Information retrieval from classical databases from a signal-detection standpoint

A review

M H Heine

The retrieval of information from classical (object/attribute) databases is discussed in the light of signal-detection theory. The approach is based on the Swetsian schema, although it is expressed in a more general form.

Keywords: information retrieval, databases, signal detection

INTRODUCTION

A conventional view of information retrieval (IR) is this: an enquirer (system user) expresses an 'information need' in some symbolic form, the 'query'. This could for example be as a Boolean expression containing index terms. The query (or 'search statement') is then input to an IR system, either by the enquirer or a proxy. The system, now almost always a computing system, has as its central features

- a database of records (stored as electrical images on hardware),
- a dynamic electronic process (software) accessing the database — possibly through inverted file representations of it.

Records in the database that 'match' the query (evaluate a Boolean expression to True), i.e., that are 'retrieved', are then transmitted to the enquirer, for example as images on a VDU, or at least are made available for such transmission. The (zero or more) retrieved records will, in general, not all be 'relevant', i.e., not all will be regarded by the enquirer as a proper response to the information need that initiated the whole process. However, if the query is a 'good' one, many or even most of the retrieved records will be relevant.

The above, informal picture can be elaborated on in manv ways, as discussed in detail by, for example, Heaps[1], Salton[2], Salton and McGill[3], and van Rijsbergen[4]. Accordingly, we limit our general comments to those closely relevant to this review. First, the *conceptualization* embodied in such a description needs careful attention. For example, one cannot, at the same time, refer to the enquirer's input to the search system as a 'query' and also use this term to connote a criterion by which relevance is to be judged. The concepts are quite different. Another example is the need to distinguish between the ranking of all records in the database (if only in principle) *prior* to retrieval from it, i.e., ranking as a means of identifying records to be retrieved, and ranking of the set of retrieved records *subsequent* to their definition as 'retrieved'. The ranking algorithms may or may not be the same. Secondly, it is clear that *fallibility* is a general concomitant of IR. If this were not so, IR would reduce to 'data retrieval'. The *scientific* problems of IR (as distinct from the problems of implementing IR in a computing environment) would not exist. A conceptual apparatus to handle this fallibility is clearly necessary, and we can imagine it as being not only central to any formal schema of IR, but as conferring on IR theory an immediate pragmatic character.* Thirdly, as is clear from various writings, although the historical roots of IR are in the retrieval of written records represented in 'bibliographical databases', any formal schema of IR should be worded so that the notion of a record as 'that which can inform' is not unduly restricted by this.

*i.e., a character determined by its being 'driven' by the usefulness of the process involved, following Durkheim

School of Librarianship & Information Studies, Newcastle upon Tyne Polytechnic, Newcastle upon Tyne, UK

Received 8 March 1984

0144–817X/84/020095–18$03.00

In this paper, we attempt to review the topic of IR from classical (object/attribute) databases from the perspective of *signal-detection theory*. The approach stems from the work of Swets[5–8]. However, it is expressed in a rather more general form that recovers Swets's schema as a special case, as discussed in selected ways by the author[9–11]. By way of clarifying the intended scope of this review the following points are noted:

- The lack of *philosophical* coherency in IR at the present time prohibits any attempt at rigour in comparing different formal schema. One thinks of such issues as: Need the *criteria by which relevance is judged* be observable? Does relevance exist *a priori*, i.e., before the enquiry? If it does, is *it*, as well as the criteria to which it appeals, required to be observable? By whom? Alternatively, should we see relevance as definable only as an *a posteriori* entity? Do IR systems actually *transmit* information (as is usually assumed, notwithstanding the difficulty of conceptualizing information), or is 'information' a *construction* created by the enquirer in response to data retrieved? And so on. The dissonance within the area of study in regard to these fundamental questions (or, if not dissonance, lack of interest in them relative to interest in problems of implementation or solely mathematical schemas) means that any broad review of IR is forced to adopt a philosophical position that may be the personal one of the reviewer.
- Where published contributions have introduced a probabilistic approach, and have not *related* the work concerned to that of Swets, the work concerned may nevertheless be so *relateable.* No claim is made in this review that all probabilistic work on IR that is relateable to Swets's work is described in it, but it is suggested that much of it can be. The demonstration of this is, however, seen as a proper burden of original authors, in view of the definitional disparities between, and occasional vagueness of, much IR writing. (For one writer 'probability of relevance' may signify 'probability that a record in the database is relevant'. To another it may mean 'probability that a record *defined by a certain system response* is relevant'. To yet another author it may denote the 'probability that a record is retrieved'. The term 'query' is also used ambiguously: search expression, or criterion of relevance? Set of keywords, or set of keywords with weights attached? Context may or may not always make clear.)

It is hoped that some interest in the Swetsian schema will be renewed by this review, especially in the area of IR experiment, and especially in concert with theoretical interest in IR combinatorics.

CONCEPTUALIZATIONS

As a basis for discussion, the following concepts are introduced:

Informational objects

We follow the convention of a set of informational objects, $s \varepsilon S$. Each such object is a structure of *signs* (which are a primitive). Such objects *can* inform an intelligent being through the latter placing constructions* on them, but whether an object *will* so inform will depend on the *effect(s)* of its data on the intelligent beings accessing it.

Example 1: A full-text record of a document may allow a person to place constructions on it, and so inform him or her.

Example 2: A record in the form of a bibliographical description of a document may inform in a 'derivative' sense: it may suggest that the document so described is as in Example 1.

Relevance

A purposeful (self-directed) user of an IR system will appeal to (access, reference) one or several criteria by which the informing constructions referred to above are *limited.* Such criteria may be personal, i.e. inaccessible to others, or they may be described to that person in some public (i.e., other-accessible) symbolic language (e.g., spoken or written natural language). We might usefully distinguish too between informing when the personal criteria of same are given *before* the object is accessed ('imposed informing') and when this is not so ('creative informing'). We shall label objects that inform as 'relevant'. When the labelling is done in respect of imposed informing, the relevance is an 'imposed relevance'; when the labelling is done for creative informing, the relevance is 'creative relevance'. Our interest in this review will primarily be in imposed relevance.

Example 3: An astrophysicist browses through a set of documents. A document on the heat balance of polar ice-caps of the Earth is interpreted as relating to the problem of describing the Solar photosphere. The object has a creative relevance for that person.

Example 4: In an IR experiment, arbiters of relevance are given written statements of information needs and are asked to flag each record in a database as 'relevant' or not. Here, the informing is imposed informing and the criteria of relevance are (arguably) public ones.

Example 5: A student carrying out a literature search at the commencement of a research study for a degree is presumably working in mixed-mode, i.e., recognizing provisional public criteria for his or her judgements.

For illuminating but inconclusive discussion in this area we refer selectively to Bookstein[14], Buckland[15], le Claire[16], Derr[17], Krikelas[18] and especially to Saracevic[19, 20], but with the suggestion that not all of the discussion is adequately formal, or sufficiently concerned with operational definition. (Even where discussion of 'relevance' is formal in a *mathematical* sense, the formality can be pseudo-rigorous from a *scientific*

*Brookes has emphasized the interpretative role of the person who is informed: 'The communication process consists solely of public displays and private interpretations of those displays'[12]; and Heather has remarked: 'Extraction of knowledge . . . is more than retrieving knowledge it is rather the generation of it'[13] viewpoint in that denotation is seen as sufficient for real-world construction, i.e., accompanying operational definition is not offered.) The commitment of some writers (e.g., Cuadra and Katter[21], Hutchinson[22]) to the notion of 'degrees of relevance' is also arguable. Possibly the concept of *types* of relevance would be more realistic. (See, for example, Buckland's distinctions between 'responsiveness', 'pertinence', and 'beneficiality'.) We refer again to Saracevic for full critical comment.

Attributes

Each informational object is directly accessible by one or more sign-sequences (keywords, terms, descriptors) which can simply be called 'attributes', $a \in A$. Here A is the set of all admissible attributes, i.e., the thesaurus, or indexing vocabulary. The set of attributes associated with (i.e., indexing) a specific informational object s_r will be denoted A_{s_r}. A numeric value can be an attribute.*

Database

A database is a binary relation $\mathscr{R}$ in the Cartesian product $S \times A$, i.e., $\mathscr{R} \subset S \times A$. A *record* in a database thus comprises an informational object and its associated set of attributes, i.e., the ordered pair (s_r, A_{s_r}).

Query

The term 'query' (and its synonyms: question, request, search statement, search expression) is subject to some variability in usage[24]. The term has been used in the past ambiguously to denote either

- a verbal criterion (putatively complete) for relevance judgements, or
- a verbal entity conveying a (necessarily incomplete) information need having its existence in some real-world, non-verbal setting.

(A vital distinction between these concepts is demonstrated by noting that in the first case the entity is *fixed*, for a given information need, wheres in the second case the entity is a *variable* — Saracevic.) The now classic Cranfield experiments[25, 26] recognized the former meaning, as did those at Aberystwyth[27, 28]. The absence in both experiments of definitions of real-world settings appears to have created ambiguity in the relevance assessments, and thus some uncertainty in the *meaning* to be attached to their experimental results[29]. (For a recent full review of the Cranfield and other experiments see Sparck Jones[30, 31].)

Belkin[32] has termed the *origin* of queries (apparently in the first sense of those given above) an 'anomalous state of knowledge'. However, to the author's knowledge this concept has not yet been operationalized.

Last, we note that some variability in the concept of query, provided it is clearly denoted, can usefully be recognized, for the second sense of query given above. Four types of query are here distinguished:

I. *Queries in narrative form.* Here the information need is seen as conveyed by a statement in natural language. (Salton has referred to these as 'initial queries'[2].) An example of a query of this type is:

THE POISONING OF CATTLE BY NITRITE IONS AND RELATED SUBJECTS

Until now such queries have been unconventional as inputs to IR software, but increasingly so-called 'intelligent software' is accepting them. (See for example Minker[33] or Doszkocs and Rapp[34].) We term such queries 'narrative-form queries'. Some categories of narrative-form query have been described by Guiliano and Jones, ranging from the phase to the paragraph.

II. *Queries as sets of attributes.* Here the information need is characterized by a set of attributes $\{a_i;\ i \varepsilon I\} \subset A$. This is equivalent to a vector $(q_1, q_2, \ldots, q_{|A|})$, where $q_i \varepsilon \{0, 1\}$ denotes presence or absence of a_i in the query. An example of a query of this type is:

{NITRITE, POISONING, CATTLE, CONCENTRATION ≥ 3.4}

We note that when the attribute is a *numeric variable* (such as 'CONCENTRATION' in the above example), each informational object will have a numeric *value* attached to it, where applicable, such as '4.2'. We assume that in such cases the set of attributes also includes a parameter and equality or inequality sign, or set of same. We refer to queries of this type as 'set-form queries'. A set-form query may derive from a narrative-form query.

III. *Queries as logical expressions.* Here, the information need is represented as a logical expression, the member elementary logical variables (ELVs) of which derive from a set form query. Examples of ELVs are:

NITRITE where **NITRITE:** = True if
NITRITE εA_{s_i} else False,

CONCENTRATION ≥ 3.4 where
CONCENTRATION ≥ 3.4: = True if
(for s_i CONCENTRATION ≥ 3.4) else False.

Examples of queries of this type *for a fixed instance of information need* are:

NITRITE ∧ (CATTLE ∨ POISONING)
(NITRITE ∨ CATTLE) ∧ POISONING
NITRITE ∧ CATTLE ∧ POISONING
NITRITE ∧ CATTLE ∧ POISONING ∧ CONCENTRATION ≥ 3.4.

In this construction, a query is usually referred to nowadays as a 'search statement', but this usage may be expected to weaken as more sophisticated IR systems are implemented. We shall refer to queries of this type as 'logical-form queries'.

IV. *Queries as sequences of logical expressions.* A form of query that is, in a sense, intermediate between the set-form query and the logical-form query will also prove useful. This is that of a *sequence* of logical expressions. One may imagine the members of such a sequence being input successively to an IR system. We shall refer to queries of this type as 'sequential-form queries'.

