: SPSS in fact.

All programs (except SPSS, of course) were written by the author, c.3000 statements (ALGOLW) being involved represented in eight discrete programs. Considerably more flexibility was written into the programs than proved to be necessary, however.

4.3 Data obtained on each retrieval process

The 240 retrieval processes defined in the experiment are characterised by the following data. As can be seen, data describing both the 2x240 observed distributions, and the (2x240)x2 modelling attempts was obtained.

(1) Descriptive statistics on each process.

(a) The numbers of relevant and non-relevant documents, and the Generality value.

(b) The mean, variance, skewness and kurtosis of the distribution for the set of relevant documents and of that for the set of non-relevant documents.

(c) The mean, variance, skewness and kurtosis of tail of the distribution ($1 < r \leq J$) for the set of non-relevant documents only and with events relabelled by $r'=r-1$ (so that the range was $r'=1,2,\ldots,J-1$).

(d) Swets's $\underline{E}$ value, and Brookes's $\underline{S}$ value, for the two distributions $f_i(j)$ taken jointly.

(e) The expected value and variance of the Euclidean distance between the origin and the points of the R-P graph.

(f) The expected value and variance of the Marczewski-Steinhaus distance.

(It is emphasised that the statistics were based on the rank values, not weight values, of events.)

(2) Statistics describing the adequacy of the modelling distributions

(a) The Kolmogorov-Smirnov statistic for the set of

relevant documents (measuring the maximum numerical difference between the CDFs of the modelling distribution and the observed distribution), and for the set of non-relevant documents.

(b) The mean, variance and range of the Euclidean distance between comparable (R,P) co-ordinates.

(c) The mean, variance and range of the difference in the distance between comparable pairs of the sets: set of retrieved documents and set of relevant documents, as measured by the Marczewski-Steinhaus metric.

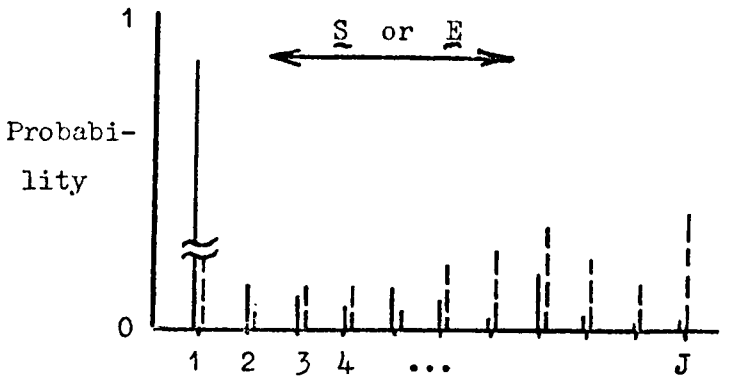
The Chi-square statistic was rejected as a measure of modelling accuracy owing to many of the expected cell frequencies being less than the minimum acceptable value of 5. Some 25,000 statistical values resulted from the above scheme of description, for the 240 retrieval processes examined. This illustrates how a relatively small number of partitionings (31) can yield, with relatively minor variation in the experiment and a modest number of properties of interest, an almost intractable number of data values.

Figures 4.3-1 and 4.3-2 illustrate the less-conventional statistics defined in the experiment.

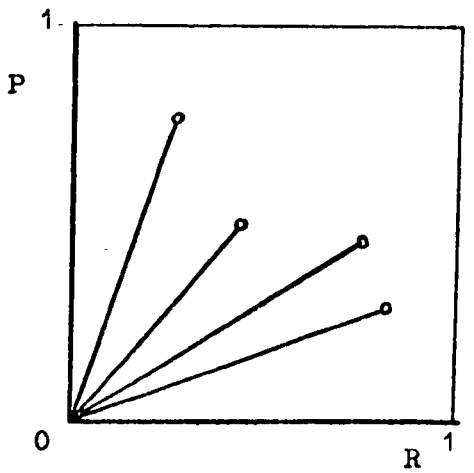
It is apparent that the data involved in such an experiment exists in various forms at different stages of the experiment. First, data appears as machine-readable copy of data-base records, in this case MEDLARS records. Secondly it appears as information resulting from analyses of sets of same, e.g. information on clustering within the set of terms attached to relevant documents. Thirdly data appears as ordered pairs of frequencies in correspondence with an array of elementary logical conjuncts, although at this

stage the latter are not themselves ordered (see Figure 4.3-3). Fourthly, data appears as ordered pairs of frequencies attached to composite logical expressions of the type we labelled e_i, i.e. attached to rank values of document weights to which the elementary logical conjuncts are mapped (see Figure 4.3-4). Fifthly, we have data as descriptive information on each retrieval process so identified. (It is at this stage that data on the success or otherwise of modelling attempts is generated.) Sixthly, data on the variation of the latter descriptive statistics is defined by examination of more than one retrieval process.

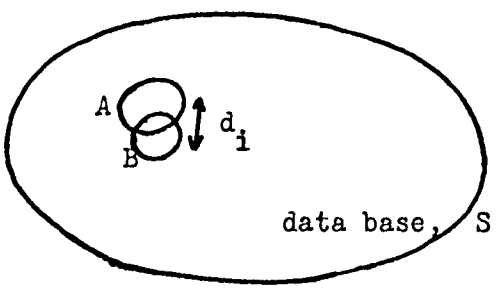
The data generated by the experiment undertaken by the author is summarised in Appendix C.



$\underset{\sim}{S}$ and $\underset{\sim}{E}$ measure the separation of the distributions f_1 (denoted ---) and f_2 (denoted ——), for a given retrieval process

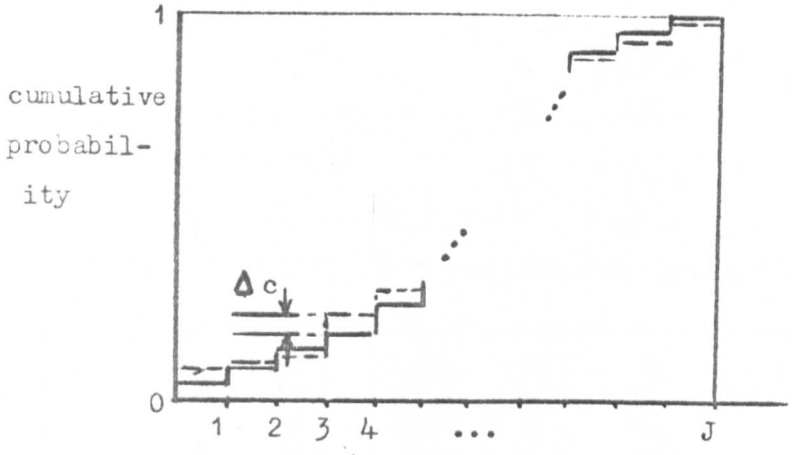


mean value of the Euclidean distances from the origin to the Recall-Precision graph, for a given process

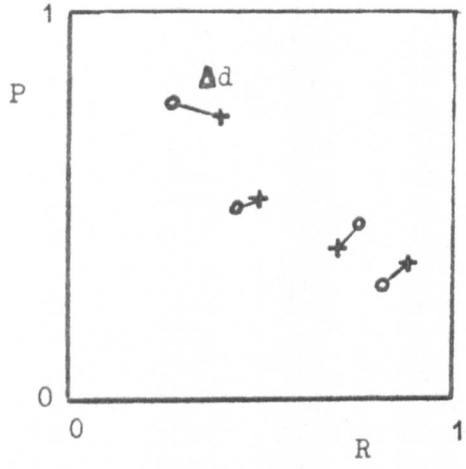


mean value of a measure of separation (d_i) between the set of relevant documents, A set of retrieved documents, B, for a given process

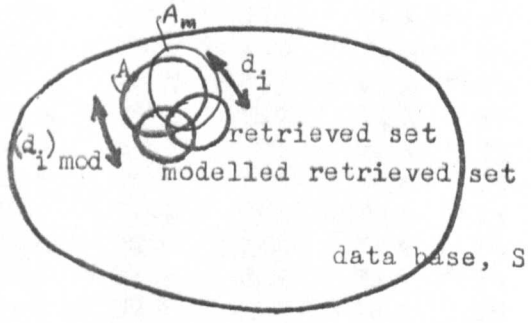
Figure 4.3-1 The less-conventional descriptive statistics evaluated in the experiment for each retrieval process, in diagrammatic form.



Kolmogorov-Smirnov statistic: maximum (unsigned) value of Δc, the separation of the CDF (observed process) and CDF (modelled process), for a given process



mean value of Euclidean distance Δd, measuring separation of R-P graph (observed) and R-P graph (modelled), for a given process



mean value of measure of separation $|d_i-(d_i)_{mod}|$ for a given process

Figure 4.3-2 The less-conventional statistics evaluated in the experiment for each retrieval process, serving to compare observed distributions $f_i(j)$ with modelling distributions $m_i(j)$, i=1,2.

"SCH1A2023"

t_1 = "NITROGENASE" .22 2.7'-4

t_2 = "CATALYSIS" .33 9.2'-4

t_3 = "AZOTOBACTER/EN" .28 7.0'-5

t_4 = "MOLYBDENUM" .28 3.4'-4

t_5 = "BINDING SITES" .22 8.4'-3

In the following, for clarity, we write T1 for t_1^1 and ¬T1 for t_1^0, etc.

T1 ∧ T2 ∧ T3 ∧ T4 ∧ T5	0	0
T1 ∧ T2 ∧ T3 ∧ T4 ∧ ¬T5	0	0
T1 ∧ T2 ∧ T3 ∧ ¬T4 ∧ T5	0	0
T1 ∧ T2 ∧ T3 ∧ ¬T4 ∧ ¬T5	0	0
T1 ∧ T2 ∧ ¬T3 ∧ T4 ∧ T5	0	0
T1 ∧ T2 ∧ ¬T3 ∧ T4 ∧ ¬T5	2	0
T1 ∧ T2 ∧ ¬T3 ∧ ¬T4 ∧ T5	0	0
T1 ∧ T2 ∧ ¬T3 ∧ ¬T4 ∧ ¬T5	2	18
T1 ∧ ¬T2 ∧ T3 ∧ T4 ∧ T5	0	5
T1 ∧ ¬T2 ∧ T3 ∧ T4 ∧ ¬T5	0	27
T1 ∧ ¬T2 ∧ T3 ∧ ¬T4 ∧ T5	0	0
T1 ∧ ¬T2 ∧ T3 ∧ ¬T4 ∧ ¬T5	0	45
T1 ∧ ¬T2 ∧ ¬T3 ∧ T4 ∧ T5	0	13
T1 ∧ ¬T2 ∧ ¬T3 ∧ T4 ∧ ¬T5	0	58
T1 ∧ ¬T2 ∧ ¬T3 ∧ ¬T4 ∧ T5	0	13
T1 ∧ ¬T2 ∧ ¬T3 ∧ ¬T4 ∧ ¬T5	0	443
¬T1 ∧ T2 ∧ T3 ∧ T4 ∧ T5	0	0
¬T1 ∧ T2 ∧ T3 ∧ T4 ∧ ¬T5	0	0
¬T1 ∧ T2 ∧ T3 ∧ ¬T4 ∧ T5	0	0
¬T1 ∧ T2 ∧ T3 ∧ ¬T4 ∧ ¬T5	1	0
¬T1 ∧ T2 ∧ ¬T3 ∧ T4 ∧ T5	0	0
¬T1 ∧ T2 ∧ ¬T3 ∧ T4 ∧ ¬T5	1	9
¬T1 ∧ T2 ∧ ¬T3 ∧ ¬T4 ∧ T5	0	303
¬T1 ∧ ¬T2 ∧ ¬T3 ∧ ¬T4 ∧ ¬T5	0	1788
¬T1 ∧ ¬T2 ∧ T3 ∧ T4 ∧ T5	0	0
¬T1 ∧ ¬T2 ∧ T3 ∧ T4 ∧ ¬T5	1	0
¬T1 ∧ ¬T2 ∧ T3 ∧ ¬T4 ∧ T5	1	5
¬T1 ∧ ¬T2 ∧ T3 ∧ ¬T4 ∧ ¬T5	2	81
¬T1 ∧ ¬T2 ∧ ¬T3 ∧ T4 ∧ T5	0	40
¬T1 ∧ ¬T2 ∧ ¬T3 ∧ T4 ∧ ¬T5	1	626
¬T1 ∧ ¬T2 ∧ ¬T3 ∧ ¬T4 ∧ T5	3	18967
¬T1 ∧ ¬T2 ∧ ¬T3 ∧ ¬T4 ∧ ¬T5	4	2268338

Figure 4.3-3 Typical data at the third stage of the experiment. The header label "SCH1A2023" identifies both the set of relevant documents and the query type. t_1, t_2, t_3, t_4 and t_5 denote the five terms comprising the query, with the specificities of the terms in both the sets of relevant & of non-relevant documents also noted. The two columns of frequencies record the frequencies of assignment of each of the 32 elementary conjuncts in the set of relevant documents and the set of non-relevant documents, respectively. Note the relatively high proportion of the elementary conjuncts that have been assigned zero probability values, for both the set of relevant documents and the set of non-relevant documents, for this query of fairly modest size. Note also that the order of the elementary conjuncts is arbitrary at this stage. A DOE has yet to be applied to generate the distributions $f_i(j)$, i=1,?, j=1,2,...,$J< 2^5$.

rank value (j)	document weight value, z	frequency for set of relevant documents	frequency for set of non-relevant documents
1	0	4	2.26'6
2	1	6	21905
3	2	6	491
4	3	2	40
5	4	0	5
6	5	0	0

(J=6)

Fig. 4.3-4 An example of a retrieval process, defined by a combination of the data given in the previous figure and an expression for mapping the elementary conjuncts to real values, i.e. a DOE. Here the particular DOE used is W1. The mean, variance etc of the two distributions (defined over the rank values of the weights) are readily found, as are the Recall-Precision graphs etc. It is these distributions (reduced to the probability distributions $f_i(j)$) that were modelled by means of the analytical forms described in Section 4.1.5.

