A Study of the Influence of Choice of Record Fields on Retrieval Performance in the Bibliographic Database*

서지 데이터베이스에서의 레코드 필드 선택이 검색 성능에 미치는 영향에 관한 연구

김 희 섭(Heesop Kim)**

목 차

ABSTRACTS

This empirical study investigated the effect of choice of record field(s) upon which to search on retrieval performance for a large operational bibliographic database. The query terms used in the study were identified algorithmically from each target set 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 possible choices of record field were recognised. Using INSPEC terminology, these were the fields: (1) Abstract, (2) 'Anywhere' (i.e., all fields), (3) Descriptors, (4) Identifiers, (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 D measure (bivariate in Recall and Precision). The main findings were that: (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 sensitive to the choice of query; and (3) the optimal choice of field for the D-measure is Title.

초 록

본 연구에서는 레코드필드 선택이 대규모 서지 데이터베이스 탐색 시 미치는 검색 성능에 대하여 관찰하였다. 실험의 구성 요소는 크게 (1) 대규모 상업용 데이터베이스 INSPEC, (2) 관련된 레코드들 (target sets이라고 정의함), (3) 4개의 키워드가 한 세트로 이루어진 4개의 서로 다른 형태의 질의어들 (CT__TF, CT__IDF, UT__TF, UT__IDF), (4) 최적의 질의를 위한 알고리즘, (5) 가능한 모든 경우의 탐색식을 생성해내는 블리언 탐색식 생성기, 그리고 (6) 실제 운영중인 웹 기반의 검색 시스템으로 이뤄졌다.
실험에서의 레코드 필드 선택은 (1) Abstract, (2) Descriptors, (3) Identifiers, (4) 'Subject' (Descriptors + Identifiers), (5) Title, (6) 'All fields'로 정의하여 독립변수로 채택하였다. 검색 성능은 재현율.정도율을 모두 반영한 Heine의 D측정에 의하여 평가되었다. 본 연구에서 얻은 주된 결과로는 (1) 필드선택은 검색성능에 중요한 영향을 미치며, (2) 각 검색 성능에서 보여준 순위는 질의어에 따라 민감한 결과를 보였고, (3) 제목(Title)필드 선택이 D측정에서 최적의 결과를 보였다.

Keywords: Information Retrieval; Retrieval Performance; Performance Evaluation; D measure; Field Searches; Bibliographic Database

* This article is part of the author's PhD thesis at the University of Northumbria at Newcastle, England.

** Senior Researcher, IT Information Center, ETRI (Electronics & Telecommunications Research Institute), Korea (khs01701@etri.re.kr)

논문접수일자 2001년 11월 12일
게재확정일자 2001년 12월 7일

1. INTRODUCTION

In bibliographic retrieval systems, developing a good search strategy requires knowledge about the nature and organisation of the target databases as well as, more importantly, how adequately the record fields and search terms are chosen and combined. Therefore, users should 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 has been undertaken.

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

1. 1 Research Aim and Objectives

This empirical study aims to investigate how the choice of record field(s) affects retrieval performance in searching operational bibliographic database. Derived from this aim, the following specific research objectives were defined: (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.

1. 2 Novelty of Research

To improve IR performance much effort has accordingly been given to the establishment of methods/models for evaluating retrieval effectiveness in terms of the relevant items that are retrieved. There are large literatures dealing with the issues and problems of research methodology supporting IR experiments (e.g., Sparck Jones, 1981; Tague, 1981; Ellis, 1984, 1996; Blair and Maron, 1990; Salton, 1992; Tague-Sutcliffe, 1992, 1996; Saracevic, 1995; Robertson and Beaulieu, 1997; Pors, 2000).

The following criteria recognising that several aspects of methodology in this area remain controversial: (1) the need to conduct retrieval in a realistic environment, (2) the need to utilise appropriate performance measures, and (3) the need to evaluate the significance of experimental results by applying appropriate statistical tests.

In response to the above criteria, and to detailed reading of previous studies, the methodology used here was centred on the following decisions: (1) the bibliographic database used should be a large-scale operational one, rather than a laboratory experimental collection, for real-world validity, (2) relevance judgements should be represented using objective 'real world need' as embedded in cognitive behaviour and as evidenced in (for example) citation behaviour rather than by using 'judgements of relevance' expressed against solely verbal descriptions of need, (3) the generation of sets of query terms should be algorithmic and derived from the sets of documents defined by such relevance judgements, (4) the search methodology should enable the effects of different record field choices to be investigated, and (5) the combinatorial generation of all possible logical forms of search statement should be supported, to avoid randomness arising from arbitration on search logic.