It is emphasized that in each of the above types of query, the individuation of the query is *variable* for a *fixed* instance of information need, i.e., for occasions when what we have termed 'imposed informing' obtains. (When the informing of the database is 'creative' then the query is *a fortiori* variable.) The lay phrase 'relevance to a

*This is evident in the everyday use of databanks. For a general theoretical framework we refer to Lipski and Marek[23] question (query)' has obscured this essential property of queries.

Sets of relevant objects

Although most IR workers readily accept that it makes sense to flag retrieved objects as relevant or not-relevant, the jump from this to the view that it makes sense to talk of a set of relevant objects *in the database* is not so easily taken. To refer to such a set is, it can be seen, to make the assumption that a database is *transferring information* rather than *allowing information to be created.* A strict adherence to the latter view entails rejection of the former view (since meaningless). Such a view might be seen as supported by the impracticality, in an experimental situation, of asking arbiters of relevance to scan, say, 2×10^6 records for 'relevance'. This issue is not faced by solely mathematical accounts but a stand on it is needed, especially following the sceptical review by Cooper[35]. The position taken in this review is that a subset of a database *can* usefully be identified as 'capable of informing' (or 'potentially informing') even if not all of the subset can be retrieved by a specific logical-form query. The justification for this is simply the thought-experiment that *in principle* an arbiter of relevance can explore all of the database to identify the subset, and that reasonable experimental proxies for this procedure can be identified. However, as this view seems to be supportable only when relevance-criteria are specified in advance of the query being formulated (whether the criteria are public or private), interest henceforth is restricted to what has been termed 'imposed informing'. A set of relevant objects, each referencing a given information need, will be denoted by $X \subseteq S$.

Sets of retrieved objects

Little conceptual difficulty arises here, since the set of retrieved objects is defined algorithmically. A set of retrieved objects is simply the set of all objects in S that evaluate a given logical-form query to True. This set is denoted by $Y \subseteq S$.

Retrieval effectiveness

The usefulness of the information imparted to an enquirer by an IR system, or more-intuitively the 'quality of response' of an IR system, can be conceptualized in various ways. Interest in this problem is restricted here to the portrayal of the *overlap of sets X and Y.* This assumption is hardly free from criticism but it is made

- in view of the indirect, empirical support for it given by Cleverdon and Kidd[36], and
- in order to allow *some* constructive discussion of the problem to develop.

For further, detailed discussion of the general concept of retrieval effectiveness, we refer selectively to King and Bryant[37], Lancaster and Climenson[38], Salton[2] (Chapter 6), Vickery[39], and van Rijsbergen[40]. Papers offering novel theoretical departures include those of Cooper[41], Gebhardt[42], Guazzo[43], Ludwig and Glockmann[44] and Radecki[45]. It is suggested that consensus is likely to arise in this area, only when formal theory is predicated on agreement as to what it means 'to be informed'.

In the following, the main interest will be in the *variability* in set Y, for a fixed set X, when different logical-form queries are defined (on the basis of a fixed set-form query). Global variability in the set-form query will also be recognized as a key, but neglected variable. To offer concentrated discussion on the relationship between sets X and Y on this basis seems most appropriate.

Lastly, we note two practical points. First, our interest is restricted to instances of informing where there is *one* set X for a specified set of retrieved objects, X. In practice, a group of system users may *share* a logical-form query (group-profile), and in such cases a *set* of (say) R-values will pertain for any set Y[46]. Secondly, the choice of database will influence the membership of sets X and Y. With over 1700 databases now available, accessible through some 250 host systems, these figures increasing at approximately 19 per cent and 15 per cent per annum, respectively*, the *choice* of set S, conditioned on a choice of host, will strongly influence the values of any retrieval effectiveness measure employed.

A note on 'control'

For completeness, it is noted that a complete schema of IR would include conceptualization of the *control structure* of IR. What is it that orientates the relation $\mathscr{R}$ towards the users of the database? What orientates an information percipient towards the (invisible) target set, X, i.e., which gives intention or purpose to this activity? The concept of 'control', although prevalent in the cybernetics literature (if not a defining characteristic of the latter) has largely been missing from the IR literature to date. Yet the increasing interest in machines that do the bidding of humans at fairly sophisticated levels (i.e., 'intelligent' software) seems to require that control should become an explicit concept. Following provisional work of the author it is noted that it may be useful to distinguish between communication processes that are: 'transmitter-driven' (the classical Shannon schema, but with the receiver controlled by an extra-database signal, the 'metasignal', from the transmitter); 'percipient-driven' (transmission in this case is an undefined or vacuous concept, and information-extraction is controlled totally by the system user); 'search-system driven' (control of perception resides in the search software); or 'intermediary-driven' (the procedure controlling $\mathscr{R}$ also controls the system user[11, 47]. In such schemas, 'intelligence' can conveniently be defined as *incompleteness* in system description, i.e., as an asserted semiotic property, rather than as a semi-mystical entity. (Relateable discussion is offered by Lipski[48].)

The concepts we have discussed above will form the basis of a formal schema to be described in the next section. The interest will be in an integrated schema that incorporates all of

- concepts of IR effectiveness,
- concepts of object, attribute and database, and
- concepts of set-form, sequential-form and logical-form query.

*Data are taken from *Directory of online databases* (Cuadra Associates; 4th ed. + supplement, 1983)

The motivations for the schema are

- improved understanding of the IR process, especially through the identification of necessary truths in the schema,
- improvement in the design of software supporting IR, especially in regard to the identification of appropriate program modules, and module-interactions, and
- improvement to the design of IR experiments, again so as to improve software function but additionally to support system-user education policies.

The essential kernel of the schema is due to Swets, but with Swets's original schema extended so as to recognize a discrete, rather than continuous outcome space, and with logical-form queries explicitly incorporated in it.

SIGNAL-DETECTION THEORY OF IR WHEN THE OUTCOME SPACE IS DISCRETE (SIGN-DETECTION THEORY)

First we define the following probabilities:

The probability that a relevant object is retrieved, R (Recall):

$$R = \frac{|X \cap Y|}{|X|} \qquad \text{(definition)}$$

The probability that a non-relevant object is retrieved, F (Fallout):

$$F = \frac{|(S \backslash X) \cap Y|}{|S \backslash X|} \qquad \text{(definition)}$$

The probability that a retrieved object is relevant, P (Precision):

$$P = \frac{|X \cap Y|}{|Y|} \qquad \text{(definition)}$$

The probability that a non-retrieved object is relevant, M (Miss-fraction):*

$$M = \frac{|(S \backslash Y) \cap X|}{|S \backslash Y|} \qquad \text{(definition)}$$

The probability that an object in the database is relevant, G (Generality)†

$$G = \frac{|X|}{|S|} \qquad \text{(definition)}$$

The probability that an object in the database is retrieved, C (retrievality)‡:

$$C = \frac{|Y|}{|S|} \qquad \text{(definition)}$$

A range of other probabilities have been defined in connection with the problem of characterizing retrieval effectiveness, and we refer for further comment on them to Cooper[50], Farradene[51], Rees[52], Robertson[49] and Swets[5]. One interesting development has been the use of effectiveness definitions that appeal to the 'symmetric difference set', $X \triangle Y$. Work by Vickery[53], Jardine and van Rijsbergen[54] and Heine[46] in this connection has been integrated and clarified by van Rijsbergen[55], via a general univariate function that maps the ordered pair (R, P) to $[0, 1]$. The form of this function is $1 - (\alpha P^{-1} + (1 - \alpha)R^{-1})^{-1}$, where $\alpha \varepsilon [0, 1]$, the value $\alpha = ½$, for example, generating a class of measures including the metric

$$\begin{aligned} D &= \frac{|A \triangle B|}{|A \cup B|} = \frac{R + P - 2RP}{R + P - RP} = \\ &= \frac{F(1 - G) + G(1 - R)}{F(1 - G) + G} = \frac{C + G - 2PC}{C + G - PC} \\ &= \frac{C + G - 2RG}{C + G - RG}. \end{aligned}$$

— i.e., the Marczewski–Steinhaus metric. The latter has recently been further investigated by Bollmann and Cherniavsky[56].

The probabilities defined relate to one specific instance of information need (i.e., one set X), and one set of retrieved objects (i.e., one set Y). The arithmetic relationships between them, such as

$$\frac{R}{P} = \frac{C}{G}; \text{ or } P = \frac{GR}{F(1 - G) + RG} = \frac{CPR}{F(R - CP) + CPR}$$

$$\text{and } R = \frac{CP}{M(1 - C) + PC}$$

will not necessarily obtain for *statistics* of R-values, P-values, etc., of course.

On the basis of the preceding constructs, the problem we address is: how can we characterize, using R, F, P etc., the effectiveness of *all possible logical-form queries* defined by a particular combination of $\mathscr{R}$, X and Q_j? The term 'communication envelope' will be chosen to denote this combination, i.e., the triple $<\mathscr{R}, X, Q_j>$.* The term 'envelope' is used since the characterization being sought 'contains' the effects of all triples $<\mathscr{R}, X, \mathbf{L_{jk}}>$, where $\mathbf{L_{jk}}$ is a logical-form query derivable from Q_j. That is, if Q_j is the set $\{a_i;\ i \in I\} \subset A$ then $\mathbf{L_{jk}}$ is a logical expression the admissible ELVs of which are:

$$\{\mathbf{a}_i^{\mathbf{k}_i};\ k_i \in \{0, 1\},\ i \in I\}.$$

The ELVs are assigned values for each informational object $s_r \in S$ according to the rule:

$\mathbf{a}_\mathbf{i}^\mathbf{1}$: = True if $\mathbf{a}_\mathbf{i} \in A_{s_r}$ else False; $\mathbf{a}_\mathbf{i}^\mathbf{0}$: = Not $\mathbf{a}_\mathbf{i}^\mathbf{1}$.

Example 6: Suppose $Q_j = \{a_{101}, a_{262}, a_{265}\}$. Then $\mathbf{L_{jk}}$ might be, for example, $\mathbf{a}_{\mathbf{101}}^{\mathbf{1}} \wedge (\mathbf{a}_{\mathbf{262}}^{\mathbf{0}} \vee \mathbf{a}_{\mathbf{265}}^{\mathbf{1}})$, or $(\mathbf{a}_{\mathbf{101}}^{\mathbf{0}} \wedge \mathbf{a}_{\mathbf{262}}^{\mathbf{1}}) \vee \mathbf{a}_{\mathbf{265}}^{\mathbf{0}}$. There are in this case 256 distinct forms for $\mathbf{L_{jk}}$, i.e., the system user may implement this number of communication processes.

*This is not a conventional definition or term, as neither is C. The probability was not included in the review by Robertson[49], but is suggested by symmetry as between (R, F) and (P, M)

†In the past, confusingly referred to as 'question generality'

‡Introduced by Heine[46] by symmetry with G

*The subscript to Q is just to emphasize that a set-form query is itself variable for a fixed instance of information need, i.e., it reminds us that 'relevance to a query' is (or can usefully be seen to be) meaningless

As indicated in the above example, $2^{2^{|Q_j|}}$ different instances of IR are defined by a communication envelope. However, not all these instances will necessarily generate distinct value-vectors of the variable-vector $(R, F, P, \ldots)$, and in some cases the value of P will be indeterminate.

To generate $\{\mathbf{L}_{\mathbf{jk}}]$ algorithmically, for a specified set-form query Q_j, we can proceed as follows:

I. Define a set of elementary logical conjuncts (ELCs), $\mathbf{e}_\mathbf{n} \in E$:

$$E = \{ \bigwedge_{i \in I} \mathbf{a}_\mathbf{i}^{\mathbf{k}_\mathbf{i}},\ k_i \in \{0, 1\}\}.$$

There are $2^{|Q_j|}$ ELCs, i.e., $n \in N = \{1, 2, \ldots, 2^{|Q_j|}\}$.