4.4 Analysis of the results of the experiment

In this section we selectively comment upon and analyse the results given in Appendix C, in accordance with the objectives of the experiment given at the beginning of Section 4. As mentioned there, the main aims were to generate hypotheses through obtaining the data as have been given, and to establish a methodology for experiments consistent with the extended formalism. The main hypotheses of interest, in the author's view, are those describing orderings of the four DOEs used, for each of the two methods of query generation employed. The DOEs determine logical search expressions of immediate practical application, and hence choice of DOE is of value in determining the optimum logical search expressions for a given query in set form. With this purpose in mind, the three definitions of the effectiveness of a retrieval process, serving to order W1, W2 and W4 (the three operational DOEs), were chosen to be: (1) Brookes's S measure redefined over rank values, (2) expected value of Euclidean distance from origin to $\{(R,P)\}$, and (3) expected value of Marczewski-Steinhaus metric. These correspond, respectively, to the variables named in Appendix C as DS1OA, DS1OB and DS1OC. For each of the three DOEs W1, W2 and W4, a hypothesis in the form of an ordering of the two methods of query generation was also generated from the experimental data. The following notes first offer general comments on the experimental data, then identify the hypotheses we have just described, and lastly comment on more incidental findings: the adequacy of the two modelling functions used, the existence of correlations between selected pairs of descriptive statistics, and the estimated maximum

Recall value for a query of size n=5.

The first, very simple but basic observation is that variation in the statistics used to describe the retrieval processes is observed. This point may seem an obvious one but it was obscured in Swets's work where such variation was rendered 'invisible' by his working with averaged data, i.e. with a fictitious composite retrieval process for each data base. Hypotheses describing the variation in each of the descriptive statistics DS01 to DS10 can be written 'post hoc' in the sense that we can hypothesise that if all partitionings (of the type examined) of the data base were considered, then for a given DOE and query type the statistic will have a sampling distribution with population mean equal to the sample mean, and population variance equal to the (unbiased)sample variance. For example, referring to Table C-1, we see that in the sample of 31 retrieval processes considered, the statistic DS10A, i.e. Brookes's measure of effectiveness $\underset{\sim}{S}$, is so distributed that we may infer the population mean and population variance to be 1.391 and 0.238, respectively. (This is for queries of type QFORM1 and the DOE W1.) These are point estimates with estimated standard errors of $\sqrt{0.238/31}$ and $0.238\sqrt{2/31}$ respectively, if it is assumed that in the population of processes $\underset{\sim}{S}$ is distributed Normally and the sampling distributions for these two statistics (mean and variance of $\underset{\sim}{S}$) for this size of sample are also Normal. (These values for the estimated standard errors also assume that the sample was obtained with replacement of processes allowed.) The assumption is also made that the sample concerned is a random one which, as described in Section 4.1.2, is not sound: the sample was partly 'quota' in character, and partly designed to identify

'extreme' processes through a large variation in Generality. The information given on the skewness and kurtosis of this and other statistics was also obtained but is not reproduced since the data is probably not of much value for inference purposes for samples as small as these. The information given in Tables C-1 to C-8 in effect generates hypotheses for 20 statistics of interest, which can be contested by later experiments. They do not refute hypotheses suggested by Swets since none were made by him consistent with the formalism we have introduced. However, the implication behind Swets's work, that the features of individual retrieval processes remain constant (from process to process) is expressed directly in the data given as the magnitude of the variance of particular statistics. Swets's assertion that the CDFs of the distributions for an 'averaged process' can be approximated by the CDFs of Normal density functions is also carried over into a description here of the adequacy of the discrete, binomial model as a representation of the discrete distributions of each process. But the evaluation of the binomial model here takes into account the probabilities attaching to the all-negated elementary logical conjunct, which were totally ignored by Swets. Whereas the influence of G on the statistics was not treated by Swets (and could not be, since only averaged data were considered), the relationships between the various statistics and G were rendered observable in the present approach. Tables C-11 to C-14 record correlations of this type, which again represent a construction of hypotheses which can be contested in later work.

Inferences based on the data given here for some statistics, e.g. DS09 to DS18, depend on a knowledge of the sampling dis-

tributions of the statistic concerned. To the author's knowledge none of the sampling distributions of interest is known, nor does the Central Limit Theorem offer proof that the distributions will be approximately Normal in the case of DS13 and DS16 (where the statistic is a sum) since it is not clear that the values being summed are drawn from the same population distribution. Accordingly, for all hypotheses inferred, and for hypotheses inferred from data for statistics DS09 to DS18 in particular, a source of uncertainty of unknown extent is involved. In view of this it was considered that this data should be seen primarily from a scientific standpoint rather than from a rigorous statistical one, notwithstanding the application of some statistical procedures to the data.

For the DOEs W1, W2 and W4, the sample variance of each of the statistics DS01 to DS10 is less in the case of the set of six retrieval processes which <u>appeared</u> to be homogeneous (by virtue of the relevance judges having a common subject area of interest) than in the case of <u>all</u> the processes. The one exception was for statistic DS08, DOE W1, and queries of type QFORM2. The homogeneity postulated was thus largely borne out.

The effectiveness of the two queries used with each DOE can be compared using the three criterion variables: DS10A, DS10B and DS10C, with the following result. In the case of each DOE and each criterion variable the queries of type QFORM1 performed <u>better</u> than the queries of type QFORM2. The improvement in performance was however not always statistically significant as measured by the t-test (independent samples, unequal population variances). The detailed results and estimated significance values are as given in Table 4.4-1. Clearly queries formed from terms identified on the

basis of the simple ratio: (frequency of term in set of relevant documents) / (frequency of term in whole data base), are in general superior to those formed from terms identified from clustering terms attached to relevant documents, when the clustering is of the prescribed type, for the data base concerned and the type of information need concerned.

On the assumption that queries of type QFORM1 are indeed (sub)-optimal for the two types of query examined, we can seek to identify which of the DOE employed is (sub)optimal for such queries. (The two sub-optima are components of a jointly-defined optimum process, as previously described.) We do this by the following steps =

(1) estimating significance values for the inequalities:

$$\begin{pmatrix}\text{mean value of DS1OA}\\ \text{determined by WI}\end{pmatrix} > \begin{pmatrix}\text{mean value of DS1OA}\\ \text{determined by WJ}\end{pmatrix} \quad ;$$

$I, J = 1,2,3,4;\ I \neq J$

- and similarly for DS1OB, and -DS1OC (choosing the minus sign to reflect better effectiveness for <u>smaller</u> values of DS1OC). The t-test for independent samples and unequal population variances was used in arriving at these significance values (Snedecor and Cochran, 1967:114).

Criterion: mean value of DS10A determined by the DOE specified is higher for queries of type QFORM1 than for queries of type QFORM2:

W1 :	Significance	=	0.025
W2 :	"	=	0.05
W3 :	"	=	0.10
W4 :	"	=	0.05

Criterion: mean value of DS10B determined by the DOE specified is higher for queries of type QFORM1 than for queries of type QFORM2:

W1 :	Significance	<	0.01
W2 :	"	=	0.05
W3 :	"	=	0.25
W4 :	"	=	0.05

Criterion: mean value of DS10C determined by the DOE specified is less for queries of type QFORM1 than for queries of type QFORM2:

W1 :	Significance	>	0.40
W2 :	"	=	0.40
W3 :	"	=	0.25
W4 :	"	>	0.40

Table 4.4-1 Estimated one-tailed significance values for the superiority of QFORM1 queries over QFORM2 queries, for mean retrieval effectiveness values given by the criterion variables DS10A, DS10B and DS10C. Values have been rounded to conventional, tabulated significance values.

(2) Ordering the three DOEs W1, W2 and W3 by applying a threshold value of 0.05 to the estimated significance values so obtained.

The results of step 1 and step 2 are summarised in Tables 4.4-2 and 4.4-3. The orderings given in Table 4.4-3 reflect the efficiencies of the three analytical DOEs in determining logical search expressions appropriate to the data base partitionings we have defined*. We conclude that it appears that the optimum generators of logical search expressions, under the conditions of the experiment, are:

W1 for effectiveness as measured by Brookes's measure (redefined);

W1 for effectiveness as measured by mean Euclidean distance from the origin to $\{(R,P)\}$ (but not significantly superior to W2 or W4);

W4 for effectiveness as measured by the mean value of the Marczewski-Steinhaus metric (but not significantly superior to W2).

---

* The reason that likelihood ratio weighting did not always throw W3 to the first position is presumably that arbitrary weights have to be assigned to elementary conjuncts which are almost-impossible or CAI.

Criterion variable: DS10A

W1	>	W2	=	Significance	< 0.005
W1	>	W3	=	"	= 0.25
W1	>	W4	=	"	= 0.005
W3	>	W2	:	"	= 0.025
W3	>	W4	:	"	< 0.05
W4	>	W2	:	"	> 0.40

Criterion variable: DS10B

W1	>	W2	:	Significance	= 0.10
W3	>	W1	:	"	< 0.001
W1	>	W4	:	"	= 0.10
W3	>	W2	:	"	< 0.001
W2	>	W4	:	"	> 0.40
W3	>	W4	:	"	< 0.001

Criterion variable: DS10C

W2	>	W1	:	Significance	= 0.010
W3	>	W1	:	"	< 0.001
W4	>	W1	:	"	= 0.010
W3	>	W2	:	"	= 0.025
W4	>	W2	:	"	> 0.40
W3	>	W4	:	"	= 0.025

Table 4.4-2 Estimated one-tailed significance values (rounded to conventional values) for the hypotheses shown, obtained using the t-test (independent samples, unequal variances). In all cases the queries were those of type QFORM1. The notation WI > WJ here denotes 'mean value of criterion variable is higher for WI than for WJ' for the variable-DS10C.

Orderings including W3:

Criterion variable DS10A	: $W1 \gtrsim W3 \succ W4 \gtrsim W2$	
Criterion variable DS10B	: $W3 \succ W1 \gtrsim W2 \gtrsim W4$	$(W1 \gtrsim W4)$
Criterion variable DS10C	: $W3 \succ W4 \gtrsim W2 \succ 1$	

Orderings excluding W3:

Criterion variable DS10A	: $W1 \succ W4 \gtrsim W2$	
Criterion variable DS10B	: $W1 \gtrsim W2 \gtrsim W4$	$(W1 \gtrsim W4)$
Criterion variable DS10C	: $W4 \gtrsim W2 \succ W1$	

Table 4.4-3. Orderings of the DOEs determined by the application of a threshold value of 0.05 to the significance values given in Table 4.4-2, for QFORM1 queries.

Tables C-1 to C-8 also contain information on the accuracy of the functions used to model the observed distributions. The two basic statistics used to measure modelling effectiveness were, as previously mentioned, mean Euclidean distance between comparable (R,P) co-ordinates, and mean value of the difference between the distance separating the set of relevant documents and the set of retrieved documents. In both these cases perfect modelling would be associated with values of zero for each type of variable (for each threshold value), and hence zero also for the mean values of these variables. Our interest is thus in whether the observed mean values for the samples considered are such that the samples could have been drawn from a population with zero mean. The estimated significance values for this hypothesis, for each of these variables and for the various groupings of retrieval process, are as in Table 4.4-4. Data for W3 have been excluded, and the significance values are two-tailed. It has been assumed that the population standard deviations equal the unbiased sample standard deviation. The better the model is at representing the observed data, the larger the significance value will be. Thus, for example, adopting the 0.05 level of significance as the criterion of acceptability, the binomial model is an unacceptable representation of individual retrieval processes defined by the QFORM1 queries and W1, when we take mean Euclidean distance between comparable (R,P) co-ordinates as measuring model effectiveness (since for variable DS13B, for processes so defined, $0.031 < 0.05$). But the binomial model is an acceptable model as measured by the difference between mean values of the Marczewski-Steinhaus metric, in this case (since $0.447 > 0.05$.) If we judge the accuracy of the modelling for all

classes of process by the DS16 statistic we see that the binomial and uniform-discrete models are both successful in all cases. But if the DS13 statistic were chosen, success is limited to only one class of process, that defined by QFORM1 and W1, and for only one of the models - the uniform-discrete model. Accordingly, accepting DS13 as the proper indication of model effectiveness leads to the conclusion that there is considerable room for testing other analytical functions as models, whereas accepting DS16 implies the opposite. DS13 would thus appear to be the more exacting test of model effectiveness and the author would suggest its use, rather than DS16, in future work.

