The Influence of Choice of Record Field on Retrieval Performance for Bibliographic Database

Heesop Kim
IT Information Center, Electronics & Telecommunications Research Institute, Daejon, Korea
E-mail: khs01701@etri.re.kr

This empirical study investigated the effect of choice of record field(s) on retrieval performance for a large operational bibliographic database. The query terms used in the study were identified algorithmically from each target set rather than chosen by third party relevance judges. This was done in four different ways: (1) controlled terms derived from index term frequency weights, (2) uncontrolled terms derived from index term frequency weights, (3) controlled terms derived from inverse document frequency weights, and (4) uncontrolled terms based on inverse document frequency weights. Six choices of record field were recognised. Using INSPEC terminology, these were: (1) 'Abstract' field, (2) 'Anywhere' (i.e., all fields), (3) the 'Descriptors' field, (4) the 'Identifiers' field, (5) 'Subject' (i.e., 'Descriptors' plus 'Identifiers'), and (6) 'Title'. The study was undertaken in an operational web-based IR environment using the INSPEC bibliographic database. The retrieval performances were evaluated using the D measure (bivariate in Recall and Precision). The main findings were: (1) there exist significant differences in search performance arising from choice of field, using 'mean performance measure' as the criterion statistic; (2) the rankings of field-choices for each of these performance measures is significantly sensitive to the choice of query; and (3) the optimal choice for the D-measure is the 'Title' field.

1. Introduction

In bibliographic retrieval systems, a good search strategy requires decisions about which query terms to use, which record fields to search against, and which search algorithm to use (e.g. which Boolean operators to use). These decisions may or may not be properly influenced by the choice of database being searched. In particular, users may need to pay careful attention to record field choice in optimising their search statements, and choose fields according to the search performance criterion they have in mind. Although the user's choice of one or more record fields on which to search is a fundamental search decision, little investigation in this field of IR system evaluation appears to have been undertaken.

This was the motivation for the present study which sought first to establish a convincing methodology for investigating search performance where 'choice of record field' is recognised as a variable, and secondly to obtain results by applying that methodology, under carefully described and controlled experimental conditions.

Further to the above general aim, the following specific research objectives were recognised: (1) To discover whether differences in search performance arising from different choices of record field(s) are significant; (2) To rank different choices of record field(s) for their effectiveness, for different search performance variables.

2. Overview of the Experimental Design

The research question: "How does the choice of field affect retrieval performance?" is reformulated more precisely as the following hypothesis-pair:

***H*$_0$**: No difference in retrieval performance exists among the choice of record field (e.g., the mean value of a performance measure is the same for all six search variants, i.e., ***H*$_0$**: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$),

***H*$_1$**: Differences in retrieval performance exist among the choice of record field (i.e., ***H*$_1$**: at least one $\mu \neq$ another μ)

where:

μ_1 is the mean of a chosen performance measure for the 'Abstract' field,

μ_2 is the mean of a chosen performance measure for the 'Anywhere' field choice,

μ_3 is the mean of a chosen performance measure for the 'Descriptor' field,

μ_4 is the mean of a chosen performance measure for the 'Identifier' field,

μ_5 is the mean of a chosen performance measure for the 'Subject' field choice, and

μ_6 is the mean of a chosen performance measure for the 'Title' field.

To test the hypothesis-pair defined in the previous section the experimental design adopted is outlined in **Figure 1**.

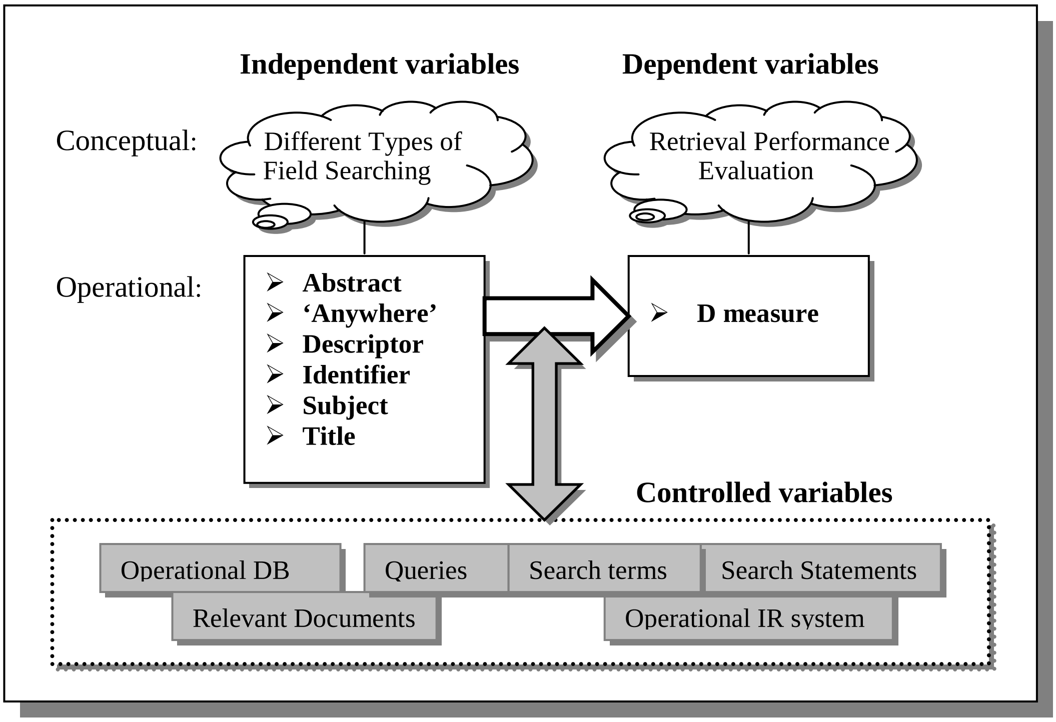


Figure 1. Overview of the Experiment Setting

2.2. Controlled variables

An Operational Database – INSPEC

The operational INSPEC database was used as the test database. At the time of testing, the database covered publications within the time span 1969 to February 1999. In August 1998, the IEE (The Institution of Electrical Engineers) announced that INSPEC then contained over 6 million bibliographic records, and that each year over 4,000 scientific and technical journals and some 2,000 conferences publications were being scanned; it was growing at a rate of 330,000 records each year with a weekly frequency of update (The Institution of Electrical Engineers, 1998b). The database is clearly a major disciplinary one, covering a major part of the World's literature in physics, computer and control engineering, and electrical and electronic engineering.

Relevant Documents and Target Sets

We defined a 'relevant document' as 'a cited reference in given review document.' A 'target set' was then defined as the set of relevant documents so defined and also available in the INSPEC database. In essence, the cognitive behaviour of end-users (the authors of review papers) was used as an operational definition of 'relevance'. This had the advantages over conventional approaches to defining 'relevant documents' by third party judges of: (1) resting the definition on a definite, given 'real-life' information need (namely the need to identify sources most appropriate to the information that the author of the review paper was reviewing), rather than an artificial need expressed solely in verbal terms and void of a real-life context of need, and (2) having such judgements made by persons who were reasonably expert and experienced in their field (on the assumption that reviewing authors are usually of such a character), thus minimising as much as seems reasonable to do the loss of potentially relevant papers not included in the review because of ignorance of them by its author (This method was used previously by Heine (1984) in the context of retrieval from Medline).