II. Define a (new) combination N_C of the members of N.

III. Define a (new) logical-form query by

$$\mathbf{L}_{\mathbf{jk}} := \bigvee_{n \in N_c} \mathbf{e}_\mathbf{n}. \text{ End}$$

More succinctly:

$$\{\mathbf{L}_{\mathbf{jk}}\} = \{ \bigvee_{n \in N_c} (\wedge\, \mathbf{a}_\mathbf{i}^{\mathbf{k}_\mathbf{i}};\ k_i \in \{0, 1\})_n,$$
$$N_c \in N = \{1, 2, \ldots, 2^{|Q_j|}\}\}$$

We now associate each ELC, $\mathbf{e}_\mathbf{n}$, with two probabilities r_n and f_n. At the same time, for later use, we define three functions M_1, M_2 and M_3 which map objects in X, $S\backslash X$, and S, respectively to E. Then:

$$r_n = \frac{|\{s_r;\ s_r \in X,\ \mathbf{e}_\mathbf{n}\}|}{|X|} = \frac{|\{s_r;\ s_r \in X,\ M_1(s_r) = \mathbf{e}_\mathbf{n}\}|}{|X|}$$

$$f_n = \frac{|\{s_r;\ s_r \in S\backslash X,\ \mathbf{e}_\mathbf{n}\}|}{|S\backslash X|}$$
$$= \frac{|\{s_r;\ s_r \in S\backslash X,\ M_2(s_r) = \mathbf{e}_\mathbf{n}\}|}{|S\backslash X|}$$

where it is understood that a True value for $\mathbf{e}_\mathbf{n}$ is required for s_r. Since, by a standard result in logic, the $\mathbf{e}_\mathbf{n}$ partition any set of objects, we have:

$$\sum_{n=1}^{|N|} r_n = 1 = \sum_{n=1}^{|N|} f_n.$$

These probabilities give us a simple basis for calculating the values of R and F (and hence the values of P, D etc.) for any given logical-form query. All we need do is *sum* the r_n (f_n) values that attach to the $\mathbf{e}_\mathbf{n}$ of which $\mathbf{L}_{\mathbf{jk}}$ is a disjunction. That any $\mathbf{L}_{\mathbf{jk}}$ *can* be expressed as disjunction of ELCs is again a standard result (see, e.g., Hohn,[57] p. 46).

Example 7: Suppose $\mathbf{L}_{\mathbf{jk}}$ is expressible as $\mathbf{e}_4 \vee \mathbf{e}_5 \vee \mathbf{e}_{12}$. (Here, $n = 12$ implies $|Q_j| > 2$.) Then $R = r_4 + r_5 + r_{12}$, and $F = f_1 + f_5 + f_{12}$. P is deducible from the values of R, F and G.

More formally, for $\mathbf{L}_{\mathbf{jk}}$ defined by combinations $N_c \in N$, we have:

$$R = \sum_{n \in N_c} r_n, \text{ and } F = \sum_{n \in N_c} f_n.$$

The probabilities r_n and f_n might usefully be thought of as 'micro-probabilities' in that they define components of the 'macro-probabilities' R and F.

Recalling our concern to describe the net effect of $\{\mathbf{L}_{\mathbf{jk}}\}$ for a given relation in $A \times S$, the probability distributions induced by R, F, P etc. are now introduced. Since the values of these variables are real numbers (which happen, in fact, to be in the interval [0, 1], since they are probabilities), they define random variables when $\mathbf{L}_{\mathbf{jk}}$ explores all its possible forms. In other words, when $\mathbf{L}_{\mathbf{jk}}$ is allowed to vary, constrained by Q_j, the variables R, F, P etc. each map $\{\mathbf{L}_{\mathbf{jk}}\}$ to [0, 1]. Induced distributions of the $\mathbf{L}_{\mathbf{jk}}$ in this interval are thus defined. Similarly the random vectors (R, F) and (R, P) define bivariate distributions of $\{\mathbf{L}_{\mathbf{jk}}\}$ in $[0, 1] \times [0, 1]$. We term these latter distributions the '*RF* probability surface' and '*RP* probability surface'. They and the univariate distributions of R, F, P etc. offer characterizations of the communication envelope. More succinct characterizations are the mean and variance of R (etc.), or the centre-of-mass of the *RP* probability surface. Intuitively, such distributions characterize the variability latent in an IR process when the *semantics* of a query is fixed (i.e., a set-form query is defined), but the (Boolean) *syntax* of the query is not fixed.*

The above schema introduces what could be termed the 'sign-detection schema'. To complete it, we note that:

- The actual (human) syntactic *mechanism* is not explicitly included in the schema, just as in classical signal-detection theory the mechanism that determines the response of a receiver is not explicitly known. Instead, the distributions of R (etc.) describe the effects of (i.e., range of action of) this mechanism.
- A syntactic mechanism is *necessary* for communication within the communication envelope to take place, and clearly there is scope for a *technological* entity (a software procedure) to act as a proxy for the human user in selecting 'good' logical-form queries. We pursue this in the next section. However, it can be noted in passing that such an entity is a *procedural* analogue of the *real-valued* (i.e., numeric) entity that appears in classical signal-detection theory, namely the 'threshold parameter'.

Selection of logical-form queries by means of *a priori* procedures

The set $\{\mathbf{L}_{\mathbf{jk}}\}$ is clearly too large, for $|Q_j| > 2$ say, for the IR system user to adopt a trial-and-error approach to identifying useful members of it. In the following, we draw attention to the ways that can be used at present to limit the exploration of $\{\mathbf{L}_{\mathbf{jk}}\}$.

Weighting expressions as a means of selecting logical-form queries

One way to limit the logical-form queries is that in which the members of E are *ordered*, and preference is given to disjunctions of early-ranking $\mathbf{e}_\mathbf{n}$. We shall discuss this method in detail, both because it is related to 'document weighting algorithms' (and in view of the now vast

That is, they show the effect of an IR system choosing a logical-form query 'at random', when the attributes of the query are determined literature on this topic), and because it lays a basis for reconciling sign-detection theory with classical signal-detection theory.

The $\mathbf{e_n}$ can be ordered weakly or strongly. To order them strongly is to define a *permutation* of the $\mathbf{e_n}$, to order them weakly a *composition of a permutation.* A 'weighting function' is a function of the $\mathbf{e_n}$ into N (strong ordering) or into $\{1, 2, \ldots, k\}$, $k < |N|$ (weak ordering). It is defined by a 'weighting expression' which attaches (in general) real-number values to the $\mathbf{e_n}$. The actual numeric weights defined by a weighting expression have no significance other than as means of ordering the $\mathbf{e_n}$. (So that to define the 'mean weight', say, is pointless.) Since the literature on weighting in an IR context is concerned with *object* weighting, rather than the weighting of set-form query ELCs, we now relate this type of weighting to the weighting just described. The reviews of object-weighting by Evans[59, 60], Sager and Lockemann[61], and Noreault *et al.*[62] are noted, as is also the classic paper by Angione[63] which first asserted on a *general* basis the equivalence of information retrieval by weighting-means and logic-means. (But see also earlier work by Brandhorst[64] and Iker[65].)

Define object-weighting functions W_1 (with domain X), W_2 (with domain $S\backslash X$), and W_3 (with domain S), mapping to the real line ($z \in Re$). Each such function is a composition-function of

- a function mapping objects in its domain to E, i.e., the function M_1, M_2 or M_3, and
- a function, W_L, mapping E to Re, i.e., the $\mathbf{e_n}$-weighting function described above.

Thus

for *relevant* objects: $W_1(X) = (W_L \circ M_1)(X) \rightarrow Re$

for non-relevant objects: $W_2(S\backslash X) = (W_L \circ M_2)(S\backslash X) \rightarrow Re$

for *all* objects: $W_3(S) = (W_L \circ M_3)(S) \rightarrow Re$

Thus an object weighting function determines both an order on the objects in its domain (i.e., an order on the $s_r \in \ldots$), and an order on the ELCs in E. It does so via $W_i^{-1}(z)$ and $W_L^{-1}(z)$, respectively. A sequence of z-values ('weights') is thus equivalent to a sequence of ELCs when $W_L^{-1}(z)$ determines a strong order in E. If $W_L^{-1}(z)$ determines a weak order in E, then each z-value in $W_i(\ldots)$ will be equivalent to a subset of E, the members of which disjoin.† (If wished, the subscripts to $\mathbf{e_n}$, hitherto arbitrary, can be chosen so as to reflect those ordering.) It is concluded that, at least for univariate weighting expressions, weighted-object retrieval is not a means of retrieval *alternative* to retrieval using logical-form queries but a means exactly equivalent to it. (This result was first asserted, though not proved, by Angione, who drew attention to the related literature of 'threshold logic'.) For an object to qualify for a weight-value, some logical expression must evaluate to True. There is however some residual uncertainty in regard to whether *multivariate* weighting can be related to Boolean searching in a simple way. The recent introduction of 'weighted Boolean searching'[66, 67] has also cast some doubt on Angione's general assertion, and it seems fair to say the matter is not yet resolved[68, 69].

Before proceeding to discuss the probability distributions induced by the functions W_i on the real line, it is noted that some ELC-weighting functions have a richer domain than E. This determines an extension of the schema which we only touch on here. Suppose we define the (very large) domain of ELCs:

$$H = \{\mathbf{b_m}\} = \{\wedge\ \mathbf{a}_\mathbf{i}^{\mathbf{k_i}};\ \text{where } a_i \in A_{s_r}\backslash Q_j,\ s_r \in S\}$$

Then a modified ELC-weighting function, with domain $E \times H$, can be defined. We shall term such ELC-weighting functions, and the former type (with domain E) Class I weighting functions. Examples of both types are as follows. In each case, $j = 1$, 2 or 3.

Class I weighting functions

(a) Given $\mathbf{e_n}$ evaluates to True: $W_j(s_r) = \sum_{i \in I} k_i$. This is the Cranfield Experiment's 'coordination level' weight.

(b) Given $\mathbf{e_n}$ evaluates to True: $W_j(s_r) = -\sum_{\substack{i \in I \\ k_1 = 1}} d_i \log_b(d_i)$. This is the expression advanced by Miller[70] and Sparck Jones[71], Miller's in fact being more complicated. Here d_i is the 'specificity' of the attribute a_i, defined as $|\{s_r;\ a_i \in A_{s_r}\}| \,/\, |S|$, and b is arbitrary.

Class II weighting functions

(a′) Given $\mathbf{e_n}$ and $\mathbf{b_m}$ each evaluate to True (i.e., $\mathbf{e_n} \wedge \mathbf{b_m}$ evaluates to True):

$$W_j(s_r) = \sum_{i=1}^{|A|} w_i v_i / \Big(\sum_{i=1}^{|A|} w_i^2 \sum_{i=1}^{|A|} v_i^2\Big)^{1/2}$$

where $w_i = 0$ iff $a_i \notin Q_j$, and $v_i = 0$ iff $a_i \notin A_{s_r}$.

This is Salton's 'cosine correlation measure'[72]. The variables w_i and v_i here are *attribute* weights, and may be either binary, or take on positive values reflecting d_i, i.e., the rarity of a_i in the relation in $A \times S$.

The weighting functions $W_i (i = 1, 2, 3)$ induce distributions of the informational objects on the real line, conditional on S, Q_j, and $\mathscr{R}$ in the case of W_3, and conditional on S, X, Q_j and $\mathscr{R}$ in the case of W_1 and W_2. Call these $m_i(z)$, respectively. Thus:

For Class I weighting functions:

$$m_1(z) = \frac{|W_1^{-1}(z)|}{|X|} = \frac{|\{s_r;\ W_1(s_r) = z\}|}{|X|}$$

$$= \frac{|\cup \{s_r;\ M_1(s_r) = \mathbf{e_n} \in W_L^{-1}(z)\}|}{|X|}$$

$$= \sum \frac{|\{s_r;\ M_1(s_r) = \mathbf{e_n} \in W_L^{-1}(z)\}|}{|X|}$$

$$= \sum_{W_1(\mathbf{e_n}) = z} \frac{|\{s_r;\ s_r \in X,\ \mathbf{e_n}\}|}{|X|}$$

$$= \Sigma r_n,\ \mathbf{e_n} \in W_L^{-1}(z).$$

*But we note that there has been some movement in classical theory towards a discrete receiver outcome space (*see* Egan[58])

†Suppose $\mathbf{e}_3$, $\mathbf{e}_7$ and $\mathbf{e}_{13}$ are each mapped by W_L to the value $z = 9.5$. Then this is equivalent to a statement that objects for which $\mathbf{e}_3 \vee \mathbf{e}_7 \vee \mathbf{e}_{13}$ evaluates to True are mapped to $z = 9.5$

Similarly,

$$m_2(z) = \Sigma f_n, \ \mathbf{e_n} \in W_L^{-1}(z).$$

For Class II weighting functions:
$m_1(z)$ and $m_2(z)$ are as above, but with r_n and f_n replaced by:

$$r_{n,\ m} = \frac{|\{s_r;\ s_r \in X,\ \mathbf{e_n} \wedge \mathbf{b_m}\}|}{|X|}$$

$$f_{n,\ m} = \frac{|\{s_r;\ s_r \in S\backslash X,\ \mathbf{e_n} \wedge \mathbf{b_m}\}|}{|S\backslash X|}$$

The values of $r_{n,\ m}$ and $f_{n\ m}$ each sum to 1 for all expressions $\mathbf{e_n} \wedge \mathbf{b_m}$. The values of $m_1(z)$, $m_2(z)$ and $m_3(z)$ each sum to 1 for all z-values defined by their weighting functions $W_i(\ldots)$. The functions W_1 and W_2 each map into the set $W_3(S)$, so that in general some of the z-events defined by $W_3(S)$ will be 'almost-impossible'[73]. For such events it is assumed $m_1(z)$ and/or $m_2(z)$ equal 0.