The Kolmogorov-Smirnov statistic was also used to evaluate the modelling of observed data. With this test, it was assumed that the population distribution was in fact the modelled sample distribution, and the probability that a value for the statistic greater than or equal to that observed in a sample drawn from that population is estimated. The one-tailed significance values, grouped into intervals, are given in Tables C-9 and C-10 with the data for W3 included for completeness only. It can be seen that for the distributions pertaining to sets of relevant documents the adequacy of both analytical models is weak: most values of the statistic are too large for the hypothesis that the population could be so modelled to be sustained. For example, accepting 0.05 as the criterion significance value, 7 out of 31 of the distributions for relevant documents, for processes defined by QFORM1 and W1, cannot be modelled by a binomial model; 8 out of 31 of them cannot be modelled by the uniform-discrete model. This is in striking contrast to the success of the modelling for the

retrieval process group	statistic	estimated significance
QFORM1 queries and W1:	DS13B	0.031
	DS13U	0.082
	DS16B	0.447
	DS16U	0.373
W2:	DS13B	0.006
	DS13U	0.033
	DS16B	0.857
	DS16U	0.246
W4:	DS13B	0.006
	DS13U	0.034
	DS16B	0.857
	DS16U	0.234
QFORM2 queries and W1:	DS13B	0.003
	DS13U	0.017
	DS16B	0.280
	DS16U	0.267
W2:	DS13B	0.001
	DS13U	0.006
	DS16B	0.889
	DS16U	0.187
W4:	DS13B	0.001
	DS13U	0.006
	DS16B	0.889
	DS16U	0.180

Table 4.4-4 Estimated significance values for hypotheses that the measures of model accuracy shown come from a population with zero mean.

distributions for non-relevant documents (Table C-10) where either model will give a close correspondence between modelling distribution and observed distribution as measured by this statistic. There the almost total lack of variation in the distribution of the statistic over the significance value intervals suggests that the spike of probability used in both models, and inferred directly from each sample, is carrying most of the 'information' in the distribution. Concentrating the comparison on the modelling of the tails only of the distributions for non-relevant documents would presumably allow the worth of the two models to be contrasted more effectively. But in the author's view there is no point in attempting this in view of the influence of the magnitude of the spike of probability on the actual Fallout values; i.e. one would then be dealing only with 'conditional Fallout' values having no practical significance. The use of the Kolmogorov-Smirnov test in this context appears to be rather artificial as well (compared to the comparisons described earlier based on DS13 and DS16) in that a fictitious population, the subject of the modelling, is required to be invoked; when in fact our prime interest is specifically in the comparability of a sample model with the observed distribution in the sample.

The experimental data allows comparisons between the variation of certain random variables, in particular between G and the various statistics describing each process. Selected correlations have been given in Tables C-11 to C-14, for processes grouped as usual by DOE and query form. Data for W3 is not commented upon. It is apparent that in all classes of process except that defined by W1 and QFORM2, the correlation between the means and variances

of the distributions for the sets of relevant documents is significant, very strongly so in the case of W2 and W4. A very strong correlation also existed between the means and variances of the distributions for the sets of non-relevant documents (both whole distributions and tails) for W1, W2 and W4. In the case of all three DOEs both the mean weight and variance tended to decrease with increasing G, for relevant documents. This was also true of mean weight for non-relevant documents but not significantly so (at the 5% level.) The net effect of these changes with G is captured by the variation of each process's S value with G. In the case of all three DOEs, and both query types, S tended to decrease as G increased (true at the 5% level for an inferred population of processes.) Retrieving a smaller set of relevant documents would thus appear to be more effectively carried out in MEDLARS than retrieving a larger set, for such instances of information need and for such query types. (This conclusion is not intended to suggest that there is no lower limit below which the effect may not be observable.) Skewness and kurtosis both tended, in all cases, to increase significantly with G, with the exception of kurtosis for W1 where no significant result obtained.

Correlations such as the above were not anticipated in Swets's description of information retrieval. Their value, conceptually, is in (1) pointing to criteria by which any formalism more detailed than the one described in this thesis should be assessed, and (2) identifying concrete features of retrieval processes that should be incorporated into simulation models of retrieval processes. These comments also apply, of course, to the hypotheses implicit in the data given earlier.

In queries of size 5, and for the DOE W3, the probabilities attaching to the all-negated elementary conjunct were measured. These gave an average value of 0.252 [0.987] for r^{ooooo}_{abcdo} [s^{ooooo}_{abcde}]. Accordingly the maximum value of Recall [Fallout] for logical search expressions that are not disjoined to this elementary conjunct can be estimated to be 0.748 [0.013], for a query comprising this number of search terms. For high-Recall searching a higher value of n is accordingly necessary.

Lastly, by way of comment on the experiment as a whole, the author would suggest that the presence of algorithmically-defined queries (in set form) in the experiment, although introducing an essential control also introduces an uncertainty that needs to be remedied in future work. In practice the users of a data base will vary in the skill with which they formulate queries, and it is probably unlikely that they will use the same terms that make up (say) QFORM1 queries in the logical expressions that they input to the data base. What is clearly needed here (as well as improvement in all other remaining experimental variables) is a further experiment in which retrieval processes are defined both from user-specified queries and from algorithmically-derived queries, and a standard DOE is employed, allowing effective comparisons of query sources.

The experimental results given in this thesis relate to retrieval processes defined in particular ways. They relate (1) to instances of information need in medicine, (2) to the MEDLARS data base, and (3) to queries (defined as sets of terms) generated algorithmically in one of two ways. It is not claimed that the hypotheses generated by the experiment have a validity extending

beyond these constraints. The methodology used is however readily replicable, so that other hypotheses for other data bases can be easily found and compared against those reported here. The methodology used by the author had several known weaknesses which could be remedied in future work, however. These were as follows. First, the fact that the data base used was not entirely accessible on-line meant that the frequencies of assignment of non-relevant documents to elementary conjuncts could not be found exactly; instead a large sample of the data base (MEDLINE) had to be used to obtain estimated values. This provided a source of systematic error owing to both the known existence of obsolescence in indexing terminology, and variable rates of literature growth and decay within the data base. Secondly, the sample of sets of relevant documents used was (deliberately) not a random one: it was a quota sample designed to 'capture variability' to a maximum extent. A more statistically-oriented experimental design would perhaps involve a stratified sample of such sets.

5. CONCLUSIONS

In abbreviated form, the conclusions drawn in this thesis are as follows:

General:

1. Swets's original formalism is ambiguous and inadequate:
   - 1.1 It introduces certain key concepts ambiguously and/or with insufficient formal definition. In particular, the concepts of 'relevance' and 'question (or query)' are so introduced.
   - 1.2 It attempts, for no given reason, to describe a hypothetical retrieval process defined by a confounding of a set of retrieval processes as we have defined them, and thereby totally obscures variability in individual processes within such a set.
   - 1.3 It does not recognise variability in the query, as one determinant of the (individual) retrieval process.
   - 1.4 It does not integrate logical search expressions into the framework of the theory, i.e. into the formalism.
   - 1.5 It does not distinguish clearly between probabilistic measures of retrieval effectiveness such as Recall and Fallout, and measures of overall process effectiveness based on the two probability distributions.
   - 1.6 It almost totally ignores Precision as a measure of retrieval effectiveness of the probabilistic type, and also the Recall-Precision graph characterising each retrieval process.

1.7 It does not clearly state that the random variables defining each process are discrete with the continuous random variables introduced in the theory serving as models of them.

1.8 Hypotheses relating to the information retrieval process are suggested or implied rather than clearly put.

2. The formalism is nevertheless of considerable conceptual value when suitably interpreted and when extended in certain ways. It is then consistent and more complete, but it possesses limitations by virtue of being a macroscopic rather than a microscopic formalism. The most essential features of the revised formalism are (a) it relates to individual combinations of the triple: information need (evidenced as a partitioning of a data base), query (as a set of terms) and DOE, not to a compounding of such combinations; and (b) it explicitly incorporates the procedure of searching a data base using logical search expressions.

3. The formalism as such, and hypotheses expressed within it, require to be distinguished.

4. The data analyses of Swets cannot be compared with data described in the revised formalism, as (a) the distributions examined by Swets were both compounded and truncated, (b) the sets of relevant documents treated by

Swets were defined using an unsatisfactory methodology, (c) the results of Swets's analyses were not given in numerical form, and are solely expressed as Recall-Fallout graphs, not Recall-Precision graphs, (d) the Generality values of the sets of relevant documents were not specified, and (e) the queries forming components of the retrieval processes are not standardised, i.e. generated algorithmically from the data base partitioning. Novel data is accordingly required.

Relating to the experiment described in the thesis

1. Of the three analytical DOEs examined in the experiment, the optimal one is that based on co-ordination level, for queries generated by either of the methods described and for optimality in the retrieval process defined by either (a) mean Euclidean distance from the origin to the Recall-Precision graph, or (b) mean value of Brookes's measure of effectiveness. However this DOE was not significantly superior to the other two DOEs in the former case. The two DOEs incorporating information on term specificity were both significantly superior to co-ordination level, for queries generated algorithmically by either method, when optimality in the retrieval process was judged by the mean value of a metric distance between the sets of retrieved documents and the set of relevant documents. Neither of these two DOEs was significantly better than the other.
2. The algorithm used to generate query terms by the term-likelihood ratio method defined retrieval processes that

were more effective than those defined using query terms generated by a clustering method. This was true for all DOEs and for all three criteria for assessing process effectiveness, although the superiority of the former algorithm was not always statistically significant.

3. The probability distributions for relevant [non-relevant] documents have approximately 0.252 [0.987] of the probability concentrated on the all-negated elementary conjunct, for queries containing 5 terms and defined by the term likelihood ratio method. The most likely estimate of the maximum value of Recall [Fallout] for logical expressions not disjoined to the all-negated elementary conjunct is accordingly 0.748 [0.013]. Thus on this evidence one would appear to be able to retrieve only 75% of relevant documents from MEDLARS before there is a catastrophic increase in the number of retrieved documents, for this size and type of question. It follows that for high-Recall searching, a high value of n must be used.
4. The most efficient groups of information retrieval processes identified in the experiment had mean properties as follows. The questions are generated by the term likelihood ratio method.

Statistic	Retrieval processes defined by co-ordination level rankings	Retrieval processes defined by Miller's function ranking
mean rank value of f_1	2.493	9.324
mean rank value of f_2	1.012	1.021
variance of rank of f_1	1.275	64.92
variance of rank of f_2	0.012	0.073
Brookes's S measure	1.391	1.060
Mean distance from (0,0) to (R,P) graph	0.631	0.565
Mean value of metric distance between rel. set and retr. sets	0.962	0.935

5. The mean and variance of f_1 correlated significantly (≤ 0.05) and positively for sets of processes defined by queries of the more efficient type (i.e. defined by term likelihood ratios) for each DOE. The mean of f_1 and Generality also correlated significantly (≤ 0.016) and negatively for all four DOE for processes defined by questions that were so formed.
6. Of two pairs of modelling functions examined, it was found that, for processes defined by the more efficient method of query generation, and for 'difference between equivalent (R,P) co-ordinates' as the criterion variable, the binomial model was not an acceptable model of the distributions involved. Insofar as the binomial distribution represents the discrete distribution 'closest' to the Normal distributions considered by Swets, the hypothesis that the distributions can be modelled by Normal distributions, implicit in Swets's work although

not formally stated by him, must be considered to be in doubt. The uniform model is, on the other hand, acceptable for the co-ordination level DOE but not for the other two analytical functions based on term specificity values. Good modelling functions have yet to be discovered.