15 'review papers' recorded by INSPEC were generated in this manner using a restricted random sampling technique published between 1997 and 1998. Each of the 15 review papers was then regarded as a 'base-document of a target set'. These base-documents were then inspected in the SCI (Science Citation Index) through the U. K. BIDS (Bath Information and Data Services) Service to obtain the complete list of documents that they each cited (as an alternative to inspection in a library). Each of the documents cited by each base document so identified was then checked against the operational INSPEC database (via the Web) to identify its presence or non-presence in the database. If present, it contributed to the tally of relevant documents within the database for the appropriate base-document, i.e. it joined the 'set of relevant documents' for that base document. Documents cited by a base document but not included in the INSPEC database were deliberately eliminated to avoid subsequent errors in tallies of 'relevant documents not retrieved' in the experiment (The study chose to evaluate retrieval from INSPEC for document sets known to be within it, not to evaluate the exhaustivity of coverage of INSPEC itself).

Hence, the total number of relevant documents for each of the base-document was defined as the number of available documents in the INSPEC database among the cited documents by the author of the base-document. **Table** 1 shows the sizes of the 15 Target Sets so defined.

Table 1. Sizes of the Target Sets

Target Set ID	No of Cited Ref.	Size of Target Set	Percentage availability in database	Target Set ID	No of Cited Ref.	Size of Target Set	Percentage availability in database
#1	100	**24**	24.0 %	**#9**	128	**80**	62.5 %
#2	82	**56**	68.3 %	**#10**	76	**69**	90.8 %
#3	114	**37**	32.5 %	**#11**	54	**11**	20.4 %
#4	62	**39**	62.9 %	**#12**	79	**58**	73.4 %
#5	130	**110**	84.6 %	**#13**	90	**26**	28.9 %
#6	92	**51**	55.4 %	**#14**	101	**16**	15.8 %
#7	264	**134**	50.8 %	**#15**	116	**84**	72.4 %
#8	138	**68**	49.3 %	**TOTAL**	1625	**863**	53.1%

Queries

In our experiment, a query ('topic' in TREC terminology) was defined as 'a set of terms', rather than either a narrative sentence expressed in natural language serving as a relevance criterion (spuriously so, in our view), or a search expression (The IS&R literature unfortunately uses the term 'query' in several ways, so that prescriptive definition is necessary). The maximum size of queries was limited to four terms. The rationale for this centred on two considerations: (1) a recent survey result showed that most searchers use 2 to 4 query terms for an initial search (Kim, 1998), and (2) the reality of the 'combinatorial explosion' as a restraint on the analysis of data in this experiment using 'logical variety' as described below.

Differences in query type (i.e. in the 'type' of term defining the query) can affect the performance result (Soergel, 1985; Lancaster, 1998). The two most contrasting in character were chosen: (1) controlled terms (CT, henceforth), and (2) uncontrolled terms (UT, henceforth).

At the same time, we adopted two well-known weighting techniques to choose terms for the queries: (1) index term frequency (ITF, henceforth). [Note that the basic principle of ITF is the same as the well-known weighting technique *tf* (*term frequency*). However, they are differentiated insofar as *tf* refers to 'the frequency of term in full text documents of the collection', whereas ITF as use it here means 'the frequency of 'index term' in a target set'], and (2) inverse document frequency (IDF, henceforth) [1] (Note that the IDF varies inversely with the number of document 'n' to which a term is assigned in a collection of 'N' documents. In our case this refers to the entire INSPEC database].

Thus, the combination of two different types of index language (i.e., CT and UT) and two different types of weighting technique (i.e., ITF and IDF) generated queries of four distinct character:

[1] Also known as ICF (inverse collection frequency), which defined by Robertson, Walker and Hancock-Beaulieu (1995).

(1) CT_ITF – a query type made up of the controlled terms derived from the index term frequency weights; (2) UT_ITF - a query type made up of the uncontrolled terms derived from the index term frequency weights; (3) CT_IDF - a query type made up of the controlled terms derived from the inverse document frequency weights; and (4) UT_IDF - a query type made up of the uncontrolled terms derived from the inverse document frequency weights.

The 15 target sets were transmitted to the EINS (European Information Network Services) where the INSPEC database is also available via its online service. The reason for using the EINS rather than the Web-based INSPEC service, for this purpose, was their support for their ZOOM function. This gives a list of the most frequently occurring index terms in a target set, thus enabling the ITF function (as we have defined it) to be applied (Martin offers some relevant discussion (Martin, 1983)).

The well-known IDF (inverse document frequency) weight may perform to enhance precision (Salton and Yang, 1973; Salton and Burkely, 1988).

In the present experiment, we adopted Robertson and Sparck Jones's definition (Robertson and Sparck Jones, 1976) of IDF, in order to choose query terms. Using the notation 'N' for the number of documents in the INSPEC database; 'n' for the number of documents containing the search term in the database; 'R' for the total number of known relevant documents, known only within an experiment (that is the size of the target set), and 'r' for the number of relevant documents containing the search term in a target set, in one or other field or set of fields, the relevant weight is:.

$$w = \log \frac{\frac{(r+0.5)}{(R-r+0.5)}}{\frac{(n-r+0.5)}{(N-n-R+r+0.5)}} \qquad (1)$$

Search Statements – ELCs and Search Process

A 'search statement' is defined as a single character string, expressed in the formal query language of the search system, which activates a search of the database, that is, causes a search algorithm to scan the database and identify a set of hits.

In the present experiment this structure was based on: (1) Boolean connectives 'AND' and 'NOT', (2) specifications of one or more record fields to be searched against, i.e. (in our case), Abstract, 'Anywhere' (i.e., all fields), Descriptor, Identifier, Subject (i.e., Descriptor plus Identifier), and Title, (3) a range by publication dates of the coverage of the specific base document involved, and (4) employing exact matching rather than using some other kind of permissive syntax, i.e. we chose not to adopt such devices as role indicators, word adjacency, proximity, truncation, wildcard, etc (This was simply to bring the scope of the study within reasonable bounds).

Control of Logical variety in Search Statements

In order to free the generation of suitable search statements from arbitrariness in the choice of Boolean operators, it was decided to generate *all possible* logical forms of search statements. In this connection and with the condition of the search statement, we based the generation on ELCs (Elementary Logical Conjunctions) of each query's terms.

The four search terms that made up each query defined such ELCs as (for example):t_1 AND t_2 AND t_3 AND t_4; t_1 AND t_2 AND t_3 AND $\neg t_4$; t_1 AND t_2 AND $\neg t_3$ AND t_4; t_1 AND $\neg t_2$ AND t_3 AND t_4, and $\neg t_1$ AND t_2 AND t_3 AND $t_{4,}$ and so on. The symbol '¬' denotes 'AND NOT', usually written 'NOT'. Since each query term can be negated or not, there are 2^4 such ELCs for each query.