The novelty of the present study subsists in the following: (1) the unique conjunction of the above decisions which together determine the experimental methodology, (2) the implementation of that methodology in its procedural, software, results-generation, and results-analysis, aspects, and (3) a high degree of novelty in the results themselves.

2. RESEARCH DESIGN

In addition, although differences of study purpose and emphasis direct differences of experimental design, the difficulties of the interactive and operational (as opposed to the laboratory-based) IR system evaluations are also well known (Cleverdon, 1968; Salton, 1972; Su, 1991; Robertson, Walker and Hancock-Beaulieu, 1995; Beaulieu, Robertson and Rasmussen, 1996; Borlund and Ingwersen, 1997; Borlund, 2000). Due to its dynamic nature, a special treatment is required in the operational environments.

2. 1 Hypotheses of the Study

Our research question is: " How does the choice of field affect retrieval performance?" We reformulate this more precisely into the following hypotheses:

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.

The null hypothesis is the prediction that there is no difference between the results of the field searches being compared, whilst, the alternative hypothesis is the prediction that there is a difference between the results of the field searches being compared expressed using means, as shown.

2. 2 Overview of the Experimental Design

To test the hypotheses set in previous section and to investigate our major research question the experimental design was established and the synopsis is outlined in **Figure 1.**

As shown in **Figure 1**, the experiment was controlled with three different variable groups: (1) the controlled variable, (2) the independent variable, and (3) the dependent variable.

The independent variable and the dependent variable were defined in a conceptual level and were detailed in an operational level within the controlled variable. For example, how does 'the choice of field search (i.e., the independent variable)' affect 'retrieval performance (i.e., the dependent variable)' was defined in the conceptual level, and in the operational level, the research question was investigated retrieval performance of 'the

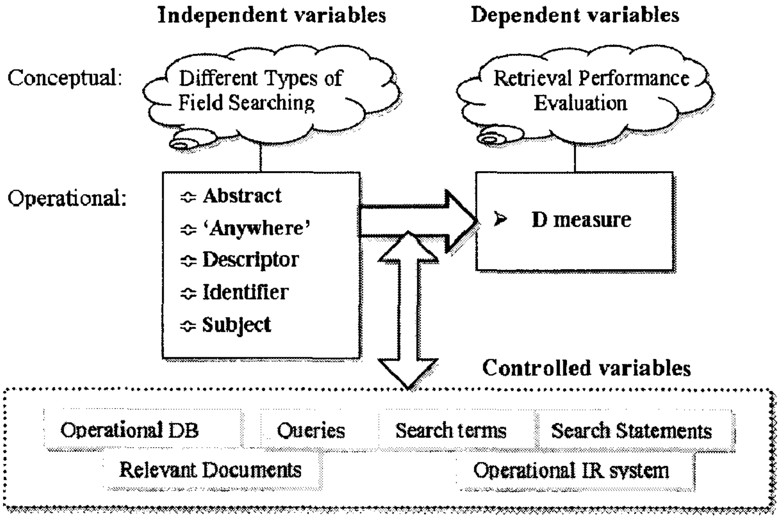


Figure 1. Overview of the Experiment Setting

six record field searches (i.e., 'Abstract', 'Anywhere', 'Descriptor', 'Identifier', 'Subject', and 'Title' field)' judging by 'D single-valued' measure within the controlled settings (i.e., under the same conditions, for example, an operational Database, Relevant documents, Queries, Search terms, Search statements, and an operational IR system).

2. 3 Controlled variables

2. 3. 1 An Operational Database - INSPEC

The operational INSPEC database was used as the test database. At the time of testing, the database covered the publications of the time span of 1969 to February 1999. In August 1998, the IEE (The Institution of Electrical Engineers) officially announced that the INSPEC contains over 6 million bibliographic records, and each year over 4,000 scientific and technical journals and some 2,000 conferences publications are scanned and is growing at the rate of 330,000 records each year. The frequency of update is weekly (The Institution of Electrical Engineers, 1998b).