A more informal, and fuller summary is as follows. We have examined the description of information retrieval as a signal detection process put forward by J.A. Swets. As a result of the examination it was found that the description had various deficiencies. The chief of these were that the description did not clearly distinguish between formalism and model; it was expressed using continuous random variables which had no counterpart in reality (but which were nevertheless of possible use in modelling observed data); and it used certain concepts and terms loosely and inconsistently - for example input signal, query, and relevance. In view of these deficiencies, a re-expression and modification of the formalism was introduced, which we have referred to as an 'extension' of the original formalism. This gave unambiguous meanings to the terms in the formalism, it distinguished clearly between formalism and modelling functions within the formalism (viewing the continuous random variables of Swets as modelling variables), and it extended the formalism by basing it clearly on outcome events related in a definite manner to the subsets of a set of terms defining a query. These subsets were also identified with elementary logical conjuncts of the

latter set. The proper expression of the formalism is through the description of what we have termed 'retrieval processes', defined by a combination of query (as SFQ), partitioning of the data base, and DOE. This is in contrast to the approach of Swets which viewed confounded data pertaining to sets of such processes as a single signal detection process (at least in one interpretation of his contribution.) In the approach adopted here, the variation of such individual processes (defined by variation in one or other of the components, such as the partitioning of the data base) then naturally becomes the object of scientific interest. This in fact was the main motivation for the experimental work we have described, although a further motivation was the conclusion, again arising from the analysis of Swets's contribution, that previous experimental work in information retrieval is of unproven validity where it has been based on the (pseudo) concept of "relevance to a question", and where data has not been 'proofed' against changes in weight values that leave the ordering of probability pairs unaffected. This thesis maintains that as the term is used 'relevance' is not an entity capable of unambiguous description in language but is a primitive entity lying outside experimental controls and describable only through its effects on human behaviour. In the extended formalism a query is seen as a variable entity, not a static one, which provides not a criterion for relevance decisions but a device with which to explore a data base. The exploration is assumed to be carried out in an algorithmic manner ('algorithmic' in a static mathematical sense, rather than in a dynamic programming sense although obviously the way in which the data base is searched by

a program will be influenced by the query) and for documents that are assumed to be flagged (as relevant or non-relevant) in some *a priori* way. The flags are not known previous to the retrieval process being defined but the query's definition is, or attempts to be, a perception of the way in which they are distributed. The Swetsian formalism, so extended, also clearly distinguishes between modelling and observed probability distributions, only the latter being essential to the definition of the retrieval process. Of particular concern in the distributions is the very large spike of probability that attaches to the all-negated elementary conjunct for non-relevant documents, which has been totally ignored in previous work. This probability determines the amount of probability distributed over the tail of the distribution, previous work having disregarded the conditional character of probability values in the tail. The Normality of the modelling distributions introduced by Swets is not seen as being of critical importance (just as modelling *per se* is not so seen) except insofar that it has prompted observations already made: that modelling needs to be contrasted with formal description, and that the Normal model is vulnerable to the criticism that it ignores the all-negated elementary conjunct of query terms. It is claimed that no individual analytical form of modelling function (or pair of same) can 'test' the Swetsian formalism. The 'testing' approach, it has been argued, must be primarily directed at hypotheses expressed in the discrete formalism and identifying values of population parameters. However, tests of whether the distributions making up individual processes are likely to have been drawn from population distributions

defined by modelling (analytical) functions may also be useful; both as a basis for further, simulation work on information retrieval, and to provide structural criteria by which theories at a deeper level of explanation than Swets's can be judged.

Stemming from the basic notion of this thesis that the individual retrieval process is the prime object of interest (both for conceptual clarity, and to provide a basic unit of observation) and must be described in a discrete formalism, are various consequences. First is the notion that 'effectiveness' is a property of (1) each process taken as a whole, and (2) of an implementation of that process defined by the application of a rank threshold value to determine retrieval and non-retrieval. We have carried this fundamental distinction through this thesis. Secondly the notion of 'optimality' in the process is suggested. We have claimed that variation in the DOE yields only a sub-optimal process (for a given partitioning of the data base by an instance of information need); just as variation in the query (as SFQ) yields only a sub-optimal process. From a signal detection point of view, optimisation must be seen as a joint procedure with query (as SFQ) and DOE jointly varying so as to maximise either some chosen measure of overall process effectiveness or, for a given threshold value, some function of the probabilistic measures R and P. Thirdly, the concept of a query as a logical search expression fits naturally into the extended formalism in a way that we have described: individual document weights are associated with disjunctions of one or more elementary logical conjuncts of the query. We have distinguished the two types of query throughout this thesis, and we recall that whereas a query as a set of terms

implies a value for an overall measure of retrieval effectiveness (for a given DOE and a given partitioning of the data base by information used), the query as a logical search expression determines values for each probabilistic measure of retrieval effectiveness. If a query (as a set of terms) is sub-optimal, the elementary conjuncts of its terms will imply, when disjoined from highest rank value to lesser rank values, a sequence of sub-optimal queries as logical search expressions. Fourthly, the introduction of bivariate or multivariate weights attaching to documents (by the retrieval process, not in its input as considered by Hutchinson) allows the formalism to be extended in a natural way, for example by allowing the age of a document to be a contributing variable to the document's weight. Fifthly, the signal detection formalism suggests a heuristic technique of retrieval whereby the 'contrast' between a set of retrieved relevant documents and a set of retrieved non-relevant documents determines a DOE that is more efficient than a predecessor DOE. (It has been suggested that linear discriminant analysis provides a DOE in this context having good prima facie value.)

An experiment has been described in Section 4 which was designed to obtain as much data as possible concerning the characteristics of processes relating to a particular data base (MEDLARS) and a particular context of information need (in medicine). The processes examined, 240 in number, were defined through controlled variation in the DOE and query (in set form), the latter being generated algorithmically in two ways. A large range of G values was involved. The experiment also intended to provide a prototype of a retrieval experiment in that, unlike some or all previous experi-

ments (1) the partitionings of the data base by instances of information need were based on objective data not influenced by the experiment and (2) probability data were stored in association with the elementary logical conjuncts of query terms to which they pertained (rather than with the weights to which such conjuncts are mapped); and (3) the queries (as sets of terms) used to define the processes were generated algorithmically from the partitioning of the data base concerned. Hypotheses of the two types we have discussed can be inferred from the data given in Section 4, though not with complete plausibility since the sample of partitionings of the data base that was used was not a random one. (The sample was (1) intended to identify extremes of variation in processes, and (2) of a quota character.) Given the absence of previous comparable data, the experiment was regarded as 'hypothesis generating' rather than 'hypothesis testing' in nature, though where the adequacy of modelling distributions was concerned the data did allow investigation of their validity. (In the latter case the conclusions are again affected by the non-randomness of the sample.) Partial orders on the set of four DOE's employed were identified, using three distinct measures of overall retrieval effectiveness.

Lastly, it is claimed that the framework of thinking embodied in the extended Swetsian formalism is useful in two respects more fundamental than ones mentioned above. First it can be said to be 'linguistically constructive' in that the expression of the formalism imposes a clear terminology relating to its components: 'relevance' in a document is seen as a binary attribute, for example, fixed for each document (for a given retrieval process)

and distinct from the labelling (i.e. identification) of documents as "relevant" and "non-relevant" implied by the imposition of a threshold value for a given process. Again a 'query' is construed as either a set of document attributes or as a logical expression, or (historically) as an expression in everyday language descriptive of an information need. Retrieval 'effectiveness' is seen as an ambiguous concept, and measures of same are required to be appropriately qualified. 'Weighting' is also seen as ambiguous: document weighting and term weighting are distinct procedures, the former being the more fundamental procedure and capable of generalisation to cover the case of vectors of weights attaching to documents. Secondly, the formalism is 'hospitable' in the sense that the main concepts of information retrieval are naturally and easily accommodated in it: the data base, the relevance judgement, the query, the DOE, and retrieval effectiveness, once certain necessary distinctions and definitions have been made. Other less-major concepts such as Precision and Recall, term specificity, heuristics, and optimality are also readily describable in it.

The extension of Swets's formalism that we have offered is a conceptually robust, simple and hospitable framework for the description of information retrieval.

5.1 Technological implications of the findings of the thesis

In the author's view, there is considerable scope for improving the nature of the dialogue between enquirer on the one hand, and data base management program on the other hand. In conventional DBMPs, the enquirer both (1) constructs a set of search terms, and (2) relates these in a logical search expression. Since it is unlikely that either is optimum, what is needed is primarily a DBMS which constructs the logical search expression for a set of query terms and a specified level of Recall. The optimum such expression is then created algorithmically by a DOE. Experiments such as the one we have described allow the optimum DOE to be identified. The extended formalism allows the interactions involved to be portrayed clearly. Secondly, a DBMP can usefully accommodate a heuristic feature such as the one described in this thesis. Again the advantage of the formalism is the clear portrayal of the heuristic process.

5.2 Suggestions for further research

It is suggested that useful work could be undertaken in the following areas: (1) the investigation of the variation of the query (as a set of terms) on the retrieval process when information need and DOE are fixed, especially the investigation of whether user-specified, rather than algorithmically-derived, queries are sub-optimum or not, and if sub-optimum to what extent so; and (2) systematic investigation, similar to the investigation described in this thesis, in respect of data bases other than MEDLINE.

Appendix A: Theorems Relevant to Compositions of the Elementary Logical Conjuncts of Query Terms

Two results related to the discussion of Section 3.3.2.3 are given here.

Since a composition is, by definition, a partition of the members of a permutation, it is of interest whether, when different compositions of a set of elementary conjuncts are considered, the Recall-Precision graphs are the same. We assert the truth of the following two theorems:

1. Two different compositions of the same permutation of elementary conjuncts may generate identical R-P graphs.
2. Two different permutations of a set of elementary conjuncts may have compositions that generate identical R-P graphs.

Intuitive proofs are as follows:

1. Consider the set of elementary conjuncts generated by a query of two terms:

$$c_1 = t_a^o \wedge t_b^o$$
$$c_2 = t_a^o \wedge t_b^1$$
$$c_3 = t_a^1 \wedge t_b^o$$
$$c_4 = t_a^1 \wedge t_b^1$$

Suppose probabilities associated with these elementary conjuncts are as follows:

$$c_1 : r_{ab}^{oo} \quad s_{ab}^{oo}$$
$$c_2 : r_{ab}^{ol} \quad s_{ab}^{ol}$$
$$c_3 : r_{ab}^{lo} \quad s_{ab}^{lo}$$
$$c_4 : r_{ab}^{ll} \quad s_{ab}^{ll}$$

Then consider the following compositions:

$C1 = (c_1 \mid c_2 \quad c_3 \mid c_4)$ and $C2 = (c_1 \mid c_2 \mid c_3 \quad c_4)$.

Suppose $r_{ab}^{10} = 0 = s_{ab}^{10}$, i.e. the event c_3 is 'compound almost impossible'. Then the arrays of logical search expressions corresponding to the above partitions are as follows, with their corresponding probability values.

For C1:

$e_1 = c_1 \qquad r_{ab}^{oo} \quad s_{ab}^{oo}$

$e_2 = c_2 \vee c_3 \qquad r_{ab}^{ol} \quad s_{ab}^{ol}$

$e_3 = c_4 \qquad r_{ab}^{ll} \quad s_{ab}^{ll}$

For C2:

$e_1 = c_1 \qquad r_{ab}^{oo} \quad s_{ab}^{oo}$

$e_2 = c_2 \qquad r_{ab}^{ol} \quad s_{ab}^{ol}$

$e_3 = c_3 \vee c_4 \qquad r_{ab}^{ll} \quad s_{ab}^{ll}$

But the arrays of probability-pairs are the same. Therefore they must generate identical R-P graphs.

2. Consider the permutations:

$$c_1 \; c_2 \; c_3 \; c_4 \; ; \; c_1 \; c_3 \; c_2 \; c_4$$

of the above example. Consider further the compositions of these:

$C1 = (c_1 \mid c_2 \quad c_3 \mid c_4) \qquad C2 = (c_1 \mid c_3 \quad c_2 \mid c_4)$

Obviously, since $e_2 = c_2 \vee c_3$ for C1, and $e_2 = c_3 \vee c_2$, both yield the same search result, i.e. both are equivalent to the search expression $e_2 = c_2$. The R-P graphs of the two compositions are accordingly identical. The same arrays of probability obtain in

each case, and again are:

$$r_{ab}^{oo} \quad s_{ab}^{oo}$$

$$r_{ab}^{ol} \quad s_{ab}^{ol}$$

$$r_{ab}^{ll} \quad s_{ab}^{ll} \ .$$

-end of intuitive proof.

The number of Recall-Precision graphs that a query can engender cannot be known with certainty without a knowledge of the probabilities $r_{\cdots}^{\cdots}$ and $s_{\cdots}^{\cdots}$. However it is possible to examine the maximum number of distinguishable R-P graphs that can be determined, as the following illustrative example shows.

Consider as before a query of just two terms. The elementary conjuncts can be grouped as follows. In effect we note some of the different compositions (i.e. ordered partitions) of the c_i.

Mappings of the partitions to exactly four expressions e_i:

The compositions here are of type $(c_i \mid c_j \mid c_k \mid c_l)$. The individual partitions contain exactly one elementary conjunct. The number of such compositions is 4! Each such composition yields up to four points on the R-P graph.

Mappings of the partitions to exactly three expressions e_i:

The compositions here are of type $(c_i \mid \ldots \mid c_j)$, or $(c_i \mid c_j \mid \ldots)$, or $(\ldots \mid c_i \mid c_j)$, where the dots denote the set (not permutation) of remaining elementary conjuncts. The number of each of these three types is 12, i.e.

$$(c_1 \mid \ldots \mid c_j), \qquad j = 2,3,4$$

$$(c_2 \mid \ldots \mid c_j), \qquad j = 1,3,4 \ .$$

$$(c_3 \mid \ldots \mid c_j), \qquad j = 2,3,4$$

$$(c_4 \mid \ldots \mid c_j), \qquad j = 1,2,3$$

for compositions of the first type. There are thus 12 partitions that can yield distinct R-P graphs. Each such R-P graph can have up to three points.

Mappings of the partitions to exactly two expressions e_i:

The compositions here are of type $(c_j \mid \ldots)$, $j = 1,2,3,4$, or of type $(\ldots \mid c_j)$, $j = 1,2,3,4$. The number of such compositions is the eight. Each R-P graph has up to two points.

Mapping to one expression e_1

Technically, one can identify the single search expression $e_1 = c_1 \vee c_2 \vee c_3 \vee c_4$, equivalent to the composition (c_1, c_2, c_3, c_4). This search expression yields just one R-P graph with just one point.