It follows from a familiar result of formal logic that these ELCs determine a *partitioning* of the database (and hence also of the chosen Target Set) into 16 different and non-overlapping (i.e. 'disjoint') subsets. (One or more of these subsets may be empty, of course.) The usefulness of this fact is that combinations ('disjunctions') of these ELCs taken one at a time, two at a time, three at a time, etc, then generate all possible logical expressions that could employ the four query terms. In the experiment, each query ELC (with one exception, see below) was presented to the INSPEC database, the INSPEC host software serving to record the number of records in it that evaluated that ELC to 'true' (i.e., more informally, 'were posted to it'. For some fuller discussion, see Heine, 1984; 2000a). However, the all-negated ELC (e.g., E_ab16 shaded row in **Table 2**) was excepted, since (as with any all-negated ELC) it retrieves almost all records of the database. Accordingly, only 15 (i.e., $2^4 - 1$) rather than 16 (i.e., 2^4) ELCs were so presented. The presentation of the 15 ELCs to the INSPEC database was done for each query and each choice of record field(s).

Table 2. A Sample of ELCs

Label	ELCs						
	Abstract Field (ab)						
E_ab01	t_1	AND	t_2	AND	t_3	AND	t_4
E_ab02	t_1	AND	t_2	AND	t_3	AND	$\neg t_4$
E_ab03	t_1	AND	t_2	AND	$\neg t_3$	AND	t_4
E_ab04	t_1	AND	$\neg t_2$	AND	t_3	AND	t_4
E_ab05	$\neg t_1$	AND	t_2	AND	t_3	AND	t_4
E_ab06	$\neg t_1$	AND	$\neg t_2$	AND	t_3	AND	t_4
E_ab07	$\neg t_1$	AND	t_2	AND	$\neg t_3$	AND	t_4
E_ab08	$\neg t_1$	AND	t_2	AND	t_3	AND	$\neg t_4$
E_ab09	t_1	AND	$\neg t_2$	AND	$\neg t_3$	AND	t_4
E_ab10	t_1	AND	$\neg t_2$	AND	t_3	AND	$\neg t_4$
E_ab11	t_1	AND	t_2	AND	$\neg t_3$	AND	$\neg t_4$
E_ab12	t_1	AND	$\neg t_2$	AND	$\neg t_3$	AND	$\neg t_4$
E_ab13	$\neg t_1$	AND	$\neg t_2$	AND	$\neg t_3$	AND	t_4
E_ab14	$\neg t_1$	AND	$\neg t_2$	AND	t_3	AND	$\neg t_4$
E_ab15	$\neg t_1$	AND	t_2	AND	$\neg t_3$	AND	$\neg t_4$
E_ab16	$\neg t_1$	AND	$\neg t_2$	AND	$\neg t_3$	AND	$\neg t_4$

Search Processes

As mentioned, all ELCs (except E_ab16) were presented to the INSPEC database for the six chosen record fields once a specific query had been chosen. **Figure 2** illustrates the procedures of constructing the ELCs and simulating the field searches.

As show in this figure, for a particular four-term query, all possible search expressions were generated from a given ELC set, by disjoining ELCs taken one at a time, two at a time, three at a time, etc, up to fifteen at a time. This generated 32,767 (i.e., 2^{15}-1 = $2^{(2^4-1)} - 1$) different searches for each query. This ensured that all possible search expressions were used, i.e. the experiment suppressed one source of experimenter arbitration (We note that, of course, additional research in a different experiment on searcher's cognitive behaviour might helpfully restrict the set of search expressions that might be used, but in view of the lack of any convincing and relevant cognitive model that was seen as lying outside the scope of the present study, i.e. we preferred not to make assumptions as to the selection of search expression grammars that users might make in practice).

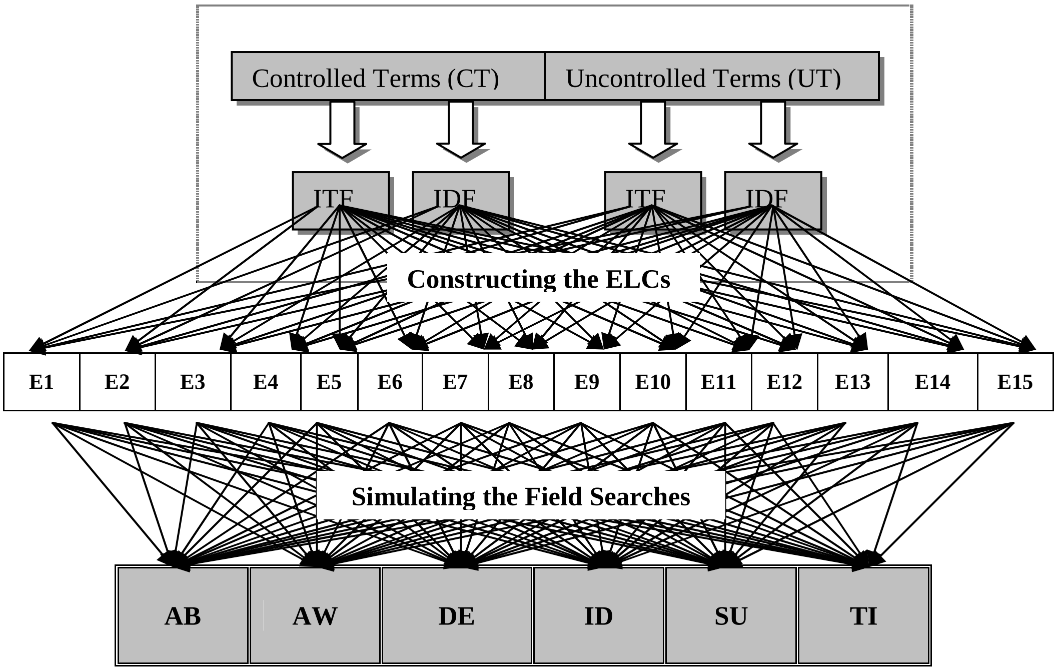


Figure 2. Construction of the ELC and Simulation of the Field Searches

An Operational IR system

For applying experimental-derived search statement to the INSPEC, an operational WebSPIRS™ (Web-based SilverPlatter®'s Information Retrieval System) IR system was used which was designed and maintained by SilverPlatter®. Permission to use this was granted by the IEE.

2.3. Independent variables – Field Searches

Independent variables are those that are manipulated or controlled by the experimenter. In this study, the main independent variables are choices from six INSPEC record fields as shown in **Table 3**.