2. 3. 2 Relevant Documents and Target Sets

To compare retrieval performance of the record field searches in the operational bibliographic database, we adopt a conventional definition of 'relevant document' as 'a cited reference in given review document.' And then 'target set' is defined as 'pre-identified relevant documents sets available in the INSPEC database derived from the cited references in given review documents'.

15 'review papers' were generated using a restricted random sampling technique published between 1997 and 1998. Each of the 15 review papers was then considered as the 'base-document of a target set'. The base-documents were then inspected in the SCI (Science Citation Index) through the BIDS (Bath Information and Data Services) to obtain the complete list of documents that they each cited. Each of the cited single documents (i.e., each document cited by each base document) was then checked against the operational INSPEC database 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.

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 (see **Table 1** for the size of the Target Sets).

2. 3. 3 Queries

In this experiment, a query ('topic' in TREC) was defined as 'a set of terms', rather

Table 1. The Size 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 %

than the narrative sentence expressed in natural language, or a search expression. (The IS&R literature tends to use the term 'query' in several ways, so this prescriptive definition is seen as 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; Kim et al., 1999), and (2) the reality of the 'combinatorial explosion' as a restraint on the analysis of data in this experiment using 'logical variety'.

Differences in query type can affect the performance result (Soergel, 1985; Lancaster, 1998), thus we chose the two most contrasting in character, but - if controversially - most commonly used, index languages: (1) controlled term (CT, henceforth), and (2) uncontrolled term (UT, henceforth).

At the same time, we adopted two well-known weighting techniques: (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 in regarding that tf implies 'the frequency of term in full text documents of the collection', whereas ITF 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.

Thus, the combination of two different types of index languages (i.e., CT and UT) and two different types of weighting techniques (i.e., ITF and IDF) produced four query types: (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 where the INSPEC database is also available via online service. The reason for using the EINS was their unique mechanism of the ZOOM function in this stage. The ZOOM command gives a list of the most frequently occurring index terms in a target set. The list may help to identify records specific to the search giving greater search precision, and improving recall (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), where 'N' is the number of documents in the INSPEC database; 'n' is the number of documents containing the search term in the database; 'R' is the total number of known relevant documents, known only within an experiment (that is the size of the target set), and 'r' is the number of relevant documents containing the search term in a target set, in one or other field or set of fields.

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

2. 3. 4 Search Statements - ELCs and Search Process

A 'search statement' is defined as a single 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.

The structure is additionally based on: (1) using the Boolean expression connectives 'AND' and 'NOT', (2) specifying the fields, i.e., Abstract, 'Anywhere' (i.e., all fields), Descriptor, Identifier, Subject (i.e., Descriptor plus Identifier), and Title, (3) ranging by publication dates of the coverage of the review paper, and (4) employing exact matching rather than using some other kind of syntax, e.g., not to adopt such as role indicators, word adjacency, proximity, truncation, wildcard, etc.

2. 3. 4. 1 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 adopted the ELCs (Elementary Logic Conjunctions).

The four search terms were systematically combined into ELCs. For example, (1) t_1 AND t_2 AND t_3 AND t_4; (2) t_1 AND t_2 AND t_3 AND $\neg t_4$; (3) t_1 AND t_2 AND $\neg t_3$ AND

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

t_4; (4) t_1 AND $\neg t_2$ AND t_3 AND t_4, and (5) $\neg t_1$ AND t_2 AND t_3 AND t_4 and so on. Note that the symbol '¬' denotes 'NOT'.