The *role* of weighting expressions (functions) in IR thus becomes clearer. It is to attach priorities (high rank-values) to the ELCs defined by a set-form query, without determining commitment to a particular logical-form query. In other words, weighting expressions serve to define *a sequence of logical-form queries.*

For Class I weighting functions, the sequence (L_j) will be:

$$(\mathbf{L_j}) : = \bigvee_{i=1}^{j} \mathbf{e_i} \text{ where } \mathbf{e_r} \prec \mathbf{e_s} \Leftrightarrow M_3(\mathbf{e_r}) > M_3(\mathbf{e_s}) \Leftrightarrow r < s.$$

For Class II functions it will be:

$$(\mathbf{L_j}) : = \bigvee_{i=1}^{j} (\mathbf{e} \wedge \mathbf{b})_i, \text{ where } (\mathbf{e} \wedge \mathbf{b})_r \prec (\mathbf{e} \wedge \mathbf{b})_s \Leftrightarrow M_3(\mathbf{e} \wedge \mathbf{b})_r > M_3(\mathbf{e} \wedge \mathbf{b})_s \Leftrightarrow r < s.$$

(We assume arbitrarily that M_3 maps *more* effective ELCs to *higher* values of z.) Thus for Class I and Class II functions, respectively, a sequence of logical-form queries $(\mathbf{L_j})$ will be associated with a sequence of ordered-pairs of Recall and Fallout values $((R_j,\ F_j))$, where

$$R_j : = \sum_{i=1}^{j} r_i \qquad \text{and } F_j : = \sum_{i=1}^{j} f_i$$

$$R_j : = \sum_{i=1}^{j} (r_{n,\ m})_i \quad \text{and } F_j : = \sum_{i=1}^{j} (f_{n,\ m})_i.$$

To return to an earlier terminology, the sequential-form query $(\mathbf{L_j})$ defined by $< S,\ Q_j,\ \mathscr{R},\ W_3(S) >$ determines a relation in the outcome space of $(R,\ F)$, i.e., in $[0,\ 1] \times [0,\ 1]$. Since R, F and G determine P and D, relations in other outcome spaces are also definable. We denote the relation defined by a sequential-form query mapping $(R,\ F)$ to its outcomes by $\mathscr{R}_{RF}$, and the relation defined by it mapping $(R,\ P)$ to its outcomes by $\mathscr{R}_{RP}$. (Other relations, e.g., that associated with $(R,\ P,\ F)$ are readily definable.) Various attempts at describing $\mathscr{R}_{RP}$ have been made. Perhaps the earliest are due to Cleverdon and Keen[25] who looked at mean Recall (etc.), and Rocchio and Salton who defined 'Normalized Recall' and related measures (*see* Salton,[74] p. 268). Other definitions have been the mean value of the Euclidean distance between $(R,\ P) = (\mathrm{O},\ \mathrm{O})$ and $(R,\ P) \in \mathscr{R}_{RP}$, and the mean value of $D(R,\ P)$, (for $(R,\ P) \in \mathscr{R}_{RP})$[9]. Another possibility would be the centre-of-mass of the relation $\mathscr{R}_{RP}$.

Before leaving the topic of object weighting, it is interesting to note:

- The general drift of interest within IR, since the 1960s, from attribute weighting ('term weighting'), to object weighting ('document weighting'), to the weighting of logical expressions. This has happened basically as (conceptually) *simpler* variables have been sought. (For example, to define attribute weights is to leave unspecified the mechanism for mapping the object to an outcome space: Why should a *sum* of such weights be assumed? Again, the notion of document weight is a clumsy one in that it leaves unspecified the procedure of grouping together, or arbitrarily ranking, documents with a common weight, and also in that it obscures the role of search logic.*
- The increasing interest in retrieval heuristics, especially when attribute dependencies are recognized. (It is perhaps not sufficiently apparent from the literature that such dependencies will *vary* according to whether one is defining objects in X, S or $S\backslash X$.) This problem appears to have two key components: how should iterations of retrieval define successive set-form queries?; and: how should iterations define successive object weighting functions, and their accompanying threshold parameters? Only large-scale *experimental* work can answer such questions in a satisfactory way. For theoretical contributions (some with an experimental component, and some giving attention to attribute-dependency) we refer selectively to: Attar and Fraenkel[75], Bookstein[76], Dillon *et al.*[77], Ide[78], Pietilainen[79], Robertson and Sparck Jones[80], Rocchio and Salton[81], Salton *et al.*[82], Smeaton[83], and Yu *et al.*[84].

Other, relateable interest has been in modification to $\mathscr{R}$ itself so as to optimize retrieval. See, for example, Cooper and Maron[85], Parker[86], Ide[78] and Tague[87]. (The two areas of procedure concerned can be labelled 'query learning' and 'object learning'.) The long-term effect of such interest will surely be the replacement of the sign-detection schema by a more general one in which the concept of 'control' becomes explicit, and the entities within the schema are primarily procedures rather than static variables, sets and relations.

Optimality in sequential-form queries

A problem of practical interest is the *optimal ordering* of the $\mathbf{e_n}$ (or of the conjuncts $\mathbf{e_n} \wedge \mathbf{b_m}$ for sequential-form queries defined by Class II weighting functions). Some optimal sequential-form queries are shown in Table 1. For related discussion we refer to Morse[88], Heine[89] and Stirling[90].

Robertson's 'Probability Ranking Principle' (PRP)[91] offers an alternative method of ranking the ELCs of a set-form query. This principle is stated ambiguously by

*The emphasis on 'monotonicity' in the literature of object weighting can, with hindsight, be seen to anticipate the 'return to logic'

Robertson, in that the means used to rank objects is given as (in our notation):

$$\phi(s_r) = \text{Prob(object is relevant} \mid \text{object is } s_r)$$
$$= \text{Prob}(s_r \text{ is relevant}).$$

But the latter statement identifies $\phi(s_r)$ with G, the Generality; and the former statement limits $\phi(s_r)$ to the values '1' (when s_r is relevant) and 'O' (s_r not relevant). Neither interpretation appears to be that intended, however. The author assumes that the PRP asserts the means of ranking is to be applied to weights, primarily, and only by derivation to objects, and that the expression intended is:

$$\phi(z) = \text{Prob}(s_r \in X \cap W_3^{-1}(z) \mid s_r \in W_3^{-1}(z))$$

$$= m_1(z)/(m_1(z) + k'm_2(z)),$$
$$\text{where } k' = G/(1 - G) > 0.$$

The contribution of Maron and Kuhns[92], to which Swets refers, appears to embody the same elliptical notation. However, $\phi(z)$ as we have interpreted it is monotonic with $m_1(z)/m_2(z)$, so that the PRP is equivalent to ranking the z-values by $m_1(z)/m_2(z)$. If the PRP is applied to the $\mathbf{e_n}$, or $\mathbf{e_n} \wedge \mathbf{b_m}$, then ranking by r_n/f_n and $r_{n,\,m}/f_{n,\,m}$ is entailed.* The PRP is thus open to the charge of obviousness, but that it should have been given an elegant theoretical foundation is a satisfying aspect of it. It does not, of course, determine a *unique* optimum sequential-form query, as demonstrated by the other means of ranking mentioned above.

Table 1. Different means of ordering the ELCs of a set-form query, so as to generate different optimal sequential-form queries

Type of optimality	Expression used to order $\mathbf{e_n}$
Recall-driven optimality	r_n
Fallout-driven optimality	$1/f_n$
Likelihood-ratio driven optimality	r_n/f_n
Precision-driven optimality	$Gr_n/[f_n(1 - G) + Gr_n]$
D-metric driven optimality	$[f_n(1 - G) + G(1 - r_n)]/[f_n(1 - G) + G]$
Euclidean-distance driven optimality	$(r_n^2 + (Gr_n/[f_n(1 - G) + Gr_n])^2)^{-½}$

Other means of generating logical-form queries

In the preceding two sections, the focus of the discussion was on the generation of logical-form queries using the sequential-form query as the generating device. The essential content of the sequential-form query was the analytical expression serving to rank the members of E. However, if only for completeness, we note here the possibility that the communication envelope can be reduced by other means. The theory (and technology) to do this is latent in the work at present being undertaken in many laboratories on 'language processing by software'. From the point of view of improvement to IR from classical databases the two key research objectives appear to be:

- OBJECTIVE 1: the identification of procedures to reduce narrative-form queries to set-form queries.
- OBJECTIVE 2: the identification of procedures to reduce set-form queries *either* to sequential-form queries *or* (directly) to logical-form queries.† (The assumption that this must be done by analytical functions of (A_{s_r}, Q_j) should be challenged, it is suggested.)

The essential interests are not in naive software modules that simply parse a narrative-form query, stripping off prepositions and case-endings (etc.), but in sophisticated algorithms that appeal to extensive, minimally-atomic, representations of knowledge, as well as clusterings of attributes and of objects in the database. Encouragement is given to the construction of such modules by the knowledge that the human mind can 'individuate' from the communication envelope to a specific member of it fairly quickly. (For example, a syntax can be imposed on a set of four attributes in, say, 60 seconds, without the human subject pondering over much upon the $2^{24} - 1$ excluded members.) This is done by what is said to be an appeal to 'aboutness' (see, for example, Maron[96]). *Representing* this procedure is of course just the problem.

Summary of query-types, and some measures of their effectiveness

In summary, given a set of objects S, a subset of relevant objects X, a set of attributes A, and a relation $\mathscr{R}$ in $S \times A$, i.e., a classical database, different query types can be distinguished. Each type of query, when input to an IR system, will yield a different type of output. The query types, and the system outputs, are closely related. Table 2 offers a summary of this picture, but excludes narrative-form queries.

COMMENTS ON THE TECHNOLOGICAL USEFULNESS OF THE SIGN-DETECTION SCHEMA

The following direct benefits to IR technology are suggested, in regard to the sign-detection schema:

- The distinction between the semantic and syntactic components of an expression of information need is clarified by the schema. Current IR technology in commercial use usually demands a logical-form query as input. An alternative technology would be one in which set-form queries were able to be input, and the search software would generate first a sequential-form

*In the reports by Robertson *et al.*[93] and van Rijsbergen *et al.*[94], the outcome event being ranked is defined ambiguously. The function concerned is defined initially as a random variable, but is later defined as an $|A|$ -dimensional vector of bits. Obviously, a *general* expression of the PRP still remains to be given

†*See*, for example, Salton *et al.*[95]

Table 2. Summary of basic IR query types

Input to IR system	Output of IR system	Retrieval effectiveness characterizations
'Set form query'		
$Q_j = \{a_i; i \in I\} \subset A$ This is associated with a set of logical expressions (see below: 'Logical-form query') defined by: $\{\mathbf{L_{jk}}\} = \{\bigvee_k \mathbf{e_n}\}$, where the $\mathbf{e_n}$ are 'elementary logical conjuncts' derived from Q_j. The $\mathbf{e_n}$ are given by $\mathbf{e_n} = \bigwedge_{i \in I} \mathbf{a_i^{k_i}}$, $k_i \in \{0, 1\}$. k indexes all possible disjunctions of the $\mathbf{e_n}$	A set of output sets $\{Y_{jk}\}$ defined by $Y_{jk} = \{s_r; L_{jk} = \text{True}\}$	The 'communication envelope', defined by $< \mathscr{R}, X, Q_j >$, can be characterized by a distribution of the $\mathbf{L_{jk}}$ induced by (R, P), or by univariate induced distributions (e.g., by R, F, P, D)
'Sequential-form query'		
This is a specific permutation of the members of $\{\mathbf{e_n}\}$, or a composition of a permutation of the $\mathbf{e_n}$. It is associated with both a sequence of logical-form queries, and a function W_L mapping $\{\mathbf{e_n}\}$ to an ordered outcome space such as the real line	A sequence of output sets (Y_u) defined by successive disjunctions in a sequence of the $\mathbf{e_n}$	A relation $\{(R_k, P_k)\}$ in the outcome space of (R, P), for example
'Logical-form query'		
This is an arbitrary member of the set $\{\mathbf{L_{jk}}\}$ defined by Q_j	A single set Y_{jk}	Values of R, P, F (etc.)

query (i.e., an optimal sequence of ELCs), and then an optimal logical-form query (as described by Heine[97]). Such technology should be rooted in a *science* of IR which would have provided optimal parameters for such procedures on the basis of studies of the communication envelope.