The number of compositions of interest, in the case n=2, is thus

$$41 + 12 + 8 + 1$$

or 45. This is also the maximum number of distinguishable Recall-Precision graphs that a query of just two terms can engender. Some of these graphs may of course be identical, which will happen if some of the c_i share paired values for $(r^{\cdot\cdot\cdot}_{\cdot\cdot\cdot}, f^{\cdot\cdot\cdot}_{\cdot\cdot\cdot})$.

An upper bound on the maximum number of distinguishable R-P graphs, for a query of n terms, is the number of compositions of 2^n objects, namely:

$$\sum_{i=1}^{2^n} {}^{2^n}C_i \cdot i!$$

This expression evaluates to 136 in the case n=2, considerably in excess of the least upper bound to this maximum number, arrived at by enumerative means above.

Appendix B: Data on the Basic Partitionings of the Data Base

Reference label	Size of set of relevant documents (in brackets: no. of references in review paper)	Size of set of all documents of eligible date of publication in MEDLARS	Generality of information need for data base concerned
1	8 (32)	2,160,462	3.703'-6
2	11 (25)	1,993,205	5.519'-6
3	57 (68)	1,887,482	30.199'-6
4	16 (24)	2,236,049	7.155'-6
5	9 (23)	2,254,754	3.992'-6
6	12 (33)	2,272,895	5.280'-6
7	13 (25)	2,309,278	5.629'-6
8	98 (199)	2,122,182	46.179'-6
9	107 (174)	2,010,216	53.228'-6
10	20 (31)	2,181,012	9.170'-6
11	33 (42)	2,327,931	14.176'-6
12	6 (27)	2,290,777	2.619'-6
13	23 (28)	2,309,280	9.960'-6
14	16 (22)	2,425,245	6.597'-6
15	44 (150)	2,065,918	21.298'-6
16	21 (26)	2,254,752	9.314'-6
17	36 (95)	2,180,962	16.506'-6
18	25 (40)	2,200,231	11,362'-6
19	20 (26)	2,217,806	9.018'-6
20	32 (59)	2,200,230	14.544'-6
21	3 (32)	2,200,231	1.363'-6

continued

22	19 (35)	1,938,096	9.803'-6
23	5 (38)	2,103,019	2.378'-6
24	17 (21)	2,200,113	7.727'-6
25	19 (43)	2,200,229	8.635'-6
26	24 (37)	2,180,963	11.004'-6
27	78 (147)	2,065,917	37.756'-6
28	18 (36)	2,290,735	7.858'-6
29	19 (28)	1,342,041	14.158'-6
30	23 (35)	1,774,029	12.965'-6
31	12 (27)	2,200,230	5.454'-6

Statistics:	mean	standard deviation (n-1)
no. of relevant documents	27.2	25.3
no. of documents	2,141,810	2.03'5
Generality	13.05'-6	12.5'-6

The median number of relevant documents was 19.

Appendix C: Main Results of the Experiment

The main results of the experiment are contained in the following tables. The variable names referred to in the tables are defined as follows:

DS01 Mean rank for relevant documents.

DS02 Mean rank for non-relevant documents.

DS03 Variance of rank for relevant documents.

DS04 Variance of rank for non-relevant documents.

DS05 Mean in tail for non-relevant documents (redefined over rank values: r'=rank-1).

DS06 Variance in tail for non-relevant documents.

DS07 Skewness (third moment about mean) of distribution for relevant documents.

DS08 Kurtosis (fourth moment about mean) of distribution for relevant documents.

DS09 Swets's E-value (redefined over rank values of weights)

DS10A Brookes's S-value (redefined over rank values of weights)

DS10B Mean value of the Euclidean distance from origin to points on the Recall-Precision graph for process.

DS10C Mean value of Marczewski-Steinhaus metric for process.

DS11X Kolmogorov-Smirnov statistic for modelling and observed distributions of relevant documents.

DS12X Kolmogorov-Smirnov statistic for modelling and observed distributions of non-relevant documents.

DS13X Mean Euclidean distance between comparable (R,P) co-ordinates.

DS14X Variance of Euclidean distance between comparable (R,P) co-ordinates.

DS15X Range of Euclidean distance between comparable (R,P) co-ordinates.

DS16X Mean difference between M-S distances separating comparable pairs of the sets: set of retrieved documents and set of relevant documents.

DS17X Variance of the difference defined for DS16X.

DS18X Range of the difference defined for DS16X.

The suffix "X" to variables DS11 to DS18 takes one of the two values "B" or "U" denoting the type of model pertaining, binomial or uniform-discrete respectively. G, as usual, denotes the Generality of the information need as reflected in the partitioning of the data base by relevance judgements.

Tables C-1 to C-8 summarise the values of six descriptive statistics (mean, variance (n-1), minimum, maximum, skewness, kurtosis) for each of the above variables, for various groupings of the 240 retrieval processes examined. The groupings are by choice of weighting function and query type, eight groups in all. The statistics in these tables are "statistics of statistics", i.e. they summarise properties of the random variables describing each retrieval process, for various sets of processes. As previously mentioned the number of retrieval processes for each DOE is 31 for processes defined by queries of type QFORM1, and 29 for those defined by queries of type QFORM2. However, also added for comparison are the equivalent values for variables DS01 to DS10A when the sets of relevant documents are restricted to those defined by the six review papers included in one issue of one journal (Section 4.1.2), although this information is not given for the weighting function W3. The tables affected record the number of processes

concerned in column 2.

Tables C-9 and C-10 summarise the significance values, grouped into intervals, of the Kolmogorov-Smirnov statistic for the eight categories of process and two types of modelling distribution considered.

Tables C-11 to C-14 summarise the values of the correlations (i.e. normalised covariances) between selected pairs of random variables for the eight categories of process. In addition, an estimated level of significance of each value is supplied indicating the probability that the sample (of paired values) is taken from a population in which each variable is distributed Normally and independently of the other. (A smaller significance value tends to refute the hypothesis.)

Lastly, we note here that the variation in the value of the "spike" of probability present in the observed distribution for non-relevant documents, attached to the all-negated elementary conjunct, was examined for the 31 queries of type QFORM1. It was found that the mean value of this probability was 0.987 with a standard deviation (n-1) of 0.029.

Table C-1. The set of processes defined by W1 and QFORM1. (31 or 6 processes [see text].)*

	n	mean	Var(n-1)	Min.	Max.
DS01	31	2.493	0.289	1.886	4.333
	6	2.217	0.058	1.886	2.611
DS02	31	1.012	0.001	1.000	1.149
	6	1.002	0.000	1.000	1.007
DS03	31	1.275	0.431	0.333	3.474
	6	1.031	0.078	0.619	1.405
DS04	31	0.012	0.001	0.000	0.128
	6	0.00	0.000	0.000	0.008
DS05	31	1.063	0.003	1.000	1.195
	6	1.061	0.002	1.000	1.108
DS06	31	0.068	0.003	0.000	0.231
	6	0.069	0.002	0.000	0.018
DS07	31	0.252	0.353	-1.004	1.632
	6	0.687	0.079	0.452	1.238
DS08	31	-0.537	0.630	-1.500	2.570
	6	-0.153	0.144	-0.705	0.472
DS09	31	2,598	0.758	1.227	4.855
	6	2.377	0.508	1.424	3.282
DS10A	31	1.391	0.238	0.619	2.798
	6	1.245	0.148	0.745	1.709
DS10B	31	0.631	0.023	0.325	0.924
	6	0.666	0.042	0.410	0.924
DS10C	31	0.962	0.001	0.881	0.999
	6	0.952	0.001	0.917	0.986
DS11B	31	0.389	0.005	0.262	0.556
DS11U	31	0.385	0.005	0.250	0.509
DS12B	31	0.008	0.000	0.000	0.101
DS12U	31	0.009	0.001	0.000	0.118
DS13B	31	0.374	0.030	0.114	0.831
DS13U	31	0.334	0.037	0.083	0.833
DS14B	31	0.212	0.034	0.022	0.783
DS14U	31	0.203	0.043	0.007	0.825
DS15B	31	0.426	0.078	0.010	0.920
DS15U	31	0.467	0.117	0.000	1.013
DS16B	31	-0.024	0.001	-0.099	0.005
DS16U	31	-0.028	0.001	-0.104	0.006
DS17B	31	0.029	0.001	-0.002	0.123
DS17U	31	0.033	0.001	-0.010	0.124
DS18B	31	0.176	0.025	0.005	0.606
DS18U	31	0.190	0.027	0.005	0.633

*Values in Tables C-1 to C-8 are rounded to the third decimal place

Table C-2. The set of processes defined by W1 and QFORM2. (29 or 6 processes [see text].)

	n	mean	var(n-1)	Min.	Max.
DS01	29	2.414	0.375	1.602	4.333
	6	1.953	0.041	1.701	2.219
DS02	29	1.018	0.000	1.000	1.093
	6	1.013	0.000	1.002	1.040
DS03	29	1.731	0.783	0.551	3.897
	6	1.255	0.332	0.695	2.176
DS04	29	0.019	0.000	0.000	0.090
	6	0.014	0.000	0.002	0.039
DS05	29	1.056	0.005	1.002	1.275
	6	1.073	0.003	1.012	1.169
DS06	29	0.056	0.004	0.002	0.262
	6	0.077	0.003	0.014	0.174
DS07	29	0.535	0.415	-0.730	1.746
	6	0.891	0.131	0.328	1.295
DS08	29	-0.545	0.706	-1.608	1.426
	6	-0.089	0.763	-1.260	0.981
DS09	29	2.023	0.629	1.019	4.777
	6	1.575	0.056	1.349	1.994
DS10A	29	1.113	0.217	0.523	2.798
	6	0.859	0.012	0.768	1.037
DS10B	29	0.514	0.027	0.244	0.835
	6	0.421	0.007	0.322	0.506
DS10C	29	0.962	0.001	0.890	0.999
	6	0.982	0.000	0.973	0.996
DS11B	29	0.405	0.006	0.262	0.584
DS11U	29	0.379	0.009	0.167	0.534
DS12B	29	0.012	0.000	0.000	0.161
DS12U	29	0.013	0.000	0.000	0.007
DS13B	29	0.371	0.016	0.133	0.758
DS13U	29	0.336	0.020	0.122	0.711
DS14B	29	0.185	0.019	0.036	0.634
DS14U	29	0.168	0.020	0.015	0.600
DS15B	29	0.342	0.085	0.013	0.958
DS15U	29	0.346	0.096	0.014	0.969
DS16B	29	-0.034	0.001	-0.104	-0.001
DS16U	29	-0.035	0.001	-0.107	-0.001
DS17B	29	0.040	0.002	0.001	0.124
DS17U	29	0.041	0.002	0.001	0.128
DS18B	29	0.211	0.036	0.006	0.625
DS18U	29	0.217	0.038	0.006	0.643

Table C-3. The set of processes defined by W2 and QFORM1. (31 or 6 processes [see text].)

	n	Mean	var(n-1)	Min.	Max.
DS01	31	9.324	15.944	4.737	23.667
	6	7.400	1.052	5.932	8.722
DS02	31	1.021	0.002	1.000	1.167
	6	1.006	0.000	1.001	1.021
DS03	31	64.915	1024.4	10.993	156.3
	6	54.688	199.7	33.439	72.531
DS04	31	0.073	0.017	0.001	0.632
	6	0.035	0.002	0.002	0.133
DS05	31	2.482	0.514	1.043	4.087
	6	2.649	0.281	2.041	3.278
DS06	31	5.082	15.016	0.326	17.869
	6	5.135	13.919	0.798	10.453
DS07	31	0.819	0.359	-0.567	1.775
	6	1.239	0.067	0.990	1.619
DS08	31	-0.058	1.490	-1.550	2.706
	6	0.580	0.412	-0.243	1.258
DS09	31	2.058	0.420	1.130	4.401
	6	1.732	0.115	1.199	2.127
DS10A	31	1.060	0.121	0.567	2.377
	6	0.884	0.031	0.614	1.080
DS10B	31	.0.565	0.046	0.187	0.957
	6	0.589	0.089	0.279	0.957
DS10C	31	0.935	0.003	0.821	0.996
	6	0.919	0.003	0.831	0.983
DS11B	31	0.459	0.009	0.289	0.740
DS11U	31	0.448	0.014	0.198	0.676
DS12B	31	0.006	0.000	0.000	0.091
DS12U	31	0.010	0.001	0.000	0.136
DS13B	31	0.442	0.026	0.099	0.708
DS13U	31	0.432	0.041	0.161	0.944
DS14B	31	0.286	0.030	0.019	0.591
DS14U	31	0.270	0.055	0.026	0.921
DS15B	31	0.688	0.092	0.009	1.011
DS15U	31	0.593	0.100	0.031	1.005
DS16B	31	0.014	0.006	-0.113	0.310
DS16U	31	-0.052	0.002	-0.160	0.000
DS17B	31	0.019	0.001	-0.039	0.140
DS17U	31	0.061	0.003	0.000	0.199
DS18B	31	2.087	5.607	0.054	10.867
DS18U	31	1.710	1.721	0.068	4.972