Table 3. Independent variables - Six field Searches

Fields	Searches
Abstract (AB)	The AB field contains a summary of the document cited in the record
'Anywhere' (AW)	It does not refer to a field, but this 'words anywhere' option searches against the free text fields in the INSPEC, including the other five fields named here. The field label is not necessary in the main search screen as a default option in WebSPIRS™ system, but we labelled as 'AW' for a simplification and unification reason with the other chosen five fields throughout this thesis.
Descriptors (DE)	The DE field contains standard (or preferred) term from the INSPEC thesaurus. Terms may be single words or hyphenated phrases. The 1999 edition of the INSPEC thesaurus contains approximately 16,000 terms of which some 8,300 terms are preferred terms (IEE, 1999b). (= *Controlled Terms*).
Identifiers (ID)	The ID field contains free language words and phrases assigned by the human INSPEC index experts. They give a more exhaustive description of the content of the document than that which is provided by the original title or by the DE field. (= *Uncontrolled Terms*).
'Subject' (SU)	The SU terms allows searching both the DE field and ID field at the same time.
Title (TI)	The TI field contains the title of the record, exactly as it appeared in the original publication

2.4. Dependent variables – Performance measures

There are well-known single-valued variables that attempt to characterise overall retrieval performance. Amongst these are: Brookes's S measure (1968), van Rijsbergen's E (β) measure (van Rijsbergen, 1979), Shaw's "harmonic mean" F (Shaw, 1986), and Heine's D (Heine, 1973). Van Rijsbergen's measure is widely used and is usefully parametric in a quantity, β, which reflects a degree of preference by the user towards a search result that is oriented towards either R or P. However, D was chosen in the present experiment since variability in user preference of this nature was excluded from the experimental design, so that β was superfluous (It also has the feature of being a metric, which may be useful in subsequent theoretical work). In addition, constituent Recall and Precision values were noted throughout.

The major features (arguably, advantages) of all these single-valued measures are: (1) they reflect retrieval effectiveness solely in terms of database subsets, i.e. they are independent of criteria such as cost of retrieval, value of information retrieved, or speed of searching, (2) they are independent of the number of documents retrieved in a particular search, i.e. they are probabilistic in nature (with the exception of Brookes's S), and (3) they express retrieval effectiveness as a single numerical value instead of two values such as Recall and Precision taken as a pair, thus allowing searches to be more easily compared on arithmetic scales.

The definition of the D measure is shown in **Table 4**. Evaluation of retrieval performance of a given set of user queries with respect to a document collection is, by convention, based on a two by two contingency table which distinguishes between the documents retrieved in answer to a given information need (confusingly referred to as a ‘query’) and those not retrieved, and between documents judged relevant to the information need and those not relevant.

Table 4. Retrieval Performance Evaluation Measures

(a) The 2-by-2 contingency table of relevant and retrieval (Swets, 1963. p.246)

	Relevant	Not Relevant	
Retrieved	a	b	a + b
Not Retrieved	c	d	c + d
	a + c	b + d	a + b + c + d

(b) The performance measures used in the present study – D measure

Symbol	Evaluation Measure	Formula	Explanation
D	Single measure	$1 - \left(\dfrac{1}{\dfrac{1}{\frac{a}{a+c}} + \dfrac{1}{\frac{a}{a+b}} - 1} \right),$ when $\dfrac{a}{a+c} \neq 0, \dfrac{a}{a+b} \neq 0$ (2)	The lower the D value, the better IR performance, since D is a measure of the ‘distance’ between the set of relevant documents and the set of retrieved documents.

3. Data Collection

3.1. Searching using the six choice of fields

The form of a search statement ELC for E_ab13 using CT_ITF query in Target Set #5 is, for example, as follows:

Find: (DIELECTRIC MEASUREMENT in **ab**) **NOT** (MICROWAVE MEASUREMENT ***in*** **ab**) **NOT** (PERMITTIVITY MEASUREMENT ***in*** **ab**) **NOT** (COAXIAL CABLES ***in*** **ab**) **AND** (**PY**=1969-1997)

Each of the six field searches was performed for each Target Set and each type of query. (The reader is reminded that we use the term ‘query’ to stand for a set of search terms that is variable against a fixed instance of information need expressed through citation behaviour, and not as a verbal search criterion.) All the results were noted on an ELC search result sheet, and the process of searching was repeated with the same fashion for the 15 Target Sets, a total of 5,400 searches. That is 15 (number of the Target Sets) x 4 (number of the query types) x 6 (number of choices of fields) x 15 (number of

ELCs per query). The summary of each query type for this particular Target Set is presented in **Table 5(a)** including the following information: (i) a Target Set reference number ('ID'), (ii) the four search terms used, (iii) the choice of fields used in the search, (iv) the range of publication dates particular to this Target Set, and (v) the size of the target set (i.e., '**a + c**') which was pre-identified. The size of each set of retrieved documents, i.e., the search result (i.e., '**a + b**'), and the size of retrieved relevant documents set (i.e., '**a**'), are presented in **Table 5(b)** for each ELC derived from this query. **Table 5** presents the sample of the search results for the query type CT_ITF in 'Abstract' field for Target Set #5.

Table 5. ELC Searches – 'Abstract' field for CT_ITF query type

(a) Basic information

Target Set ID	**#05**	
Query Type - CT_ITF	(t_1) (t_2) (t_3) (t_4)	MICROWAVE MEASUREMENT PERMITTIVITY MEASUREMENT COAXIAL CABLES DIELECTRIC MEASUREMENT
Publication data Coverage	1969 – 1997	
Total number of relevant documents (a + c)	110	