- The schema draws attention to the global significance of Q_j, the set-form query. Perhaps too much has been written to little point on object (document) weighting. The reviewer suggests that research is now urgently required on the ability of system users to choose optimal set-form queries, possibly building onto Taylor[98] and Kochen and Badre[99]. On such a basis the technologies of both logical-form query generation, and of improvement to $\mathscr{R}$, should be enhanced. So should strategies of user-education in regard to database use.
- It is believed that the distribution of $\{\mathbf{L_{jk}}\}$ on the outcome-space of (R, P), i.e., what we have termed the *RP* probability surface, provides a useful industrial standard. That is, it embodies a method of testing a database for effectiveness (as distinct from testing search software for efficiency). A complete standard in this regard would, however, need also to define an acceptable way of generating sets of informing objects in the database. Possibly a sampling method based on what has been termed the 'Virtual Attribute Technique'[10] may be of use here. This simply involves defining a set X by $\{s_r; a_i \in A_{s_r}\}$, but excluding the attribute a_i thereafter from the relation $\mathscr{R}$, i.e., making it 'virtual'. The a_i could be sampled from the Zipfian distribution of attribute-frequencies.

RECOVERY OF THE CLASSICAL SIGNAL-DETECTION SCHEMA

The classical theory of signal-detection theory embodies a schema in which the outcome-space is continuous rather than discrete. The sets X and $S\backslash X$ are thus continuous also. The receiver-induced distributions, $g_1(y)$ and $g_2(y)$, $y \in Re$, are such that the probability that a transmitting object [non-transmitting object] is mapped to $(y, y + dy)$ is given by $g_1(y)dy\,[g_2(y)dy]$. If a threshold variable, x say, is defined, such that only objects mapped to y-values greater than x are *identified* as transmitting, then:

$$R(x) = \int_x^\infty g_1(y)\ dy \text{ and } F(x) = \int_x^\infty g_2(y)\ dy.$$

Such statements can be generated by the sign-detection schema in two ways. First, we could say that the definite integrals stand as approximations for summations, for example:

$$\int_x^\infty g_1(y)\ dy \doteq \sum_{y > x} m_1(y)$$

where it is understood $m(y) = 0$ when $y \notin W_3(S)$. Under this interpretation the density functions of classical theory are *modelling* functions. Secondly, we could see the functions g_i as *limiting forms* of the distributions m_i, where S and X are infinite sets. However, we seem to need also to assume that the set-form query becomes infinite. Alternatively, we could see the g_i as modelling (or being limiting forms of) the distributions of $m_i(z)$ over the integers, according to the rank values of the members of $W_3(S)$. This tentative and indicative comment requires a much fuller analysis, of course, perhaps resting on an appeal to the Stieltjes integral (Robertson, *pers. comm.*), and by examining again the treatment given to discrete receiver outcome-spaces in the classical literature.

In the first, novel attempt at relating IR to (classical) signal-detection theory, Swets choose to model m_1 and m_2 by Normal density functions. However, the large 'spike' of probability that attaches to the all-negated ELC, i.e., to

$$\mathbf{e}^0 = \bigwedge_{i \in I} \mathbf{a}_i^0 = \text{Not}\left(\bigvee_{i \in I} \mathbf{a}_i^1 \right)$$

for non-relevant objects, and which also attaches to the lowest-ranking value of z, was ignored by Swets. (This is assuming that $\mathbf{e}^0$ is mapped by W_L to the lowest z-value.) The explanation of this spike is simply that most of the objects in the set $S \backslash X$ will be assigned *none* of the attributes in Q_j, for most instances of IR. (Technically, it seems to be required that $|Q_j| \ll |A|$, and $|X| \ll |S|$ for the spike to appear.) The values $R \doteq 1$ and $F \doteq 1$ are thus associated with a logical-form query that disjoins to $\mathbf{e}^0$. These values were excluded from Swets's figures showing observed (R, F)-data in consequence of the method of scaling R- and F-values employed by Swets. The point of this criticism is simply to demonstrate a limitation of the classical signal-detection schema employing continuous modelling functions, but it is offered with hindsight and does not diminish the central value of Swets's work.

Lastly, two criticisms of classical signal-detection theory itself are

- Its mapping of a transmitter's effects on a receiving apparatus is over-specific (it addresses only one permutation of the receiver's possible responses).
- It excludes a portrayal of the control of the receiver. (What is it that orientates a receiver to a transmitter, i.e., endows the receiver with purpose?)

FURTHER COMMENTS ON THE *RF* AND *RP* PROBABILITY SURFACES

It was noted in the last section that when $|Q_j| \ll |A|$ and $|X| \ll |S|$ (which conditions will usually obtain in practice), the value of f that corresponds to $\mathbf{e}^0$ will be close to unity. The values of f_n for all other ELCs must accordingly be close to zero. On this basis, it seems reasonable to distinguish two populations of logical-form queries. The populations are defined and distinguished by the criterion: Does **L** involve (contain) a disjunction with $\mathbf{e}^0$? The two populations will be referred to as 'the population of useful [useless] logical-form queries'. An example of a useful logical-form query is thus $\mathbf{L} = \mathbf{e}_r \vee \mathbf{e}_s \vee \mathbf{e}_t$; an example of a useless one is $\mathbf{L} = \mathbf{e}^0 \vee \mathbf{e}_r \vee \mathbf{e}_s$. The two populations map to the outcome space of (R, F) in clearly distinct ways. If we denote the values of r and f attaching to $\mathbf{e}^0$ by r^0 and f^0, then the population of useful logical-form queries will be characterized by $F \doteq 1 - f^0 \doteq 0$, and will be distributed over $R \in (0, 1 - r^0)$, while the population of useless logical-form queries will be characterized by $F \doteq f^0 \doteq 1$ and will be distributed over $R \in (0, 1)$. A sketch of the *RF* probability surface that will obtain when *both* populations are present is shown in Figure 1. It is an intriguing conjecture that the classical signal-detection schema's portrayal of receiver response, which appears as a marginal distribution of the *RF* surface, may be a combinatorial response to a deeper randomness, as is portrayed by sign-detection theory. That is, if one relabels useful logical-form queries as 'signal', and useless logical-form queries as 'noise', then what we conceive of as 'observation of an entity' can be interpreted as measurement of 'the *probability* of what *can* be observed', i.e., a value of R.

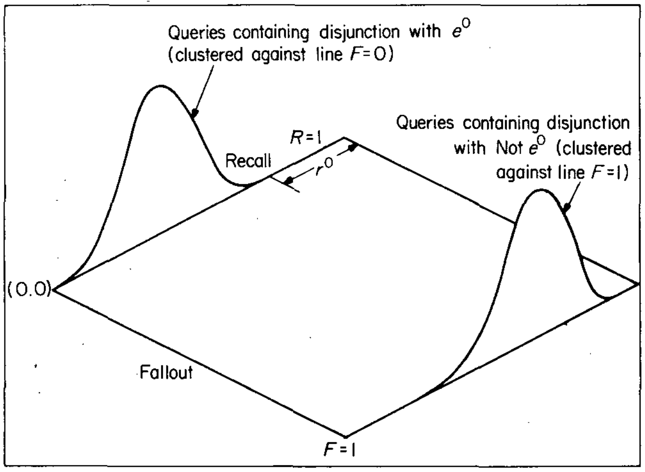
Queries containing disjunction with e^0 (clustered against line F=0)
Queries containing disjunction with Not e^0 (clustered against line F=1)
R=1
Recall
r^0
(0.0)
Fallout
F=1

Figure 1. RF probability surface with both populations present

In the case of the *RP* probability surface, useless logical-form queries map to near the line $P = 0$, since $F \doteq 1$ implies:

$$P = \frac{GR}{GR + (1 - G)} \doteq 0, \text{ for } G \doteq 0, \text{ i.e.,}$$

$$\text{for } |X| \ll |S|.$$

The other population will be distributed in a less-predictable manner over the outcome space of (R, P), i.e., over $[0, 1] \times [0, 1]$. However, an important constraint to the mapping needs to be recognized, in that all logical-form queries will be mapped to points that lie 'within' a perimeter-relation defined by the particular sequential-form query:

$$(\mathbf{L_j}) = \bigvee_{i=1}^{j} \mathbf{e_i}, \text{ where } \mathbf{e_r} \prec \mathbf{e_s} \Leftrightarrow r_r/f_r > r_s/f_s.$$

As suggested above, a major *scientific* task is the description of the *RP* probability surface, and it is hoped that serious work on describing and modelling this surface will be undertaken in the near future. Without such work, little understanding is possible of the sensitivity of IR to choice of set-form query, or sequential-form query, and IR *technology* will be functioning on a basis of flimsy real-world assumption. Once sound data has been obtained, perhaps two modelling approaches could be pursued. One would involve modelling the surface using an analytical function of two real variables. The other would involve modelling the values of r_n and f_n and then deriving a model surface by simulation. To illustrate the latter approach, let us suppose the r_n are modelled by:

$$r_n = r(\mathbf{e_n}) = r\left(\bigwedge_{i \in I} \mathbf{a}_i^{k_i} \right)$$

$$\propto -\left(\sum_{i \in I} k_i \log(d_i) \right). \text{ Bin } \left(|Q_j|, \sum_{i \in I} k_i \right)$$

where Bin(M, N) is a binomial distribution (with parameter to be supplied) over the events $M = 0, 1, 2, \ldots, N,$

and as before d_i is the specificity of attribute a_i. A constant of proportionality will need to be chosen so that:

$$\sum_{n=1}^{2^{|Q_j|}} r_n = 1$$

The values of d_i might be obtained by sampling from a Zipfian distribution of attribute frequencies in the database, or from a negative-exponential distribution. However, they are set equal here to the arbitrary values 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}, for a set-form query of size 4 attributes. The mean of the binomial distribution will be set at 2.8. The values of f_n (but excluding f^0) will be modelled by:

$$f_n = f(\mathbf{e_n}) \propto (1 - f^0) \prod_{i \in I} (k_i d_i)$$

and f^0 itself will be modelled by the *ad hoc* value 0.98. The value of G is set to 0.01, i.e., one object in a hundred in the database is relevant. The *RP* probability surface so modelled is portrayed in Figure 2a. The *RF* probability surface (independent of G) is portrayed in Figure 2b. These figures show the surfaces as isometric ones, with points grouped into cells as shown. A more literal portrayal of the *RP* probability surface (strictly, *relation*) is shown in Figure 3, but with all points for which $P < 0.01$ excluded.

By way of contrast, Figures 4a and 4b show the *RP* probability surfaces and *RF* probability surface for concatenated data arising from a study of 31 instances of communication through MEDLINE, and Figures 5a and 5b show these surfaces for a particular communication envelope. (These are intended to be indicative only — a detailed report is in preparation.) Source data for Figures 4 and 5 are given in Heine[11].