Table C-4. The set of processes defined by W2 and QFORM2. (29 or 6 processes [see text].)

	n	Mean	var(n-1)	Min.	Max.
DS01	29	9.082	18.281	3.531	23.667
	6	5.932	1.836	4.318	7.875
DS02	29	1.030	0.001	1.001	1.141
	6	1.024	0.000	1.003	1.055
DS03	29	80.283	1356.3	17.880	156.3
	6	51.687	846.77	26.269	100.11
DS04	29	0.104	0.019	0.002	0.632
	6	0.092	0.002	0.015	0.141
DS05	29	2.065	0.699	1.043	4.494
	6	2.364	0.325	1.387	2.984
DS06	29	3.732	12.033	0.242	14.31
	6	4.700	6.567	1.524	8.635
DS07	29	0.939	0.470	-0.567	2.166
	6	1.458	0.283	0.658	2.166
DS08	29	-0.01	2.499	-1.715	4.416
	6	1.322	4.364	-1.222	4.416
DS09	29	1.753	0.488	0.961	4.401
	6	1.358	0.047	1.095	1.725
DS10A	29	0.908	0.139	0.487	2.377
	6	0.707	0.012	0.571	0.879
DS10B	29	0.482	0.037	0.234	0.866
	6	0.370	0.013	0.258	0.579
DS10C	29	0.938	0.002	0.820	0.995
	6	0.970	0.000	0.950	0.991
DS11B	29	0.529	0.005	0.364	0.764
DS11U	29	0.489	0.017	0.202	0.689
DS12B	29	0.009	0.000	0.000	0.042
DS12U	29	0.015	0.000	0.000	0.079
DS13B	29	0.490	0.022	0.238	0.706
DS13U	29	0.435	0.025	0.163	0.939
DS14B	29	0.321	0.028	0.074	0.565
DS14U	29	0.246	0.032	0.040	0.094
DS15B	29	0.669	0.085	0.030	1.086
DS15U	29	0.454	0.089	0.004	1.005
DS16B	29	0.012	0.007	-0.120	0.307
DS16U	29	-0.059	0.002	-0.176	-0.005
DS17B	29	0.025	0.001	-0.028	0.149
DS17U	29	0.069	0.003	0.005	0.215
DS18B	29	2.262	6.099	0.180	10.707
DS18U	29	1.855	1.930	0.172	4.472

Table C-5. The set of processes defined by W3 and QFORM1. (31 processes)

	Mean	var(n-1)	Min.	Max.
DS01	6.280	1.262	3.692	8.917
DS02	2.732	0.394	1.000	3.036
DS03	9.333	42.833	1.000	27.724
DS04	0.038	0.009	0.000	0.506
DS05	1.976	0.086	1.001	2.559
DS06	0.056	0.037	0.000	1.024
DS07	0.354	0.134	-0.313	1.119
DS08	-1.031	0.159	-1.668	-0.098
DS09	2.468	0.940	1.308	5.886
DS10A	1.290	0.245	0.671	2.999
DS10B	0.777	0.022	0.551	1.067
DS10C	0.906	0.003	0.784	0.993
DS11B	0.376	0.006	0.267	0.667
DS11U	0.306	0.007	0.214	0.600
DS12B	0.564	0.036	0.000	0.687
DS12U	0.718	0.062	0.000	0.952
DS13B	0.335	0.016	0.125	0.644
DS13U	0.374	0.034	0.079	0.806
DS14B	0.225	0.025	0.029	0.606
DS14U	0.270	0.041	0.009	0.770
DS15B	0.822	0.072	0.303	1.075
DS15U	0.806	0.073	0.174	1.001
DS16B	-0.084	0.003	-0.215	-0.005
DS16U	-0.093	0.003	-0.216	-0.007
DS17B	0.107	0.007	0.005	0.320
DS17U	0.118	0.007	0.007	0.320
DS18B	0.974	0.374	0.053	2.603
DS18U	1.061	0.583	0.078	3.720

Table C-6. The set of processes defined by W3 and QFORM2. (29 processes)

	Mean	var(n-1)	Min.	Max.
DS01	5.456	1.089	3.526	7.545
DS02	2.869	0.268	1.000	3.114
DS03	6.647	22.495	0.819	18.616
DS04	0.046	0.002	0.000	0.185
DS05	2.026	0.004	2.000	2.318
DS06	0.044	0.008	0.000	0.451
DS07	0.531	0.258	-0.329	1.538
DS08	-0.826	0.597	-1.733	1.202
DS09	2.068	0.991	1.089	5.886
DS10A	1.106	0.257	0.582	2.999
DS10B	0.744	0.034	0.481	1.075
DS10C	0.920	0.004	0.762	0.992
DS11B	0.414	0.007	0.225	0.667
DS11U	0.353	0.009	0.223	0.600
DS12B	0.587	0.027	0.000	0.687
DS12U	0.691	0.042	0.000	0.866
DS13B	0.338	0.018	0.181	0.644
DS13U	0.347	0.026	0.093	0.687
DS14B	0.239	0.031	0.054	0.661
DS14U	0.234	0.031	0.013	0.624
DS15B	0.807	0.078	0.382	1.093
DS15U	0.710	0.107	0.137	0.982
DS16B	-0.075	0.004	-0.238	-0.008
DS16U	-0.080	0.004	-0.238	-0.008
DS17B	0.097	0.008	0.009	0.357
DS17U	0.102	0.009	0.009	0.357
DS18B	0.675	0.193	0.059	1.820
DS18U	0.722	0.232	0.059	1.840

Table C-7. The set of processes defined by W4 and QFORM1. (31 or 6 processes [see text].)

	n	Mean	var(n-1)	Min.	Max.
DS01	31	9.610	15.892	4.737	23.667
	6	7.702	1.473	5.932	9.472
DS02	31	1.021	0.002	1.000	1.167
	6	1.007	0.000	1.001	1.025
DS03	31	67.994	1194.3	14.205	156.3
	6	54.520	234.66	31.619	73.604
DS04	31	0.076	0.017	0.001	0.632
	6	0.042	0.004	0.003	0.116
DS05	31	2.628	0.611	1.043	4.087
	6	2.915	0.460	2.041	3.914
DS06	31	5.280	15.403	0.326	17.869
	6	5.506	14.505	0.798	11.222
DS07	31	0.788	0.363	-0.567	1.739
	6	1.181	0.097	0.791	1.619
DS08	31	-0.118	1.412	-1.577	2.706
	6	0.533	0.414	-0.258	1.258
DS09	31	2.087	0.436	1.114	4.401
	6	1.830	0.211	1.199	2.355
DS10A	31	1.075	0.125	0.558	2.377
	6	0.937	0.060	0.614	1.246
DS10B	31	0.565	0.046	0.187	0.958
	6	0.595	0.085	0.279	0.958
DS10C	31	0.934	0.003	0.811	0.997
	6	0.918	0.004	0.828	0.987
DS11B	31	0.462	0.010	0.290	0.761
DS11U	31	0.447	0.014	0.199	0.676
DS12B	31	0.006	0.000	0.000	0.091
DS12U	31	0.010	0.001	0.000	0.136
DS13B	31	0.454	0.027	0.099	0.761
DS13U	31	0.429	0.041	0.161	0.970
DS14B	31	0.299	0.033	0.019	0.659
DS14U	31	0.267	0.055	0.026	0.964
DS15B	31	0.698	0.088	0.044	1.016
DS15U	31	0.602	0.097	0.050	1.005
DS16B	31	0.014	0.006	-0.102	0.310
DS16U	31	-0.053	0.002	-0.170	0.000
DS17B	31	0.019	0.001	-0.041	0.137
DS17U	31	0.062	0.003	0.000	0.205
DS18B	31	2.127	5.614	0.054	10.867
DS18U	31	1.773	1.963	0.068	5.115

Table C-8. The set of processes defined by W4 and QFORM2. (29 or 6 processes [see text].)

	n	Mean	var(n-1)	Min.	Max.
DS01	29	9.205	18.699	3.469	23.667
	6	5.941	1.815	4.318	7.875
DS02	29	1.031	0.001	1.001	1.142
	6	1.024	0.000	1.003	1.055
DS03	29	82.864	161.94	16.582	156.3
	6	51.712	846.56	26.075	100.11
DS04	29	0.107	0.020	0.002	0.632
	6	1.024	0.000	1.003	1.055
DS05	29	2.111	0.708	1.043	4.548
	6	2.381	0.349	1.387	3.070
DS06	29	3.677	9.917	0.242	10.872
	6	4.697	6.652	1.524	8.635
DS07	29	0.925	0.459	-0.567	2.166
	6	1.451	0.275	0.658	2.166
DS08	29	-0.060	2.410	-1.714	4.416
	6	1.289	4.200	-1.222	4.416
DS09	29	1.765	0.505	0.961	4.401
	6	1.360	0.046	1.095	1.725
DS10A	29	0.915	0.144	0.487	2.377
	6	0.709	0.012	0.571	0.879
DS10B	29	0.484	0.037	0.234	0.900
	6	0.376	0.014	0.258	0.579
DS10C	29	0.937	0.003	0.810	0.998
	6	0.970	0.000	0.950	0.992
DS11B	29	0.532	0.009	0.364	0.768
DS11U	29	0.487	0.017	0.261	0.703
DS12B	29	0.009	0.000	0.000	0.042
DS12U	29	0.015	0.000	0.000	0.079
DS13B	29	0.470	0.021	0.238	0.736
DS13U	29	0.436	0.025	0.192	0.969
DS14B	29	0.299	0.028	0.074	0.649
DS14U	29	0.247	0.033	0.057	0.964
DS15B	29	0.658	0.091	0.030	1.123
DS15U	29	0.484	0.103	0.004	1.005
DS16B	29	0.012	0.007	-0.108	0.307
DS16U	29	-0.060	0.002	-0.186	-0.002
DS17B	29	0.026	0.001	-0.028	0.145
DS17U	29	0.070	0.004	0.002	0.227
DS18B	29	2.335	6.401	0.153	10.767
DS18U	29	1.929	2.369	0.063	5.587

Table C-9. Variation in the significance value of the Kolmogorov-Smirnov statistic, for 480 comparisons of observed distribution with modelling distribution, for sets of relevant documents. The table shows raw frequencies when values of the statistic are grouped into the intervals shown, for variation in query type, modelling distribution, and DOE.

	Significance value intervals					
	(1,0.20]	(0.20,0.15]	(0.15,0.10]	(0.10,0.05]	(0.05,0.01]	(0.01,0)
QFORM1 queries						
Binomial model						
W1	3	1	2	1	7	17
W2	3	0	0	3	4	21
W3	3	2	1	5	4	16
W4	3	0	0	3	3	22
Uniform model						
W1	5	0	2	1	3	20
W2	4	2	1	1	1	22
W3	10	1	4	7	2	7
W4	4	2	1	1	1	22
QFORM2 queries						
Binomial model						
W1	3	0	1	1	8	16
W2	1	0	0	2	1	25
W3	4	0	1	3	3	18
W4	1	0	0	2	1	25
Uniform model						
W1	5	2	1	0	4	17
W2	4	1	0	0	2	22
W3	3	1	3	4	5	13
W4	4	1	0	0	3	21

Table C-10. Variation in the significance value of the Kolmogorov-Smirnov statistic, for 480 comparisons of observed distribution with modelling distributions, for sets of non-relevant documents. The table shows raw frequencies when values of the statistic are grouped into the intervals shown, for variation in query type, modelling distribution, and DOE.

	Significance value intervals					
	(1,0.20]	(0.20,0.15]	(0.15,0.10]	(0.10,0.05]	(0.05,0.01]	(0.01,0)
QFORM1 queries						
Binomial model						
W1	31	0	0	0	0	0
W2	31	0	0	0	0	0
W3	3	0	0	0	1	27
W4	31	0	0	0	0	0
Uniform model						
W1	31	0	0	0	0	0
W2	31	0	0	0	0	0
W3	3	0	0	0	1	27
W4	31	0	0	0	0	0
QFORM2 queries						
Binomial model						
W1	29	0	0	0	0	0
W2	29	0	0	0	0	0
W3	2	0	0	0	1	26
W4	29	0	0	0	0	0
Uniform model						
W1	29	0	0	0	0	0
W2	29	0	0	0	0	0
W3	1	0	0	1	0	27
W4	29	0	0	0	0	0

Table C-11. Correlations of selected pairs of descriptive statistics for processes defined by W1 and QFORM1 (31 processes) or QFORM2 (29 processes).

Variable pairs		Queries of type QFORM1		Queries of type QFORM2	
		Correl.	estimated significance	Correl.	estimated significance
DS01	DS03	0.301	0.050	0.239	0.11
DS02	DS04	0.995	< 0.00001	0.997	< 0.00001
DS05	DS06	0.989	< 0.00001	0.991	< 0.00001
G	DS01	-0.440	0.0066	-0.567	0.00070
G	DS02	-0.210	0.13	-0.227	0.12
G	DS03	-0.261	0.078	-0.499	0.0029
G	DS05	-0.231	0.11	-0.0763	0.35
G	DS06	-0.203	0.14	-0.536	0.39
G	DS07	0.375	0.019	0.309	0.052
G	DS08	0.203	0.14	0.237	0.11
G	DS09	-0.270	0.071	-0.357	0.029
G	DS10A	-0.302	0.049	-0.344	0.034
G	DS10B	0.213	0.13	-0.441	0.008
G	DS10C	-0.258	0.081	0.324	0.043

Table C-12. Correlations of selected pairs of descriptive statistics for processes defined by W2 and QFORM1 (31 processes) or QFORM2 (29 processes).