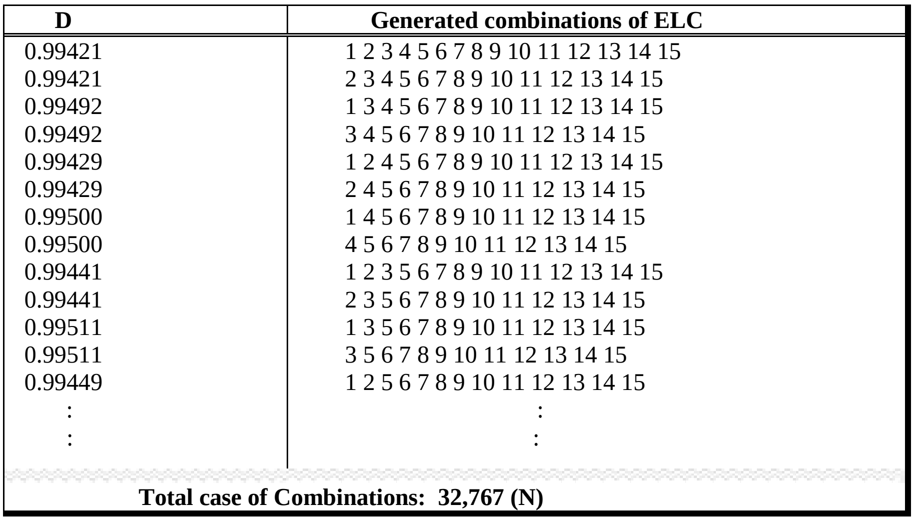
(b) Search results

Label	**ELCs**							No of Retrieved Relevant Doc. **(a)**	No of Retrieved Doc. **(a + b)**
E_ab01	t_1	AND	t_2	AND	t_3	AND	t_4	0	0
E_ab02	t_1	AND	t_2	AND	t_3	AND	$\neg t_4$	0	0
E_ab03	t_1	AND	t_2	AND	$\neg t_3$	AND	t_4	0	0
E_ab04	t_1	AND	$\neg t_2$	AND	t_3	AND	t_4	0	0
E_ab05	$\neg t_1$	AND	t_2	AND	t_3	AND	t_4	0	0
E_ab06	$\neg t_1$	AND	$\neg t_2$	AND	t_3	AND	t_4	0	0
E_ab07	$\neg t_1$	AND	t_2	AND	$\neg t_3$	AND	t_4	0	1
E_ab08	$\neg t_1$	AND	t_2	AND	t_3	AND	$\neg t_4$	0	0
E_ab09	t_1	AND	$\neg t_2$	AND	$\neg t_3$	AND	t_4	0	3
E_ab10	t_1	AND	$\neg t_2$	AND	t_3	AND	$\neg t_4$	0	1
E_ab11	t_1	AND	t_2	AND	$\neg t_3$	AND	$\neg t_4$	0	1
E_ab12	t_1	AND	$\neg t_2$	AND	$\neg t_3$	AND	$\neg t_4$	0	219
E_ab13	$\neg t_1$	AND	$\neg t_2$	AND	$\neg t_3$	AND	t_4	6	136
E_ab14	$\neg t_1$	AND	$\neg t_2$	AND	t_3	AND	$\neg t_4$	0	989
E_ab15	$\neg t_1$	AND	t_2	AND	$\neg t_3$	AND	$\neg t_4$	4	55
E_ab16	$\neg t_1$	AND	$\neg t_2$	AND	$\neg t_3$	AND	$\neg t_4$	Excepted	Excepted

3.2. Calculating D Values

To evaluate D (i.e., the dependent variable in this experiment), the 15 forms of ELC were combinatorially disjoined using a small C program so to generate all possible logical form of search statement, doing so for each triple "Target Set, query, and choice of record fields". In case of either $\frac{a}{a+c}$ =0 or $\frac{a}{a+b}$ =0, the D value was arbitrarily assigned as '1' which indicates the worst performance result (since D is a distance measure). This value was coded to the system-missing value as '1' in SPSS™.

A sample result of D values and the combination are shown in **Figure 3**.



D	Generated combinations of ELC
0.99421	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.99421	2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.99492	1 3 4 5 6 7 8 9 10 11 12 13 14 15
0.99492	3 4 5 6 7 8 9 10 11 12 13 14 15
0.99429	1 2 4 5 6 7 8 9 10 11 12 13 14 15
0.99429	2 4 5 6 7 8 9 10 11 12 13 14 15
0.99500	1 4 5 6 7 8 9 10 11 12 13 14 15
0.99500	4 5 6 7 8 9 10 11 12 13 14 15
0.99441	1 2 3 5 6 7 8 9 10 11 12 13 14 15
0.99441	2 3 5 6 7 8 9 10 11 12 13 14 15
0.99511	1 3 5 6 7 8 9 10 11 12 13 14 15
0.99511	3 5 6 7 8 9 10 11 12 13 14 15
0.99449	1 2 5 6 7 8 9 10 11 12 13 14 15
:	:
:	:

Total case of Combinations: 32,767 (N)

Figure 3. Result of D values and generated combinations of ELCs

4. Results and Discussion

4.1. Statistical Analysis

Two statistical analyses were conducted: (1) exploration and description of the performance measures, and (2) test of hypothesis. For the first gradation, several summary statistics were examined using the Descriptive procedure. The descriptive is the principal procedures for describing and exploring interval data, and provides a quick way of obtaining a range of common descriptive statistics, both of tendency and of dispersion.

The descriptive statistics was presented including such as: (1) N (i.e., number of cases – 32,767), (2) Mean (i.e., the arithmetic averages), (3) Standard Deviation (i.e., a measure of how much observations vary from the mean, expressed in the same units as the data), (4) Standard Error (i.e., a measure of variability), (5) 95% confidence interval for the mean with lower bound and upper bound, (6) Minimum (i.e., the smallest value), and (7) Maximum (i.e., the largest value). See **Table 6** for an example of the descriptive statistics of D measure for CT_ITF in Target Set #5.

Table 6. Descriptive statistics of D measure for CT_ITF – Target Set #5

	N	MEAN	STD. DEVIATION	STD. ERROR	95% CONFIDENCE INTERVAL FOR MEAN		MIN	MAX
					Lower Bound	Upper Bound		
AB	32767	.99026	1.0209E-02	5.6401E-05	.99015	.99037	.966	1.000
AW	32767	.99055	1.0404E-02	5.7476E-05	.99044	.99066	.855	1.000
DE	32767	.98994	1.0856E-02	5.9973E-05	.98982	.99006	.852	1.000
ID	32767	.98943	5.9217E-03	3.2714E-05	.98936	.98949	.974	1.000
SU	32767	.98997	1.1073E-02	6.1170E-05	.98985	.99009	.838	.999
TI	***32767***	***.98856***	***1.0146E-02***	***5.6051E-05***	***.98845***	***.98867***	***.965***	***1.000***
Total	196602	.98979	9.9450E-03	2.2429E-05	.98974	.98983	.838	1.000

As an auxiliary for the descriptive statistics, the Percentiles (i.e., values that divide cases according to values below which certain percentages of cases fall) graphs were produced to facilitate a visualised comparison between the variables. The values for the 5th 10th 25th 50th 75th 90th 95th percentiles were displayed in graphs for each case of the test result (See **Figure 4** for an example).

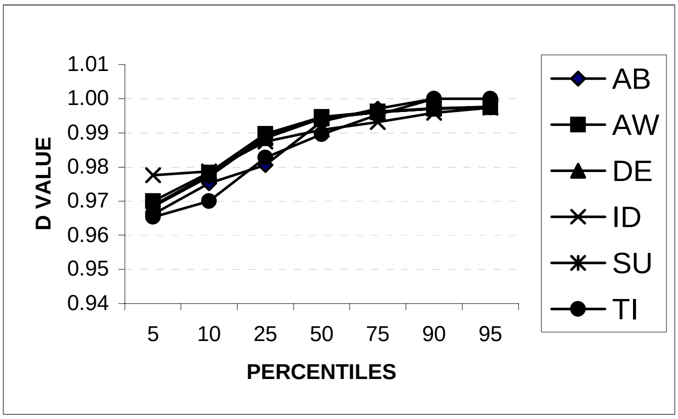


Figure 4. Percentile of D measure for CT_ITF query type – Target Set #5

In the second gradation, One-way ANOVA was used to test the null hypothesis. Analysis of variance, or ANOVA, is a method of testing the null hypothesis that several group means are equal in the population, by comparing the sample variance estimated from the group means to that estimated within the groups. In this study, One-Way ANOVA was used to test the hypothesis that several means are equal. This technique is an extension of the two-sample t-test. The ANOVA F statistic is calculated by dividing an estimate of the variability between groups by the within groups' variability: F = (variance between) / (variance within). See **Table 7** for an example of ANOVA of D measure for CT_ITF query in Target Set#5.