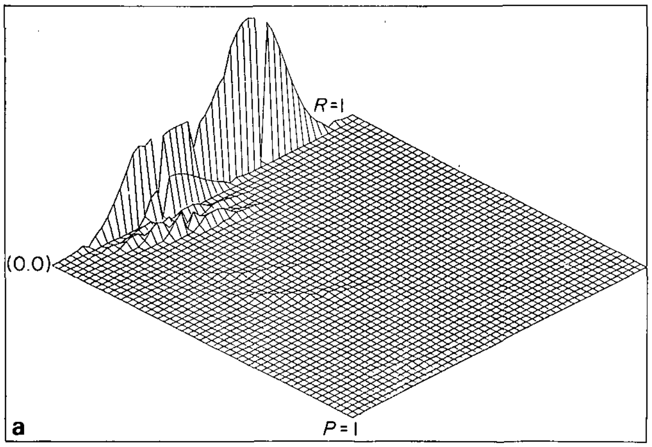

Figure 2a. RP probability surface

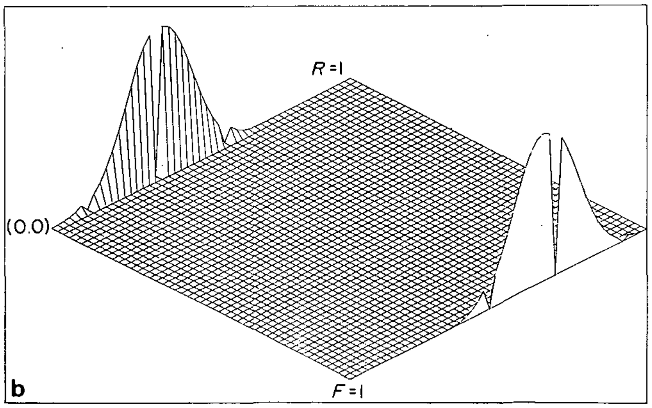

Figure 2b. RF probability surface

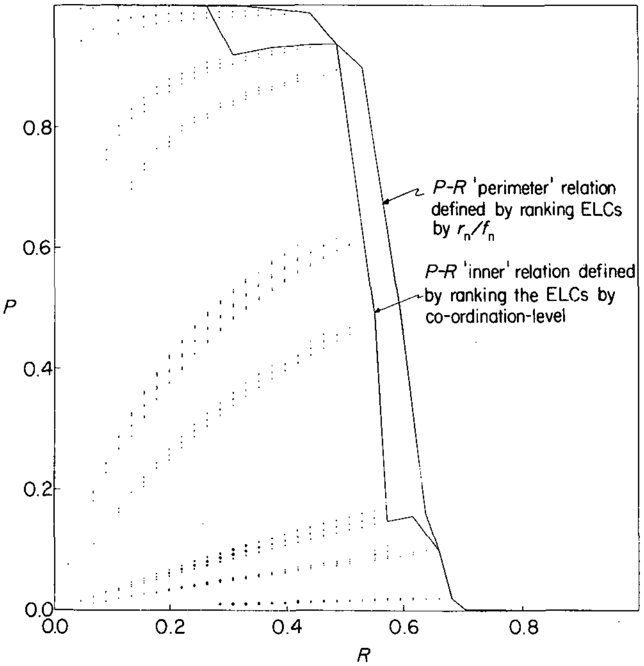

Figure 3. RP probability surface excluding all points for which P < 0.01

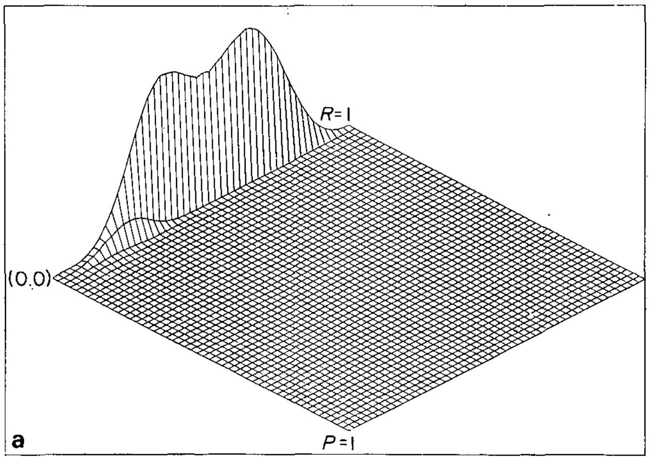

Figure 4a. RP probability surface for concatenated data arising from study of communication through MEDLINE

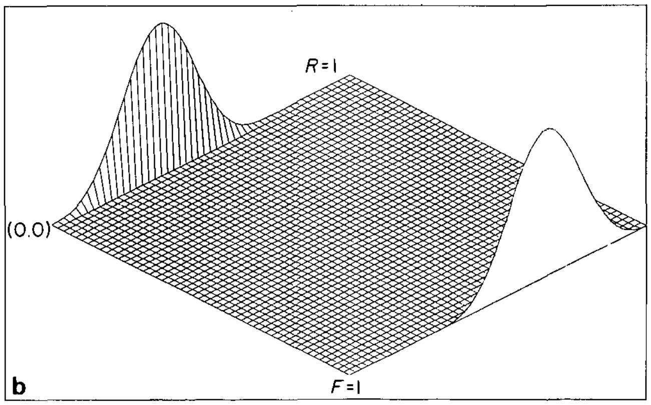

Figure 4b. RF probability surface for concatenated data arising from study of communication through MEDLINE

Lastly here it is suggested that theoretical work is needed on clarification of the relationship between combinatorial theory and the actual shape of the *RP* probability surface. Characteristically, 'hills' of point emanate from the origin and radiate towards the perimeter relation. Why should this be so? How is this related to the equation (derived from a continuous modelling function, $P = P(R, F)$):

$$\left.\frac{\partial P}{\partial F}\right|_{R = R'} = \frac{1}{(c_1 + c_2 \mathrm{F} + c_3 F^2)}$$

where $c_1 = 1/(G-1)$; $c_2 = -2$; $c_3 = (G-1)/(GR')$? A further theoretical question is: what can be predicted concerning the *RP* probability surface when the system user recognizes successively more-inclusive set-form queries, i.e., queries such that $Q_1 \subset Q_2 \subset Q_3 \subset \ldots$? This would seem to involve the joint study of combinatorics, and branching processes applied to r_n and f_n. It should be possible to relate the work of van Rijsbergen[100] to this problem.

RELATED WORK ON SIGNAL-DETECTION THEORY AND IR

Work on the effectiveness of sequential-form queries

In Swets's[5] paper, a measure of the separation of the distributions $m_1(z)$ and $m_2(z)$ was proposed. This was in the context of a novel discussion of IR in which the notion of a signal-detection schema of IR was proposed. However, there was some ambiguity in regard to whether the observed distributions or modelling density functions were the subject of the schema. From the perspective offered earlier in this review, perhaps the key factors were that a description of a *sequential-form query* was being sought, and that the actual (cardinal) values of the object weights were brought into the description (rather than the ranks of those values). The measure introduced by Swets, which is labelled here E_s, was defined by:

$$\frac{E(W_1(X)) - E(W_2(S \backslash X))}{(V(W_1(X)))^{½}}$$

— in our notation, where $E(\ldots)$ and $V(\ldots)$ denote expectation and variance, and where it was assumed $V(W_1) = V(W_2)$.

Swets's primary concern, to which his novel schema was accessory, was to predict the effect of varying a threshold parameter (determining retrieval and non-retrieval) on $\mathscr{R}_{RF}$, when E_s was held constant. We can interpret this to imply 'when constrained by a given set-form query'. The (R, F)-relation determined by variation in this threshold determined what Swets referred to as an 'operating characteristic' curve. Modelling this curve, using Normal density functions to describe the separate m_1 and m_2 distributions was perhaps Swets's main concern. He pointed out that under the Normal-density assumption, the modelling relation became a straight line when the R and F values were transformed by:

$$x = \Phi_c^{-1}(\beta) \text{ and } \Phi_c(x) = \int_x^\infty (2\pi)^{-½} \exp(-u^2/2) du$$

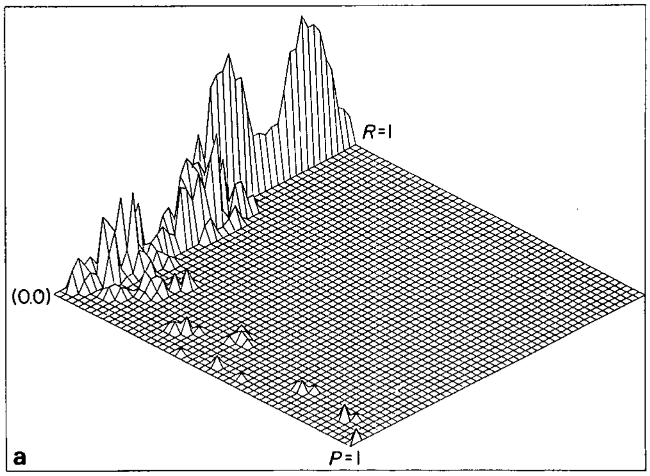

Figure 5a. RP probability surface for a particular communication envelope

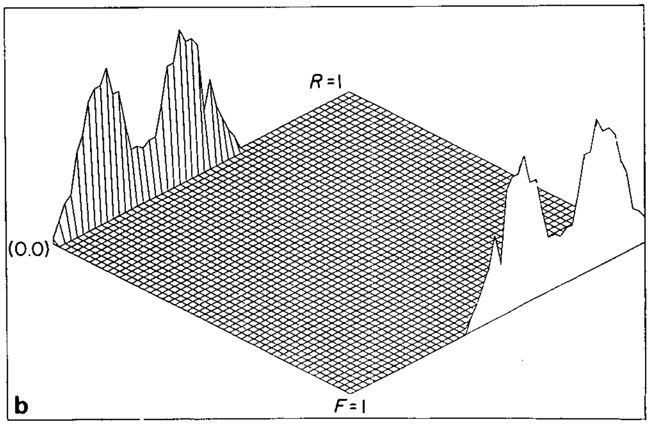

Figure 5b. RF probability surface for a particular communication envelope

where β is a sample value of R or F, and x is the threshold parameter. However, as previously remarked, one effect of this is exclusion from such plots of the point $R = 1 = F$ which will always appear in the relation and which corresponds to the whole database being retrieved. (The values of F shown in Swets's figures characteristically fall in the range (0, 0.40) (approximately).)* It should be emphasized that the variable z was largely *implicit* in Swets's analyses, in that his (R, F) data were obtained *directly* from experimental data, i.e., from search output ranked by (say) coordination level for Cranfield data, or by an unspecified weighting expression (cosine correlation?) applied to records accessed by the SMART system, then at Harvard. Only passing attention was given by Swets to either $\mathscr{R}_{RP}$ (no experimental data are reported) or the role of logical expressions in database searching. However, that Swets was aware of the *need* to incorporate logic into the schema is evident from his statement:

> The choice of an 'and' or 'or' relationship among a number 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[5] (p.248).

*Such large ranges of F, for useful logical-form queries, are not in agreement with the author's findings.[9] A study of 31 instances of communication through Medline showed F generally to lie in the much smaller range (0, 0.013). This was for algorithmically-defined set-form queries, however. Possibly the data Swets was dealing with involved arbitrary, sub-optimal set-form queries

The measure E_s was amended by Brookes[101] to:

$$S = \frac{E(W_1) - E(W_2)}{(V(W_1) + V(W_2))^{1/2}}$$

which was subsequently proved by Robertson[49] to be equivalent to an alternative measure, A, also put forward by Swets:

$$A = \tfrac{1}{2} \ \text{erfc}\left(\frac{-S}{(2)^{1/2}}\right) \ ; \ \text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2)dt.$$

Thus either S or A is redundant. However, perhaps S should be preferred (if $\mathscr{R}_{RF}$ is to be characterized at all) since it relates to a satisfying geometrical property of $\mathscr{R}_{RF}$ when the relation is transformed using Φ_c^{-1} and when the Normal density models apply. This is that S measures the distance from the point $R = 0.5 = F$ to the line through the relation (when transformed). But this is still to retain loyalty to the portrayal of retrieval effectiveness by R and F, and it is to ignore the failure of the Normal density model to accommodate the point $R = 1 = F$.