Variable pairs		Queries of type QFORM1		Queries of type QFORM2	
		Correl.	estimated significance	Correl.	estimated significance
DS01	DS03	0.649	0.00004	0.558	0.008
DS02	DS04	0.824	<0.00001	0.894	<0.00001
DS05	DS06	0.813	<0.00001	0.927	<0.00001
G	DS01	-0.387	0.016	-0.553	0.0009
G	DS02	-0.241	0.096	-0.190	0.16
G	DS03	-0.309	0.046	-0.650	0.00007
G	DS05	0.0788	0.34	0.122	0.26
G	DS06	-0.167	0.18	-0.0360	0.43
G	DS07	0.413	0.010	0.553	0.0009
G	DS08	0.385	0.016	0.690	0.0002
G	DS09	-0.335	0.033	-0.384	0.020
G	DS10A	-0.340	0.031	-0.369	0.025
G	DS10B	0.278	0.065	-0.315	0.048
G	DS10C	-0.263	0.076	0.553	0.0009

Table C-13. Correlations of selected pairs of descriptive statistics for processes defined by W3 and QFORM1 (31 processes) QFORM2 (29 processes).

Variable pairs		Queries of type QFORM1		Queries of type QFORM2	
		Correl.	estimated significance	Correl.	estimated significance
DS01	DS03	0.629	0.00008	0.713	0.00001
DS02	DS04	0.0621	0.37	0.248	0.098
DS05	DS06	0.470	0.0038	0.975	<0.00001
G	DS01	0.458	0.0048	0.223	0.12
G	DS02	-0.0217	0.45	0.222	0.12
G	DS03	0.792	<0.00001	0.594	0.0003
G	DS05	-0.673	0.00002	-0.0814	0.33
G	DS06	-0.140	0.23	0.0686	0.36
G	DS07	0.400	0.013	0.428	0.010
G	DS08	0.555	0.0006	0.423	0.011
G	DS09	-0.247	0.090	-0.329	0.041
G	DS10A	-0.276	0.066	-0.351	0.031
G	DS10B	-0.278	0.44	-0.659	0.00005
G	DS10C	-0.284	0.44	+0.456	0.0065

Table C-14. Correlations of selected pairs of descriptive Statistics for processes defined by W4 and QFORM1 (31 processes) or QFORM2 (29 processes)

Variable pairs		Queries of type QFORM1		Queries of type QFORM2	
		Correl.	estimated significance	Correl.	estimated significance
DS01	DS03	0.571	0.00039	0.514	0.0022
DS02	DS04	0.822	< 0.00001	0.904	< 0.00001
DS05	DS06	0.769	< 0.00001	0.888	< 0.00001
G	DS01	-0.395	0.014	-0.561	0.0008
G	DS02	-0.243	0.094	-0.195	0.16
G	DS03	-0.320	0.040	-0.638	0.0001
G	DS05	0.101	0.29	0.111	0.28
G	DS06	-0.149	0.21	-0.0266	0.45
G	DS07	0.394	0.014	0.557	0.0009
G	DS08	0.359	0.024	0.703	0.00001
G	DS09	-0.311	0.044	-0.379	0.021
G	DS10A	-0.317	0.041	-0.364	0.026
G	DS10B	0.282	0.062	-0.319	0.045
G	DS10C	-0.262	0.077	0.353	0.030

References

ANDERSON, T.W. (1962) and BAHADUR, X.X. Classification into two normal distributions... Annals of mathematical statistics, 33, pp. 420-31.

ANGIONE, P.V. (1975) On the equivalence of Boolean and weighted searching based on the convertibility of query forms. Journal of the American Society for Information Science, 26, pp. 112-24.

BARHYDT, G.C. (1967) The effectiveness of non-user relevance assessments. Journal of documentation, 23, pp. 146-9.

BARKER, F.H. (1972), VEALE, D.C. and WYATT, B.K. Towards automatic profile construction. Journal of documentation, 28, pp.44-55.

BARKLA, J.K. (1969) The University of Sheffield biomedical information project. Information scientist, 3, pp. 13-30.

BARNES, R.C.M. (1964) The present state of information retrieval by computer. H.M.S.O. (also as U.K. Atomic Energy Establishment, Research Group, Report AERE-R-4514)

BARR, D.R. (1971) and ZEHNA, P.W. Probability. Belmont, Calif., Brooks/Cole.

BECKER, P.W. (1968) Recognition of patterns using the frequencies of occurrence of binary words. [D. Tech. Thesis] Technical University of Denmark.

BELNAP, N.D. (1976) and STEEL, T.B. The logic of questions and answers. New Haven, Yale Univ. Press. (Bibliography by EGLI, U. and SCHLEICHERT, H. The theory of questions and answers, as Appendix)

BENWELL, R.G. (1974) Automatic query construction. [M.Sc thesis] Computing Lab., University of Newcastle upon Tyne.

BOOKSTEIN, A. (1974) The anomalous behaviour of precision in the Swets model and its resolution. Journal of documentation, 30, pp. 374-80.

BOOKSTEIN, A. (1976) and COOPER, W.S. A general mathematical model for information retrieval systems. Library quarterly, 46, pp. 153-67.

BOOKSTEIN, A. (1977) When the most 'pertinent' document should not be retrieved - an analysis of the Swets model. Information processing & management, 13, pp. 377-83.

BOURNE, C.P. (1966) Evaluation of indexing systems. Annual review of information science and technology, 1, pp. 171-90.

BRANDHORST, R.T. (1966) Simulation of Boolean logic constraints theory to the use of term weights. American documentation, 17, pp. 145-6.

BROOKES, B.C. (1968) The measures of information retrieval effectiveness proposed by Swets. Journal of documentation, 24, pp. 41-54.

BROOKES, B.C. (1970) Obsolescence of special library periodicals: sampling errors and utility contours. Journal of the American Society for Information Science, 21, pp. 320-9.

BROOKES, B.C. (1972) The Shannon model of IR systems. Journal of documentation, 28, pp. 160-2.

CLEVERDON, C. (1966) and KEEN, M. Factors determining the performance of indexing systems: vol. 2 test results. Aslib.

CLEVERDON, C.W. (1967) The Cranfield tests of index language devices. Aslib proceedings, 19, pp. 173-94.

CLEVERDON, C.W. (1972) On the inverse relationship of recall and precision. Journal of documentation, 28, pp. 195-201.

CLEVERDON, C.W. (1976) and KIDD, J.S. Redundancy, relevance and value to the user in the outputs of information retrieval systems. Journal of documentation, 32, pp. 159-73.

COOMBS, C.H. (1970), DAWES, R.M. and TUFSKY, A. Mathematical psychology. Englewood Cliffs, N.J., Prentice Hall. [see esp. Chap. 6: The theory of signal detectability]

COOPER, W.S. (1968) Expected search length: a single measure of retrieval effectiveness based on weak ordering action of retrieval systems. American documentation, 19, pp. 30-41.

COOPER, W.S. (1973) On selecting a measure of retrieval effectiveness [2 pts]. Journal of the American Society for Information Science, 24, pp. 87-100, 413-24

COOPER, W.S. (1976) The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. Information processing & measurement, 12, pp. 367-75.

COOPER, W.S. (1977) The suboptimality of retrieval rankings based on probability of usefulness. [Private communication to S.E. Robertson, cited by Robertson (1977b)]

CORMACK, R.M. (1971) A review of classification. Journal of the Royal Statistical Society, A134, pp. 321-67.

DIFONDI, N.M. (1969) Statistical information retrieval system. Rome Air Development Center. (also as CFSTI Report AD-697403; the Center's Technical report RADC-TR-69-382)

EGAN, J.P. (1975) Signal detection theory and ROC analysis. N.Y., Academic Press.

EVANS, L. (1973) Methods of ranking SDI and IR outputs. London, IEE (INSPEC), (also as INSPEC Report R73/18; OSTI Report 5184)

EVERITT, B. (1974) Cluster analysis. London, Heinemann for the S.S.R.C..

FAIRTHORNE, R.A. (1964) Basic parameters of retrieval tests. In: American Documentation Institute. Parameters of information science: proceedings of the 27th annual meeting, 1964. Washington, D.C., Spartan. (pp. 343-5)

FARRADANE, J. (1974) The evaluation of information retrieval systems. Journal of documentation, 30, pp. 195-209.

FISHER, R.A. (1936) The use of multiple measurement in taxonomic problems. Annals of eugenics, 7, pp. 179-88

FOLEY, D.H. (1975) and SAMMON, J.W. An optimal set of discriminant vectors. IEEE transactions on computing, C-24, pp. 281-9.

GEBHARDT, F. (1975) A simple probabilistic model for the relevance assessments of documents. Information processing & management, 11, pp. 59-65.

GIULIANO, V.E. (1966) and JONES, P.E. Study and test of a methodology for laboratory evaluation of message retrieval systems. Cambridge, Mass., Arthur De Little, Inc..

GOFFMANN, W. (1964a) On relevance as a measure. Information storage and retrieval, 2, pp. 201-3.

GOFFMAN, W. (1964b) A searching procedure for information retrieval. Information storage and retrieval, 2, pp. 73-8.

GOFFMAN, W. (1966) and NEWELL, V.A. A methodology for test and evaluation of information-retrieval systems. Information storage and retrieval, 3, pp. 19-25.

GOOD, I.J. (1967) The decision-theory approach to the evaluation of information retrieval systems. Information storage and retrieval, 3, pp. 31-4.

GOOD, I.J. (1971) and CARD, W. The diagnostic process with special reference to errors. Methods of information in medicine, 10, pp. 176-88

GREEN, D.M. (1974) and SWETS, J.A. Signal detection theory and psychophysics. Huntingdon, N.Y., Krieger. (reprint with corrections and additions to first edition, 1966)

GREY, D.R. (1972) and MORGAN, B.J.T. Some aspects of ROC curve-fitting: normal and logistic models. Journal of mathematical psychology, 9, 1972, 128-39

GUAZZO, M. (1977) Retrieval performance and information theory. Information processing & management, 13, pp. 155-66.

HARTER, S.P. (1975) A probabilistic approach to automatic keyword indexing: part 1. Journal of the American Society for Information Science, 26, pp. 197-206

HEAPS, H.S. (1978) Information retrieval: computational and theoretical aspects. N.Y., Academic press.

HEINE, M.H. (1973a) The inverse relationship of precision and recall in terms of the Swets model. Journal of documentation, 29, pp. 81-4.

HEINE, M.H. (1973b) Distance between sets as an objective measure of retrieval effectiveness. Information storage and retrieval, 9, pp. 181-98.

HEINE, M.H. (1974) Design equations for retrieval systems based on the Swets model. Journal of the American Society for Information Science, 25, pp. 183-98.

HEINE, M.H. (1975) Measures of language effectiveness and the Swetsian hypotheses. Journal of documentation, 31, pp. 283-7

HEINE, M.H. (1977a) Incorporation of the age of a document into the retrieval process. Information processing & management, 13, pp. 35-47.

HEINE, M.H. (1977b) The 'question' as a fundamental variable in information science. In: Theory and application of information research - proceedings of the Second International Research Forum on Information Science, Copenhagen, 1977. London, Mansell, 1980.(pp. 137-145)

HELSTROM, C.W. (1960) Statistical Theory of signal detection. Pergamon. (second ed. 1968)

HILLMAN, D.J. (1964) Two models for retrieval system design. American documentation, 15, pp. 217-25.

HOEL, P.G. (1971) Introduction to mathematical statistics. 4th ed. N.Y., Wiley. (pp. 181-6)

HUTCHINSON, T.P. (1978) An extension of the signal detection model of information retrieval. Journal of documentation, 34, pp. 51-4.

IKER, H.P. (1967) Solution of Boolean equations through the use of term weights to base two. American documentation, 18, p.47

JARDINE, N. (1971a) and VAN RIJSBERGEN, C.J. The use of hierarchic clustering in information retrieval. Information storage and retrieval, 7, pp. 217-40.

JARDINE, N. (1971b) and SIBSON, R. Mathematical taxonomy. London, Wiley.

KAYE, D. (1973) A weighted rank correlation coefficient for the comparison of relevance judgements. Journal of documentation, 29, pp. 380-9.

KEEN, E.M. (1971) Evaluation parameters. In: SALTON, G. (ed.) The SMART retrieval system. Englewood Cliffs, Prentice Hall.

KEEN, E.M. (1972) and DIGGER, J.A. Report of an information science index languages test. Aberystwyth, College of Librarianship Wales.

KEEN, E.M. (1973) The Aberystwyth index languages test. Journal of documentation, 29, pp. 1-35

KING, D.W. (1971) and BRYANT, E.C. The evaluation of information services and products. Washington, D.C., Information Resources Press.