Table 7. ANOVA of D measure for CT_ITF query type – Target Set #5

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.132E-02	5	1.626E-02	165.125	.000
Within Groups	19.363	196596	9.849E-05		
Total	19.444	196601			

In this connection, once we have determined that differences exist among the means, LSD (least significant difference) in "post hoc" tests for pair-wise multiple comparisons were used to determine which means differ. Pair-wise multiple comparisons test the difference between each pair of means, and yield a matrix where asterisks (*) indicate significantly different group means at an alpha level of 0.01 in this study.

Table 8. LSD Multiple Comparisons of D measure for CT_ITF – Target Set #5

		Mean Difference	Std.	Sig.	99% Confidence Interval	
(I) FIELD TYPES	(J) FIELD TYPES	(I-J)	Error		Lower Bound	Upper Bound
AB	AW	-2.90084E-04(*)	.000	.000	-4.89803E-04	-9.03657E-05
	DE	3.1885E-04(*)	.000	.000	1.1913E-04	5.1857E-04
	ID	8.2929E-04(*)	.000	.000	6.2957E-04	1.0290E-03
	SU	2.8438E-04(*)	.000	.000	8.4658E-05	4.8410E-04

	TI	1.6934E-03(*)	.000	.000	1.4937E-03	1.8931E-03
AW	AB	2.9008E-04(*)	.000	.000	9.0366E-05	4.8980E-04
	DE	6.0893E-04(*)	.000	.000	4.0921E-04	8.0865E-04
	ID	1.1194E-03(*)	.000	.000	9.1966E-04	1.3191E-03
	SU	5.7446E-04(*)	.000	.000	3.7474E-04	7.7418E-04
	TI	1.9835E-03(*)	.000	.000	1.7838E-03	2.1832E-03
DE	AB	-3.18847E-04(*)	.000	.000	-5.18566E-04	-1.19128E-04
	AW	-6.08931E-04(*)	.000	.000	-8.08650E-04	-4.09213E-04
	ID	5.1044E-04(*)	.000	.000	3.1073E-04	7.1016E-04
	SU	-3.44702E-05	.000	.657	-2.34189E-04	1.6525E-04
	TI	1.3746E-03(*)	.000	.000	1.1748E-03	1.5743E-03
ID	AB	-8.29291E-04(*)	.000	.000	-1.02901E-03	-6.29572E-04
	AW	-1.11938E-03(*)	.000	.000	-1.31909E-03	-9.19657E-04
	DE	-5.10444E-04(*)	.000	.000	-7.10163E-04	-3.10725E-04
	SU	-5.44914E-04(*)	.000	.000	-7.44633E-04	-3.45196E-04
	TI	8.6411E-04(*)	.000	.000	6.6439E-04	1.0638E-03
SU	AB	-2.84377E-04(*)	.000	.000	-4.84095E-04	-8.46580E-05
	AW	-5.74461E-04(*)	.000	.000	-7.74180E-04	-3.74742E-04
	DE	3.4470E-05	.000	.657	-1.65248E-04	2.3419E-04
	ID	5.4491E-04(*)	.000	.000	3.4520E-04	7.4463E-04
	TI	1.4090E-03(*)	.000	.000	1.2093E-03	1.6087E-03
TI	AB	-1.69340E-03(*)	.000	.000	-1.89312E-03	-1.49368E-03
	AW	-1.98349E-03(*)	.000	.000	-2.18321E-03	-1.78377E-03
	DE	-1.37456E-03(*)	.000	.000	-1.57427E-03	-1.17484E-03
	ID	-8.64111E-04(*)	.000	.000	-1.06383E-03	-6.64393E-04
	SU	-1.40903E-03(*)	.000	.000	-1.60874E-03	-1.20931E-03
* The mean difference is significant at the 0.01 level						

4.2. Integrated Data Analysis for the Overall 15 Target Sets

Descriptive analysis

To identify the best field search result in D measure, all the individual 15 Target Sets were considered for the case of the **best mean** value of the 32767 ELC results for a given query type and choice of field. The same method, which used in the Recall and Precision, was applied to this analysis. The results were categorised based on the four query types in **Table 9**. The comparison results were shown in **Figure 5**.

Table 9. Occurrences of the best D measure performance from the overall 15 Target Sets

	CT_ITF		**CT_IDF**		**UT_ITF**		**UT_IDF**		**Total**	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Abstract (AB)	0.00	0.00	0.00	0.00	0.00	0.00	5.00	33.33	**5.00**	**8.33**
'Anywhere' (AW)	1.00	6.67	3.00	20.00	1.00	6.67	2.00	13.33	**7.00**	**11.67**
Descriptors (DE)	3.00	20.00	5.50	36.67	0.00	0.00	0.00	0.00	**8.50**	**14.17**
Identifiers (ID)	3.00	20.00	2.00	13.33	5.00	33.33	2.00	13.33	**12.00**	**20.00**
Subject (SU)	1.00	6.67	0.50	3.33	0.00	0.00	1.00	6.67	**2.50**	**4.17**
Title (TI)	7.00	46.66	4.00	26.67	9.00	60.00	5.00	33.33	**25.00**	**41.67**
Total ***(Valid)***	**15.00** *(15)*	**100.0** *(100)*	**14.98** *(15)*	**100.0** *(100)*	**14.99** *(15)*	**100.0** *(100)*	**14.98** *(15)*	**100.0** *(100)*	**60.00** *(60)*	**100.01** *(100)*

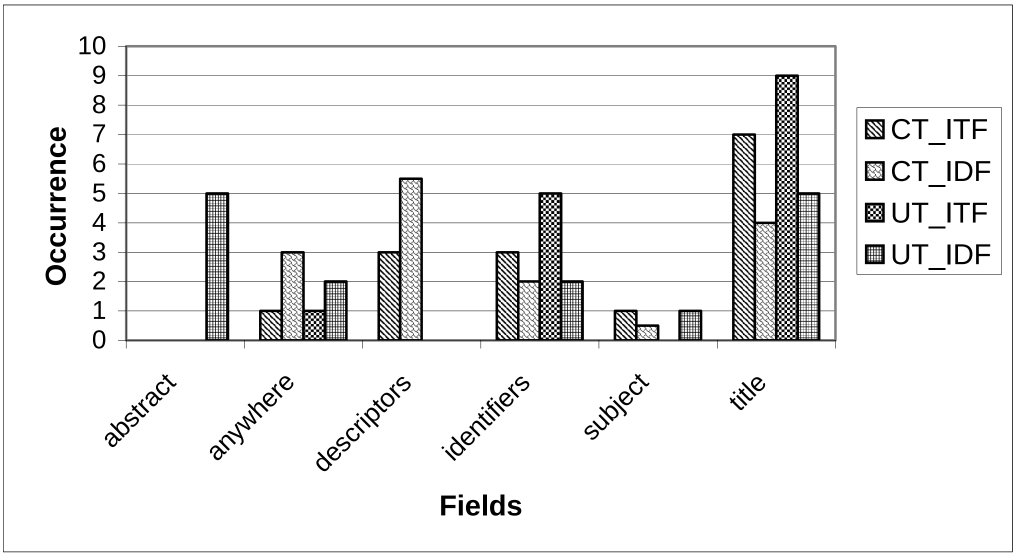


Figure 5. Comparison of the Occurrences of the best D measure Performance from the overall 15 Target Sets