Multivariate generalizations

One generalization of classical signal-detection theory as applied to IR involves a refinement to the definition of set X. Instead of relevance being a binary property, i.e., $s_r \in X$ or $s_r \notin (S\backslash X)$, a continuum of relevance values is introduced. This has been discussed by Hutchinson[22]. Whether this has a strong pragmatic value is not yet clear. As previously suggested, a *qualitative* partitioning of set X by 'types of relevance' may be more useful. Undoubtedly, experimental work is now badly needed on the influence of both 'degrees of relevance' and 'types of relevance' on the communication envelope, though within the schema of sign-detection rather than signal-detection.

A generalization of the outcome space of the receiver has also been treated, in an IR context. This is where two weighting expressions are recognized, one of which is the conventional one (defining the functions W_i), the other being the age of the object, a_i say, at the time of retrieval (defining functions W_i' say)[102]. Retrieval using both types of weight jointly is then effected by an expression such as:

$$c_1 W_3(s_r) + c_2 W_3'(s_r), \text{ where } c_1 > 0 > c_2.$$

This generalizes to any quantitative attribute, or set of same. Heine's work was weakened however by its commitment to the Normal densities of classical theory, and by its not explicitly incorporating search logic. Work is needed, it is suggested, on the reconciliation of Boolean searching with multivariate weighting expressions, perhaps building onto the work of Lipski and Marek[23]. Judgement on Bookstein's[65] view that Boolean retrieval does not provide a complete substitute for retrieval using weights is, it is suggested, in abeyance while such work remains to be done. Angione's view was clearly that a logical equivalent of weighting (by implication either univariate or multivariate) can always be found.

The precision–recall relation defined by a sequential-form query

As noted in the section on the effectiveness of sequential-form queries, the emphasis in Swets's work on sequential-form queries was on the interplay of R and F. Little attention was given to the relation defined by (R, P) for a specific sequential-form query. Possibly this was because the data investigated were 'confounded' (i.e., 'averaged'), so that G was a variable, or possibly it was to clarify the role of the measure E_s. Discussion of this relation was subsequently offered by Heine[103, 104] and Bookstein[105, 106]. Bookstein's[104] paper was, it is suggested, particularly useful in drawing attention to the significance of the ratio $m_1(z)/m_2(z)$ as an influence on the relation. However, it seems reasonable to say that interest in the relation *for a given sequential-form query* has been overtaken by interest in the *variety* of such relations when the set-form query is held constant, i.e., in the RP probability surface. The interest of the 1970s in optimum weighting functions, which was usually in the context of this relation, has probably now passed its zenith, as interest in retrieval heuristics grows. Discrimination functions, based on data from a partitioning of set Y, seem to offer a more useful line of research. What still seems to be limited, however, in the author's view, is strong interest in optimizing the membership and size of the set-form query. It is this, of course, which provides the global constraint on the effectiveness of any sequential-form, or logical-form query.

Application of signal-detection theory to feedback

The paper by Yu *et al.*[84] appears to be unique in relating Swets's work to the problem of improving IR through feedback of relevance judgements. The original, dual-Normal density functions of Swets are used to generate theorems on the effectiveness of feedback-modified queries. (A 'query' for this purpose is a set-form query.) Apart from this interest, these authors describe relationships between the values of $E(W_2)$ and $E(W_1)$, and $V(W_2)$ and $V(W_1)$, for original and modified queries. As interest in retrieval heuristics grows, this work may provide a fertile basis of a useful general schema, once it has been transformed from a continuous footing to a discrete one. (An attempt in the latter direction was made by the author but on a limited basis only, and largely to accommodate dependencies of attributes in X, $S\backslash X$, and S[97]. We have already indicated in the section on weighting expressions as a means of selecting logical-form queries what appear to be the major contributions to feedback-supported IR, at the present time.)

SUMMARY — OTHER PROBABILISTIC SCHEMAS OF IR

This review has attempted to show how the interplay of probabilities that is the concern of signal-detection theory can now be viewed. The kernel of this 'paradigm' is the partitioning of a set of informational objects into an informing set (X) and its complement ($S\backslash X$). It is the *query*, in the various constructions that have been attached to it, that elicits paired probability distributions from this partitioning, which are central to the schema. As such, the query acts an *informational* role that is analogous to that of 'apparatus' in classical science, i.e., science dealing with the material world. Ideally, this review would have attempted to bring under the signal-detection (sign-detection) conceptual umbrella other probabilistic schemas of IR. However, it was considered that the extent of discussion that would be needed would imbalance this review, and also be presumptuous in view of

- existing theoretical uncertainties that have been indicated,
- the need for frequent *interpretations* of published contributions.

We refer instead, and selectively, to Maron and Hillman[107], Uhlmann[108], Landry[109], Turski[110], Pawlak[111], Bookstein and Swanson[112], Dabrowski[113], Gebhardt[45], Ludwig and Glockmann[44], Bookstein and Cooper[114] and Bookstein[115]. For further citations, the indicative review by Robertson[116] and the analytical review by van Rijsbergen *et al.*[94] are referred to.

CONCLUSIONS

What we have termed 'sign-detection theory' is a construction that has arisen naturally from classical signal-detection theory. The construction is a useful one, in that it provides an integrating device for otherwise separate concepts of database, search-expression, relevance and retrieval effectiveness. In particular, it allows the *individual* search expression (or 'logical-form query' as we have termed it) to be portrayed within a broader picture. The notions of object weighting (at least in the univariate-weight case) and ranking of objects are included in it. It provides a clear basis for experimental design in IR, and also a motivation for experiment. Its weaknesses, on the other hand, are perhaps in its commitment to the idea that 'relevance' is inevitably an *a priori* entity, i.e., its assumption that informing cannot be a creative process, and its not incorporating the control structures that in the 'real IR world' determine retrieval, transmission (when it exists), and database construction. However, whatever limitations we may recognize, the debt owed to Swets for providing an hospitable prototype schema, and to Angione for pointing to the necessary link between search logic and weighting methods, is a major one and deserves wider recognition.

REFERENCES

1 **Heaps, H S** *Information retrieval: computational and theoretical aspects* Academic Press, USA (1978)

2 **Salton, G** *Dynamic information and library processing* Prentice Hall, USA (1975)

3 **Salton, G and McGill, M J** *Introduction to modern information retrieval* McGraw Hill, USA (1983)

4 **van Rijsbergen, C J** *Information retrieval* 2nd edn Butterworths, UK (1979)

5 **Swets, J A** 'Information retrieval systems' *Science (USA)* Vol 241 (1963) pp 245–250

6 **Swets, J A** 'Effectiveness of information retrieval methods' Air Force Cambridge Research Labs., Bedford, Mass, USA (1967) (also as the Laboratories' Report AFCRL-67-0412; and published separately by Bolt, Beranek and Newman)

7 **Swets, J A** 'Signal detection as a model of information retrieval' in *La simulation du comportement humain: the simulation of human behavior: actes d'un symposium O.T.A.N., Paris, 1967* **F Brisson and M de Montmollin (Eds)** Dunod, France (1969) pp 253–267

8 **Swets, J A** 'Effectiveness of information retrieval methods' *Am. Doc.* Vol 20 (1969) pp 72–89

9 **Heine, M H** 'The extension and application of Swet's theory of information retrieval' PhD thesis, Computing Laboratory, University of Newcastle upon Tyne, UK (1981)

10 **Heine, M H** 'Simulation and simulation experiments' in *Information retrieval experiment* **K Sparck Jones (Ed.)** Butterworths, UK (1981) pp 179–198

11 **Heine, M H** 'Sign detection theory and its application' Paper presented at *Symposium on Empirical Foundations of Information and Software Science, Georgia Institute of Technology, Atlanta, Nov. 1982* (1982) [Reprinted in: *Inf. Process. & Manage.* Vol 20 (1984) pp 47–61

12 **Brookes, B C** 'Communicating research results' *Aslib Proc. (GB)* Vol 16 (1963) pp 7–21

13 **Heather, M J** 'Knowledge syngenesis from full text' (Unpublished paper presented at *BCS IRSG Res. Colloq. Sheffield, 1983*)

14 **Bookstein, A** 'Relevance' *J. Am. Soc. Inf. Sci.* Vol 30 (1979) pp 269–273

15 **Buckland, M K** 'Relatedness, relevance and responsiveness in retrieval systems' *Inf. Process. & Manage.* Vol 19 (1983) pp 237–241

16 **Le Claire, K A** 'Consumer "processing" of information: fact or fiction?' *Eur. Res.* Vol 9 (1981) pp 134–143

17 **Derr, R L** 'A conceptual analysis of information need' *Inf. Process. & Manage.* Vol 19 (1983) pp 273–278

18 **Krikelas, J** 'Information-seeking behavior: patterns and concepts' *Drexel Lib. Q.* Vol 19 (1983) pp 5–20

19 **Saracevic, T** 'The concept of "relevance" in information science: a historical review' in *Introduction to information science* **T Saracevic (Ed.)** Bowker, USA (1970) pp 111–151

20 **Saracevic, T** 'Relevance: a review' *J. Am. Soc. Inf. Sci.* Vol 26 (1975) pp 321–343

21 **Cuadra, C A and Katter, R V** *Experimental studies of relevance assessments* (3 vols) Systems Development Corp., USA (1967) [*See also* their 'Opening the black box of relevance' *J. Doc. (GB)* Vol 23 (1967) pp 291–303]

22 **Hutchinson, T P** 'An extension of the signal detection model of information retrieval' *J. Doc. (GB)* Vol 34 (1978) pp 51–54

23 **Lipski, W and Marek, W** 'Information systems: on queries involving cardinalities' *Inf. Syst. (GB)* Vol 4 (1979) pp 241–246

24 **Rees, A M and Saracevic, T** 'Conceptual analysis of questions in information retrieval systems' *Proc.*

Am. Doc. Inst. Part 2 (1963) pp 175–177

25 **Cleverdon, C W, Mills, J and Keen, E M** *Factors determining the performance of indexing systems* (2 vols) Aslib, UK (1966)

26 **Cleverdon, C W** 'The Cranfield tests of index language devices' *Aslib Proc. (GB)* Vol 19 (1967) pp 173–194

27 **Keen, E M and Digger, J A** 'Report of an information science index languages test' Aberystwyth, College of Librarianship Wales, UK (1972)

28 **Keen, E M** 'The Aberystwyth index languages test' *J. Doc. (GB)* Vol 29 (1973) pp 1–35

29 **Heine, M H** 'The "question" as a fundamental variable in information science' in *Theory and application of information research* **O Harbo and L Kajberg (Eds)** Mansell, UK (1977) pp 137–145

30 **Sparck Jones, K** 'Retrieval system tests 1958–1978' in *Information retrieval experiment* **K Sparck Jones (Ed.)** Butterworths, UK (1981) pp 213–255

31 **Sparck Jones, K** 'The Cranfield tests' in *Information retrieval experiment* **K Sparck Jones (Ed.)** Butterworths, UK (1981) pp 256–284

32 **Belkin, N J** 'Ineffable concepts in information retrieval' in *Information retrieval experiment* **K Sparck Jones (Ed.)** Butterworths, UK (1981) pp 44–58

33 **Minker, J** 'Information storage and retrieval: a survey and functional description' *ACM SIGIR Forum* Vol 12 (1977) pp 1–108

34 **Doszkocs, T E and Rapp, B A** 'Searching MEDLINE in English: a prototype user interface with natural language query, ranked output, and relevance feedback' *Proc. ASIS Annu. Meet.* Vol 16 (1979) pp 131–139

35 **Cooper, W S** 'The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness' *Inf. Process. & Manage.* Vol 12 (1976) pp 367–375

36 **Cleverdon, C W and Kidd, J S** 'Redundancy, relevance and value to the user in the outputs of information retrieval systems' *J. Doc. (GB)* Vol 32 (1976) pp 159–173

37 **King, D W and Bryant, E C** *The evaluation of information services and products* Information Resources Press, USA (1971)

38 **Lancaster, F W and Climenson, W D** 'Evaluating the efficiency of a document retrieval system' *J. Doc. (GB)* Vol 24 (1968) pp 16–40

39 **Vickery, B C** *Techniques of information retrieval* Butterworths, UK (1970)

40 **van Rijsbergen, C J** 'Retrieval effectiveness' *Prog. Commun. Sci.* Vol 1 (1979) pp 91–118

41 **Cooper, W S** 'Expected search length: a single measure of retrieval effectiveness based on a weak ordering action of retrieval systems' *Am. Doc.* Vol 19 (1968) pp 30–41