KOCHEN, M. (ed.) (1967a) The growth of knowledge: readings in organization and retrieval of information. N.Y., Wiley.

KOCHEN, M. (1967b) Adaptive mechanisms in digital "concept" processing. In: KOCHEN, M. (ed.) The growth of knowledge: readings... N.Y., Wiley. (pp. 185-203)

KOCHEN, M. (1974a) Principles of information retrieval. Los Angeles, Melville.

KOCHEN, M. (1974b) and BADRE, A.N. Questions and shifts of representation in problem-solving. American journal of psychology, 87, pp. 369-83.

KORFHAGE, R.R. (1966) Logic and Algorithms. N.Y., Wiley.

LANCASTER, F.W. (1968) and CLIMENSON, W.D. Evaluating the economic efficiency of a document retrieval system. Journal of documentation, 24, pp. 16-40.

LANCE, G.N. (1967) and WILLIAMS, W.T. A general theory of classificatory sorting strategies. Computer journal, 9, pp. 373-80; 10, pp. 271-7.

LANDRY, B.L. (1971) A Theory of indexing: indexing theory as a model for information storage and retrieval. [PhD thesis] Computer and Information Science Research Center, Ohio State University.

LINE, M.B. (1973) and SANDISON, A. 'Obsolescence' and changes in the use of literature with time. Journal of documentation, 30, pp. 283-350.

LUDWIG, B.M. (1975) and GLOCKMAN, H.P. The formal analysis of document retrieval systems. Journal of the American Society for Information Science, 26, pp. 51-5

LUHN, H.P. (1957) A statistical approach to mechanised encoding and searching of literary information. IBM journal of research & development, 1, pp. 309-17.

MARCZEWSKI, E. (1958) and STEINHAUS, H. On a certain distance of sets and the corresponding distance of functions. Colloquium mathematica (Warsaw), 6, pp. 319-27

MARON, M.E. (1960) and KUHNS, J.L. On relevance, probabilitic indexing and information retrieval. Journal of the Association for Computing Machinery, 7, pp. 216-44.

MATHAI, A.M. (1975) and RATHIE, P.N. Basic concepts in information theory statistics: axiomatic foundations and applications. New Delhi, Wiley Eastern.

MATTHEWS, F.W. (1967) and THOMSON, L. Weighted term search. Journal of chemical documentation, 7, pp. 49-56.

MATTHEWS, F.W. (1970) Weighted term search. Proceedings of the ASIS annual meeting, 33rd, 1970. 7, pp. 315-7.

MEADOWS, A.J. (1974) Communication in science. London, Butterworths.

MILLER, W.L. (1971) A probabilistic search strategy for MEDLARS. Journal of documentation, 27, pp. 254-66.

MOSS, R. (1973) [Letter]. Journal of documentation, 29, pp.109-13.

MOTT, T.H. Jr. (1972), ARTANDI, S. and STRUMINGER, L. Introduction to PL/1 programming for library and information science. N.Y., Academic.

MUNSON, W.A. (1954) and KARLIN, J.E. The measurement of human channel transmission characteristics. Journal of the Acoustical Society of America, 26, pp. 542-53.

NEYMAN, J. (1933) and PEARSON, E.S. On the problem of the most efficient tests of statistical hypotheses. Philosophical transactions of the Royal Society of London, A231, pp. 289-

NILSSON, N.J. (1965) Learning machines - foundations of trainable pattern classifying systems. N.Y., McGraw-Hill.

ODDY, R.N. (1974) Reference retrieval based on user induced dynamic clustering. PhD thesis Computing Lab., University of Newcastle upon Tyne.

ODDY, R.N. (1977) Information retrieval through man-machine dialogue. Journal of documentation, 33, pp. 1-14.

PAICE, C.D. (1977) Information retrieval and the computer. London, Macdonald and Jane's.

QUINE, W.V.O. (1959) On cores and prime implicants of truth functions. American mathematical monthly, 66, pp. 755-60.

RADECKI, T. (1976a) New approach to the problem of information system evaluation. Information processing & management, 12, pp. 319-26.

RADECKI, T. (1976b) Mathematical model of information retrieval system based on the concept of fuzzy thesaurns. Information processing & management, 12, pp. 313-18.

RADECKI, T. (1977) Mathematical model of time-effective information retrieval system based on the theory of fuzzy sets. Information processing & management, 13, pp. 109-16

REES, A.M. (1963) and SARACEVIC, T. Conceptual analysis of questions in information retrieval systems. In: Proceedings of the American Documentation Institute, part 2. Washington, D.C. (pp. 175-77)

REES, A.M. (1967a) and SCHULTZ, D.G. A field experimental approach to the study of relevance assessments in relation to document searching. Case Western Reserve University, School of Library Science. 2 vols.

REES, A.M. (1967b) Evaluation of information systems and services. Annual review of information science and technology, 2, pp.63-86.

REISIG, G.H.R. (1972) Optimization of economical utilization of scientific information. Kybernetes, 1, pp. 199-205.

ROBERTSON, S.E. (1969) Parametric description of retrieval tests. (2 pts) Journal of documentation, 25, pp. 1-27, 93-107

ROBERTSON, S.E. (1972) Term specificity [letter]. Journal of documentation, 28, pp. 164-5.

ROBERTSON, S.E. (1974) Specificity and weighted retrieval. Journal of documentation, 30, pp. 41-6.

ROBERTSON, S.E. (1975) A theoretical model of the retrieval characteristics of information. [PhD thesis]. School of Library, Archive & Information Studies, University College, London.

ROBERTSON, S.E. (1976) and SPARCK JONES, K. Relevance weighting search terms. Journal of the American Society for Information Science, 27, pp. 129-46.

ROBERTSON, S.E. (1977a) Theories and models in information retrieval. Journal of documentation, 33, pp. 126-48.

ROBERTSON, S.E. (1977b) The probability ranking principle in IR. Journal of documentation, 33, pp. 294-304.

ROCCHIO, J.J. (1965) and SALTON, G. Information search optimization and iterative retrieval techniques. AFIPS Fall Joint Computer Conference: proceedings, 27(1), pp. 293-305.

ROCCHIO, J.J. (1966) Document retrieval systems - optimization and evaluation. [PhD thesis] Harvard Computation Lab., Harvard University.

SAGER, W.K. (1976) and LOCKEMANN, P.C. Classification of ranking algorithms. International forum on information and documentation (FID), 1, pp. 12-25.

SALTON, G. (1966) et al. Information storage and retrieval. Cornell University, Dept. of Computer Science. (also as the Dept's Scientific report, ISR-11)

SALTON, G. (1968) Automatic information organisation and retrieval. N.Y., McGraw Hill.

SALTON, G. (1971a) The SMART retrieval system - experiments in automatic document processing, Englewood Cliffs, Prentice Hall.

SALTON, G. (1971b) The performance of interactive information retrieval. Information processing letters, 1, pp. 35-41.

SALTON, G. (1972) The "Generality" effect and the retrieval evaluation for large collections. Journal of the American Society for information science, 23, pp. 11-22.

SALTON, G. (1973) and YANG, C.S. On the specification of term values in information retrieval. Journal of documentation, 29, 351-72.

SALTON, G. (1975a) Dynamic information and library processing. Englewood Cliffs, N.J., Prentice Hall.

SALTON, G. (1975b), WONG, A. and YANG, C.S. A vector space model for automatic indexing. Communications of the Association for Computing Machinery, 18, pp. 613-20.

SALTON, G. (1975c), YANG, C.S. and YU, C.T. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 26, pp. 33-44.

SARACEVIC, T. (1966) Comparative effects of titles, abstracts and full texts on relevance judgements. Proceedings of the ASIS annual meeting, 32d, 1966. 6, pp. 293-9.

SARACEVIC, T. (ed.) (1970a) Introduction to information science. N.Y., Bowker.

SARACEVIC, T. (1970b) The concept of 'relevance' in information science: a historical review. In: SARACEVIC, T. (ed.) Introduction to information science - N.Y., Bowker (pp. 111-51)

SMITH, M. (1953) and WILSON, E.A. A model of the auditory threshold and its application to the problem of the multiple observer. Psychological monographs, 67(9) = no. 359.

SNEDECOR, G.W. (1967) and COCHRAN, W.G. Statistical methods. 6th ed. Iowa State Univ. Press.

SOMMAR, H.G. (1969) and DENNIS, D.E. A new method of weighted term searching with a highly structured thesaurus. Proceedings of the ASIS annual meeting, 32nd, 1969. 6, pp. 193-8.

SPARCK JONES, K. (1971) Automatic keyword classification for information retrieval. London, Butterworths.

SPARCK JONES, K. (1972) A statistical interpretation of term specificity and its application to retrieval. Journal of documentation, 28, pp. 11-21.

SPARCK JONES, K. (1973) Index term weighting. Information storage and retrieval, 9, pp. 619-33.

SPARCK JONES, K. (1975) A performance yardstick for test collections. Journal of documentation, 31, pp. 266-72.

STAMPER, R. (1973) Information in business and management decisions. London, Batsford.

SWETS, J.A. (1963) Information retrieval systems. Science, 241, pp. 245-50.

SWETS, J.A. (ed.) (1964) Signal detection and recognition by human observers: contemporary readings. N.Y., Wiley.

SWETS, J.A. (1967a) Effectiveness of information retrieval methods. Air Force Cambridge Research Labs., Bedford, Mass.. (also as the Laboratories' Report, no. AFCRL-67-0412; & also published by Bolt, Beranek and Newman, Cambridge, Mass.)

SWETS, J.A. (1967b) Signal detection as a model of information retrieval. In: BRISSON, F. and MONTMOLLIN, M. de La simulation du comportement humain: the simulation of human behavior: actes d'un symposium O.T.A.N., Paris, 1967. Paris, Dunod. 1969. pp. 253-67

SWETS, J.A. (1969) Effectiveness of information retrieval methods. American documentation, 20, pp. 72-89

SWETS, J.A. (1973) The relative operating characteristic in psychology. Science, 182, pp. 990-1000.

TAGUE, J. (1978) and FARRADANE, J. Estimation and reliability of retrieval effectiveness measures. Information processing & management, 14, pp. 1-16

TAHANI, V. (1976) A fuzzy model of document retrieval systems. Information processing & management, 12, pp. 177-87.

TANNER, W.P.Jr. (1954) and SWETS, J.A. A decision-making theory of visual detection. Psychological review, 61, pp. 401-9.

TAYLOR, R.S. (1968) Question-negotiation and information-seeking in libraries. College & research libraries, 29, pp. 178-94.

TURSKI, W.M. (1971) On a model of information retrieval system design based on thesaurus. Information storage and retrieval, 7, pp. 89-94.

UHLMANN, W. (1968) Document specification and search strategy using basic intersections and the probability measure of sets. American documentation, 19, pp. 240-6

VAN RIJSBERGEN, C.J. (1973) and SPARCK JONES, K. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. Journal of documentation, 33, pp. 251-7

VAN RIJSBERGEN, C.J. (1974) Foundation of evaluation. Journal of documentation, 30, pp. 365-73.

VAN RIJSBERGEN, C.J. (1977) A theoretical basis for the use of co-occurrence data in information retrieval. Journal of documentation, 33, pp. 106-19.

VAN RIJSBERGEN, C.J. (1979a) Information retrieval. 2nd ed. London, Butterworths.

VAN RIJSBERGEN, C.J. (1979b) Retrieval effectiveness. Progress in communication sciences, 1, pp. 91-118.

VAN RIJSBERGEN, C.J. (1980). S.E. ROBERTSON and M.F. PORTER. New models in probabilistic information retrieval. Computer laboratory, University of Cambridge.

VERNIMB, C. (1977) Automatic query adjustment in document retrieval. Information processing & management, 13, pp. 339-53.

VICKERY, B.C. (1970) Techniques of information retrieval. London, Butterworths.

VICKERY, B.C. (1973) Information systems. London, Butterworths.

VICTOR, N. (1974), TRAMPISCH, H.J. and ZENTGRAF, R. Diagnostic rules for qualitative variables with interactions. Methods of information in medicine, 13, pp. 184-6

WALD, A. (1950) Statistical decision functions. N.Y., Wiley.

WHITTLE, P. (1970) Probability. Harmondsworth, Penguin.

WILLIAMS, J.H. (1965) Results of classifying documents with multiple discriminant functions. In: Statistical association methods for mechanized documentation. (Ed. M.E. Stevens) National Bureau of Standards. (pp. 217-24)

WORDSWORTH, H.M. (1959) and BOOTH, R.E. [no title given]. Western Reserve University. (also as Technical note no. 7; and AFOSR-TN-418) [as cited by Swets (1963)]

YU, C.T. (1976), LUK, W.S. and CHEUNG, T.Y. A statistical model for relevance feedback in information retrieval. Journal of the Association for Computing Machinery, 23, pp. 273-86.

YU, C.T. (1977) and SALTON, G. Effective information retrieval using term accuracy. Communications of the Association for Computing Machinery, 20, pp. 135-142.

ZUNDE, P. (1967) and SLAMECKA, V. Distribution of indexing terms for maximum efficiency. American documentation, 18, pp. 104-8.

ZUNDE, P. (1971) Structural models of complex information sources. Information storage and retrieval, 7, pp. 1-18.