‘Title’ field search gave the best performance among the six chosen field searches in terms of the occurrence of the best mean value in D measure performance. The outcomes in descending order for the query types were as follows: (1) UT_ITF (9 occurrences – 60.0%), (2) CT_ITF (7 occurrences – 46.7%), (3) UT_IDF (5 occurrences – 33.3%) (4) CT_IDF (4 occurrences – 26.7%), respectively. Not surprisingly, the uncontrolled terms were dominated since ‘Title’ field is one of the uncontrolled indexing fields in the INSPEC. The overall result implies that ‘Title’ field search can be considered as well-balanced field both for its specificity and its exhaustivity when compare with other chosen field in this study

‘Identifiers’ field search gave the second-best in D measure performance. The outcomes in descending order for the query types were as follows: (1) UT_ITF (5 occurrences – 33.3%), (2) CT_ITF (3 occurrences – 20.0%), (3) UT_IDF (2 occurrences – 13.3%), and (4) CT_IDF (2 occurrences – 13.3%), accordingly. The main reason for UT_ITF query type’s domination may cause since ‘Identifiers’ field is assigned with the uncontrolled terms in the INSPEC.

‘Descriptors’ field search gave the third-best result in D measure performance among the six chosen field searches. The outcomes for the query types were as follows: (1) CT_IDF (6 occurrences – 40.0%), and (2) CT_ITF (3 occurrences – 20.0%). It clearly shows that the controlled terms (i.e., CT_ITF and CT_IDF) perform better than the uncontrolled terms (i.e., UT_ITF and UT_IDF) for ‘Descriptors’ field choice. It may because the field is assigned with the uncontrolled index terms in the INSPEC.

‘Anywhere’ (i.e., all fields) search gave the fourth-best result in D measure performance. The outcomes for the query types were as follows: (1) UT_IDF (2 occurrences – 13.3%), (2) CT_IDF (2 occurrences – 13.3%) (3) CT_ITF (1 occurrence – 6.7%), and (4) UT_ITF (1 occurrence – 6.7%), respectively.

As already discussed, this multiple field choice provides a high exhaustivity but not a high enough specificity. Although the choice of 'Anywhere' fields may cause a dilemma, but there s evident that it provides an advantage in the initial stage.

'Abstract' field search gave the fifth-best result in D measure performance among the six chosen field searches. The outcome for the best mean value was found in a specific query type for UT_IDF (5 occurrences – 33.3%), exclusively. Although the overall performance was disappointing, it may be more suitable for the uncontrolled terms than for the controlled terms. 'Subject' (i.e., 'Descriptors' plus 'Identifiers') field search gave the worst result in D measure performance in this study. The outcome for the query types were as follows: (1) CT_ITF (1 occurrence – 6.7%), (2) CT_IDF (1 occurrence – 6.7%), and (3) UT_IDF (1 occurrence – 6.7%), respectively.

Overall, the D performance is indicated on this evidence to be given in this study in descending order: 'Title' > 'Identifiers' > 'Descriptors' > 'Anywhere' > 'Abstract' > 'Subject'.

AVOVA analysis

The inferential statistical test, ANOVA at a level of significance 0.05 was carried out to test between the hypotheses:

- $\boldsymbol{H_0}$: No differences exist between the mean-D values for the six choices of field(s).
- $\boldsymbol{H_1}$: Differences exist between the mean-D values for the six choices of field(s).

All the results of D measure from the 15 Target Sets showed that the p-value (i.e., significance value) was less than 0.05. Accordingly, the null hypothesis $\boldsymbol{H_0}$ was rejected and the alternative hypothesis $\boldsymbol{H_1}$ accepted. The result was thus significant beyond the 95% level.

5. Conclusions and Further Research

This study produced a large mass of data and it proved not to be easy to distinguish that which is significant from that which is of little novelty and/or value. However, obtaining such a large quantity of data was regarded as necessary if the study was: (1) to go beyond the 'construction of demonstrator' stage, and (2) to generate and evaluate hypotheses that had a validity for at least one database, i.e. to try to avoid the traps of obtaining anecdotal data with little transferable value and of doubtful statistical significance.

A number of conclusions have been drawn, principally expressed in terms of the D single measure. In stating these conclusions, the author emphasises that, in common with most, of not all, experiments in information storage and retrieval, their more general validity is contingent on the validity of the experimental design employed. We have attempted to justify the different aspects of this design and also the choice of the INSPEC database as a test vehicle given its status as a modern, fully operational, bibliographical database having a worldwide scope

First, the overall D measure results of the comparative evaluation with the performance of the chosen six field searches may be confirmed as suggesting that the following order of preference regarding choices of record field: TI > ID > DE > AW > AB > SU for data aggregated for all query types. In other words, an end-user is best advised, when searching INSPEC, to choose the 'Title' field, with 'ID' as second preference, and so on. (This is assuming that the end-user's interests are equally balanced as between Recall and Precision.)

Secondly, for data separated for each query type, this ranking of field-choice in order of preference becomes: TI > DE > ID > AW = SU > AB for the CT_ITF query type; DE > TI > AW > ID > SU > AB for the CT_IDF query type; TI > ID > AW >AB = DE = SU for the UT_ITF query type; and AB = TI >AW = ID > SU > DE for the UT_IDF query type. The variability of these 'preference orders' reflects sensitivity in retrieval effectiveness as measured by D to the searcher's choice of query terms - assuming throughout that four-term queries are defined, and also that users' choices of query terms are consonant with the optimal query terms used in the present study.

For the present, the research findings give a number of ideas to be pursued in the future research.

First, there appears to be a major need for a study that employs queries defined by real users in real-problem contexts since this experiment adopted an algorithmic method to produce the queries although, that said, users' search behaviour should arguably be viewed as capable of benign influence by experimental results such as those we have obtained in this study.

Secondly, further studies being considered might investigate the combination of search field rather than isolation of search field, although it would involve a more complicated experimental design. (We attempted this to only a limited extent, with our use of the 'Anywhere' choice.)

Thirdly, because this study has provided a rich set of data, it would be worth focusing further analyses of the data on the indexing languages, for example, controlled index terms versus uncontrolled index terms and then compare the results with many earlier studies (Lancaster, 1986; Shaw, 1994; Boyce and McLain, 1989; Muddamalle, 1998; Voorbij, 1998). For example, Muddamalle (1998) concludes that the best performance could be achieved by the two in combination.

Fourthly, although the present study has, in a sense, 'maintained faith' with the conventions of using probabilistic measures to characterise retrieval performance, there are strong criticisms, rehearsed for many years but arguably not yet responded to with any vigour by the retrieval community, of such measures, and in particular the Recall measure. (For a recent review, see Heine (2000b). Constructs more relevant to the assessment of retrieval undertaken in a 'user-learning' environment, i.e. those recognising heurism, are badly needed, one tentative example based on logic vectors being put forward by Heine (2000b).