42 **Gebhardt, F** 'A simple probabilistic model for the relevance assessments of documents' *Inf. Process. & Manage.* Vol 11 (1975) pp 59–65

43 **Guazzo, M** 'Retrieval performance and information theory' *Inf. Process. & Manage.* Vol 13 (1977) pp 155–165

44 **Ludwig, B M and Glockmann, H P** 'The formal analysis of document retrieval systems' *J. Am. Soc. Inf. Sci.* Vol 26 (1975) pp 51–55

45 **Radecki, T** 'Retrieval system models of documents indexed by weighted descriptors' Prace Naukowe Politechniki Wroclawskiej, no. 4, Central Library and Information Centre, Technical University of Wroclaw (1980)

46 **Heine, M H** 'Distance between sets as an objective measure of retrieval effectiveness' *Inf. Storage & Retr.* Vol 9 (1973) pp 181–198

47 **Heine, M H** 'The flow of control in a communication process' Paper presented at *10th International Congress on Cybernetics, Namur, Aug. 1983* [In *Cybernetica* Vol 27 (1984) pp 57–64]

48 **Lipski, W** 'On semantic issues connected with incomplete information databases' *ACM Trans. Database Syst.* Vol 4 (1979) pp 262–296

49 **Robertson, S E** 'Parametric description of retrieval tests' *J. Doc. (GB)* Vol 25 (1969) pp 1–27, 93–107

50 **Cooper, W S** 'On selecting a measure of retrieval effectiveness' *J. Am. Soc. Inf. Sci.* Vol 24 (1973) pp 87–100, 413–424

51 **Farradene, J** 'The evaluation of information retrieval systems' *J. Doc. (GB)* Vol 30 (1974) pp 195–209

52 **Rees, A M** 'Evaluation of information systems and services' *Annu. Rev. Inf. Sci. & Technol.* Vol 2 (1967) pp 63–86

53 **Vickery, B C** (1966) [In **Cleverdon, Mills and Keen** (1966) see Reference 25 q.v.]

54 **Jardine, N and van Rijsbergen, C J** 'The use of hierarchic clustering in information retrieval' *Inf. Storage & Retr.* Vol 7 (1971) pp 214–240

55 **van Rijsbergen, C J** 'Foundation of evaluation' *J. Doc. (GB)* Vol 30 (1974) pp 365–373

56 **Bollmann, P and Cherniavsky, V S** 'Measurement-theoretical investigation of the MZ-metric' in *Information retrieval research* **R N Oddy** *et al.* **(Eds)** Butterworths, UK (1981) pp 256–267

57 **Hohn, F E** *Applied Boolean algebra* (2nd edn) Macmillan, UK (1966)

58 **Egan, J P** *Signal detection theory and ROC analysis* Academic Press, USA (1975)

59 **Evans, L** 'Methods of ranking SDI and IR outputs' Inspec Report R73/18; OSTI Report no. 5184, UK (1973) [This contains review material not in **Evans**[60]]

60 **Evans, L** 'Methods of ranking SDI and IR outputs: final report' Inspec Report R75/23; OSTI Report no. 5232, UK (1975)

61 **Sager, W K and Lockemann, P C** 'Classification of ranking algorithms' *Int. Forum Inf. & Doc. (FID)* Vol 1 (1976) pp 12–25

62 **Noreault, T, McGill, M and Koll, M B** 'A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment' in *Information retrieval research* **R N Oddy** *et al.* **(Eds)** Butterworths, UK (1981) pp 57–76

63 **Angione, P V** 'On the equivalence of Boolean and weighted searching based on the convertibility of query forms' *J. Am. Soc. Inf. Sci.* Vol 26 (1975) pp 112–124

64 **Brandhorst, R T** 'Simulation of Boolean logic constraints theory to the use of term weights' *Am. Doc.* Vol 17 (1966) pp 145–146

65 **Iker, H P** 'Solution of Boolean equations through the use of term weights to base two' *Am. Doc.* Vol 18 (1967) p 47

66 **Bookstein, A** 'Fuzzy requests: an approach to weighted Boolean searches' *J. Am. Soc. Inf. Sci.* Vol 31 (1980) pp 240–247

67 **Salton, G, Fox, E and Wu, H** 'Extended Boolean information retrieval' Technical Report TR 82-511, Department of Computer Science, Cornell University, USA (1982)

68 **Bookstein, A** 'On the perils of merging Boolean and weighted retrieval systems' *J. Am. Soc. Inf. Sci.* Vol 29 (1978) pp 156–158

69 **Radecki, T** 'Reducing the perils of merging Boolean and weighted retrieval systems' *J. Doc. (GB)* Vol 38 (1982) pp 207–211

70 **Miller, W L** 'A probabilistic search strategy for MEDLARS' *J. Doc. (GB)* Vol 27 (1971) pp 254–266

71 **Sparck Jones, K** 'A statistical interpretation of term specificity and its application to retrieval' *J. Doc. (GB)* Vol 28 (1972) pp 11–21

72 **Salton, G** *Automatic information organisation and retrieval* McGraw Hill, USA (1968)

73 **Barr, D R and Zehna, P W** *Probability* Brooks/Cole, USA (1971)

74 **Salton, G** *The SMART retrieval system — experiments in automatic document processing* Prentice Hall, USA (1971)

75 **Attar, R and Fraenkel, A S** 'Experiments in local metrical feedback in full-text retrieval systems' *Inf. Process. & Manage.* Vol 17 (1981) pp 115–126

76 **Bookstein, A** 'Information retrieval: a sequential learning process' *J. Am. Soc. Inf. Sci.* Vol 34 (1982) pp 331–342

77 **Dillon, M, Ulmschneider, J and Desper, J** 'A prevalence formula for automatic relevance feedback in Boolean systems' *Inf. Process. & Manage.* Vol 19 (1983) pp 27–36

78 **Ide, E** 'New experiments in relevance feedback' in *The SMART retrieval system — experiments in automatic document processing* **G Salton (Ed.)** Prentice Hall, USA (1971) Chap. 16 [*See also* the chapter by **Ide and Salton** in this work (Chap. 18)]

79 **Pietilainen, P** 'Local feedback and intelligent automatic query expansion' *Inf. Process. & Manage.* Vol 19 (1983) pp 51–58

80 **Robertson, S E and Sparck Jones, K** 'Relevance weighting of search terms' *J. Am. Soc. Inf. Sci.* Vol 27 (1976) pp 129–146

81 **Rocchio, J J and Salton, G** 'Information search optimization and iterative retrieval techniques' *AFIPS Fall Joint Comput. Conf. Proc.* Vol 27 (1965) pp 293–305

82 **Salton, G, Fox, E A, Buckley, C and Voorhees, E** 'Boolean query formulation with relevance feedback' Technical Report TR 83-539, Department of Computer Science, Cornell University, USA (1983)

83 **Smeaton, A F** 'Relevance feedback and a fuzzy set of search terms in an information retrieval system' *Inf. Technol.* Vol 3 (1984) pp 15–24

84 **Yu, C T, Luk, W S and Cheung, T Y** 'A statistical model for relevance feedback in information and retrieval' *J. ACM* Vol 23 (1976) pp 273–286

85 **Cooper, W S and Maron, M E** 'Foundations of probabilistic and utility-theoretic indexing' *J. ACM* Vol 25 (1978) pp 67–80

86 **Parker, L M P** 'Towards a theory of document learning' *J. Am. Soc. Inf. Sci.* Vol 34 (1983) pp 16–21

87 **Tague, J M** 'User-responsive subject control in bibliographical retrieval systems' *Inf. Process. & Manage.* Vol 17 (1981) pp 149–159

88 **Morse, P M** 'Optimal linear ordering of information items' *Oper. Res. (USA)* Vol 20 (1972) pp 741–751

89 **Heine, M H** 'Measures of language effectiveness and the Swetsian hypotheses' *J. Doc. (GB)* Vol 31 (1975) pp 283–287

90 **Stirling, K H** 'The effect of document ranking on retrieval system performance: a search for an optimum ranking rule' *Proc. Am. Soc. Inf. Sci.* Vol 12 (1975) pp 105–106

91 **Robertson, S E** 'The probability ranking principle in IR' *J. Doc. (GB)* Vol 33 (1977) pp 294–304

92 **Maron, M E and Kuhns, J L** 'On relevance, probabilistic indexing and information retrieval' *J. ACM* Vol 7 (1960) pp 216–244

93 **Robertson, S E, van Rijsbergen, C J and Porter, M F** 'Probabilistic models of indexing and searching' in *Information retrieval research* **R N Oddy** *et al.* **(Eds)** Butterworths, UK (1980) pp 35–56

94 **van Rijsbergen, C J, Robertson, S E and Porter, M F** *New models in probabilistic information retrieval* Computer Laboratory, University of Cambridge, UK (Chap. 3) (1980)

95 **Salton, G, Buckley, C and Fox, E A** 'Automatic query formulations in information retrieval' Technical Report TR 82-524, Department of Computer Science, Cornell University, USA (1982)

96 **Maron, M E** 'On indexing, retrieval and the meaning of about' *J. Am. Soc. Inf. Sci.* Vol 28 (1977) pp 38–43

97 **Heine, M H** 'A simple, intelligent front end for information retrieval systems using Boolean logic' *Inf. Technol.* Vol 2 (1982) pp 247–260

98 **Taylor, R S** 'Question-negotiation and information seeking in libraries' *Coll. & Res. Libr. (USA)* Vol 29 (1968) pp 178–194

99 **Kochen, M and Badre, A N** 'Questions and shifts of representation in problem-solving' *Am. J. Psychol.* Vol 87 (1974) pp 369–383

100 **van Rijsbergen, C J** 'A theoretical basis for the use of co-occurrence data in information retrieval' *J. Doc. (GB)* Vol 33 (1977) pp 106–119

101 **Brookes, B C** 'The measures of information retrieval effectiveness proposed by Swets' *J. Doc. (GB)* Vol 24 (1968) pp 41–54

102 **Heine, M H** 'Incorporation of the age of a document into the retrieval process' *Inf. Process. & Manage.* Vol 13 (1977) pp 35–47

103 **Heine, M H** 'The inverse relationship of precision and recall in terms of the Swets model' *J. Doc. (GB)* Vol 29 (1973) pp 81–84

104 **Heine, M H** 'Design equations for retrieval systems based on the Swets model' *J. Am. Soc. Inf. Sci.* Vol 25 (1974) pp 183–198

105 **Bookstein, A** 'The anomalous behaviour of precision in the Swets model and its resolution' *J. Doc. (GB)* Vol 30 (1974) pp 374–380

106 **Bookstein, A** 'When the most "pertinent" document should not be retrieved — an analysis of the Swets model' *Inf. Process. & Manage.* Vol 13 (1977) pp 377–383

107 **Maron, M E and Hillman, D J** 'Two models for retrieval system design' *Am. Doc.* Vol 15 (1964) pp 217–225

108 **Uhlmann, W** 'Document specification and search strategy using basic intersections and the probability measure of sets' *Am. Doc.* Vol 19 (1968) pp 240–246

109 **Landry, B L** '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, USA (1971)

110 **Turski W M** 'On a model of information retrieval system design based on thesaurus' *Inf. Storage & Retr. (GB)* Vol 7 (1971) pp 89–94

111 **Pawlak, Z** 'Mathematical foundations of information retrieval' CC PAS Report no. 101, Computation Centre, Polish Academy of Sciences, Warsaw (1973)

112 **Bookstein, A and Swanson, D** 'Probabilistic models for automatic indexing' *J. Am. Soc. Inf. Sci.* Vol 25 (1974) pp 312–318

113 **Dabrowski, M** 'A general model of distribution of objects in information retrieval systems' *Inf. Syst. (GB)* Vol 1 (1975) pp 147–151

114 **Bookstein, A and Cooper, W S** 'A general mathematical model for information retrieval systems' *Libr. Q. (USA)* Vol 46 (1976) pp 153–167

115 **Bookstein, A** 'Outline of a general probabilistic retrieval model' *J. Doc. (GB)* Vol 39 (1983) pp 63–72

116 **Robertson, S E** 'Theories and models in information retrieval' *J. Doc. (GB)* Vol 33 (1977) pp 126–148