Finally, follow-up studies that adopted the methodology used this study might be valuable in the Internet environment, since in Web search engines, field searching (e.g. on 'Dublin Core' HTML fields) offers the same advantages as in traditional online bibliographic databases. However, partly because of the newness of Web search engines and partly because of the unique nature of Web resources, the options are limited (Clarke, 2000; Hattery, 1997; Notess, 1996, 1997; Hock, 1998; Vidmar, 1999; Webber, 1998). On the other hand, it seems clear that a robust and consensual evaluation methodology for the IR performance in the Internet is still required.

Acknowledgements

The author would like to thank Dr M. H. Heine, School of Information Studies, and Dr D. Gee, School of Computing, at the University of Northumbria at Newcastle, UK, for their feedback and supervision throughout this study.

References

Aitchison, T.M. and Tracy, J. M. (1969). *INSPEC: Comparative Evaluation of Index Languages - Part 1: Design*. London: Institution of Electrical Engineers. (Report no.: INSPEC/4).

Aitchison, T.M. et al. (1970). *INSPEC: Comparative Evaluation of Index Languages - Part II: Results*. London: Institution of Electrical Engineers. (Report no.: INSPEC/5).

Boyce, B. R. and McLain, J.P. (1989). "Entry Point Depth and Online Search Using a Controlled Vocabulary," *Journal of the American Society for Information Science* **40**(4): 273-276.

Brookes, B.C. (1968). "The Measure of Information Retrieval Effectiveness Proposed by Swets," *Journal of Documentation* **24**(1): 41-54.

Clarke, S.J. (2000). "Search Engines for the World Wide Web: An Evaluation of Recent Developments," *Journal of Internet Cataloging* **2**(3-4): 81-93.

Hattery, M. (1997). "Online World: the Bumpy Ride of the Web Engine," *Information Retrieval & Library Automation* **33**(5): 1-2.

Heine, M.H. (2000a). "Describing Query Expansion using Logic-induced Vectors of Performance Measures," paper presented in *SIGIR 2000*, Athens, July 2000.

Heine, M.H. (2000b). "Reassessing and Extending the Precision and Recall Concepts," *In* www.ewic.org.uk/ewic. Revised version of "Time to dump 'P and R'?" *Proceedings of the MIRA '99*: Final MIRA Conference on Information Retrieval Evaluation, Glasgow, 14-16 April 1999: 61-74.

Heine, M.H. (1984). "Information Retrieval from Classical Databases from a Signal-Detection Standpoint: A Review," *Information Technology: Research and Development* **3**(2): 95-112.

Heine, M.H. (1978). "The Signal-Detection Model of Information Retrieval," *Journal of Informatics* **2**(1): 26-33.

Heine, M.H. (1973). "Distance Between Sets as an Objective Measure of Retrieval Effectiveness," *Information Storage and Retrieval* **9**(3): 181-198.

Hock, R.E. (1998). "How to Do Field Searching in Web Search Engines: A Field Trip," *Online* **22**(3): 18-22

The Institution of Electrical Engineers (1999a). *Classification: A Classification Scheme for the INSPEC Database*. London: IEE Publishing and Information Services.

The Institution of Electrical Engineers (1999b). *Thesaurus: 1999*. Surrey, England: The Gresham Press.

The Institution of Electrical Engineers (1998a). *INSPEC Database on WebSPIRS: User Notes*. London: IEE Publishing and Information Services.

(Also available at www.iee.org.uk/publish/inspec/venders/splatter.html)

The Institution of Electrical Engineers (1998b). *INSPECMATTERS: the Newsletter of the IEE Publishing and Information Services Division*. Stevenage: IEE Publishing and Information Services.

(Also available at www.iee.org.uk/publish/inspec/)

Kim, H. (1998). *User Differences in Interactive Web-based OPAC Evaluation. Unpublished MPhil thesis,* Department of Information Studies, University of Sheffield, UK.

Lancaster, F.W. (1986). *Vocabulary Control for Information Retrieval, 2nd Edition*. Arlington, VA: Information Resources.

Lancaster, F.W. (1998). *Indexing and Abstracting in Theory and Practice, 2nd Edition*. London: Library Association Publishing.

Martin, W.A. (1983). "Methods for Evaluating the Number of Relevant Documents in a Collection," *Journal of Information Science* **6**(3): 173-177.

Muddamalle, M.R. (1998). "Natural Language versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics," *Journal of the American Society for Information Science* **49**(10): 881-887.

Notess, G.R. (1997). "Internet Search Techniques and Strategies," *Online* **21**(4): 63-66.

Notess, G.R. (1996). "Searching the Web with Alta Vista," *Database* **19**(3): 86-88.

Robertson, S.E. and Sparck Jones, K. (1976). "Relevance Weighting of Search Terms," *Journal of the American Society for Information Science* **27**(3): 129-146.

Robertson, S.E., Walker, S. and Hancock-Beaulieu, M. (1995). "Large Test Collection Experiments on an Operational, Interactive System: OKAPI at TREC," *Information Processing & Management* **31**(3): 345-360.

Salton, G. and Yang, C.S. (1973). "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation* **29**(4): 351-372.

Salton, G. and Buckley, C. (1988). "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management* **24** (5): 513-523.

Shaw, Jr., W.M. (1986). "On the Foundation of Evaluation," *Journal of the American Society for Information Science* **37**(5): 346-348.

Shaw, Jr., W.M. (1994). "Retrieval Expectations, Cluster-based Effectiveness, and Performance Standards in the CF Database," *Information Processing & Management* **30**(5): 711-723.

Soergel, D. (1985). *Organizing Information: Principles of Data Base and Retrieval Systems.* London: Academic Press.

Su, L. T. (1991). "Evaluation of Interactive Information Retrieval: Implication for Operational Systems and Practice," In Edited by M. E. Williams. *Proceedings of the 12th National Online Meeting* **12**. 1991 May 7-9, New York, NY. Medford, NJ: Learned information, Inc: 391-402.

Swets, J. A. (1963). "Information Retrieval Systems," *Science* **141**: 245-250.

Swets, J. A. (1969). "Effectiveness of Information Retrieval Methods," *American Documentation* **20**(1): 72-89.

Tague-Sutcliffe, J. M. (1996). "Some Perspective on the Evaluation of Information Retrieval Systems," *Journal of the American Society for Information Science* **47**(1): 1-3.

van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd ed*. London: Butterworths.

Vidmar, D.J. (1999). "Darwin on the Web: the Evolution of Search Tools," *Computers in Libraries* **19**(5): 22-4, 26, 28.

Voorbij, H.J. (1998). "Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences," *Journal of Documentation* **54**(4): 466-476.

Webber, S. (1998). "Search Engines and News Services: Developments on the Internet," *Business Information Review* **15** (4): 229-37.