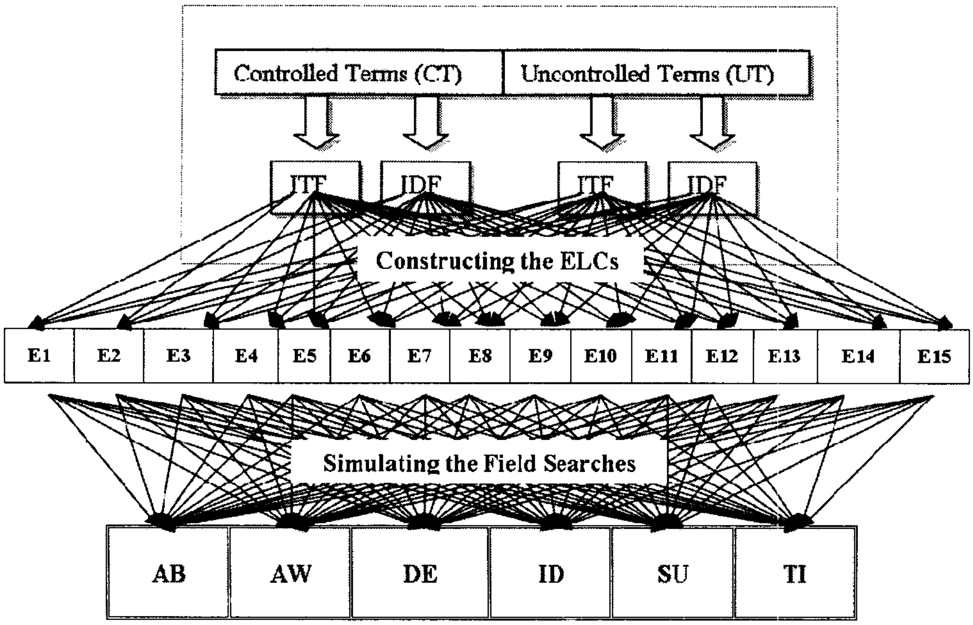
These ELCs eligible in the search statement determine a partitioning of the database and hence also of the chosen Target Set into 16 (i.e., 2^4) different and non-overlapping (i.e. disjoint) subsets. In effect, each record in the INSPEC database is examined and assigned to the appropriate member set of the partitioning by the set of ELCs, i.e. by that ELC which the record evaluates to 'true'. (For fuller discussion, see Heine, 1984; 1999; 2000). However, the all-negated ELC (e.g., E_ab16 shaded row in **Table 2)** was excepted, since it retrieves almost all records of the database, so that 15 (i.e., 2^4 -1) rather than 16 (i.e., 2^4) ELCs were presented. Each of 15 ELCs was presented to the INSPEC database for each query and choice of record field(s).

2. 3. 4. 2 Search Processes

All ELCs (except E_ab16) were presented to the INSPEC database for the six chosen record fields once other variables had been fixed. **Figure 2** illustrates the procedures of constructing the ELCs and simulating the field searches.

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	$\neg t_4$



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

As show in **Figure 2**, for a particular four-term query search expressions were generated from the relevant ELCs taken one at a time, two at a time, three at a time, etc, up to fifteen at a time, generating 32,767 (i.e., $2^{15}-1 = 2^{(2^4-1)}-1$) searches, i.e. generating search expressions using Boolean 'OR-ing' applied to the various possible combinations of ELC. This ensured that all possible search expressions were used, i.e. that the experiment suppressed one source of experimenter arbitration and experimental objectivity was maximised. (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 that was seen as lying outside the scope of the present study. Accordingly, we preferred not to make assumptions as to the selection of search expression grammars that users might make in practice.)

2. 3. 5 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 designed and maintains by SilverPlatter®. Permission to use this was granted by IEE (The Institution of Electrical Engineers).

2. 4 Independent variables - Field Searches

In the experiment, what we manipulated is called the independent variable. Each record in the INSPEC database, which was used for this study, contains an English-language title and abstract, together with full bibliographic details, which include the journal or other publication title, the author' s name and affiliation and the language of the original document.

In this study, the independent variables are six field searches that are shown in **Table 3.**

2. 5 Dependent variables - Performance measures

There are well-known single-valued measure models, for example, Brookes (1968), van Rijsbergen' s E(β) measure (van Rijsbergen, 1979), Shaw' s "harmonic mean" *F*

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

(Shaw, 1986), and Heine's D (Heine, 1973a; 1973b). Van Rijsbergen's measure is usefully parametric in a quantity, β, which can 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.

The major attractions of all these single-valued measures are as follows: (1) they reflect retrieval effectiveness alone, independently of criteria such as cost, or speed of searching the database, (2) they are independent of the number of documents retrieved in a particular search, i.e. probabilistic in nature, and (3) they can be expressible as a single number instead of two values such as Recall and Precision, thus allowing searches to be more easily compared on arithmetic scales.

Heine (1973a; 1973b) examined Recall and Precision using the Swets model (Swets, 1963; Swets, 1969) and suggests a "general measure" of retrieval performance effectiveness D, referred to the MZ (Marczewski-Steinhaus) metric. The formula and definition of D measure will be found in **Table 4.**

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.

The evaluation factors used to assess the retrieval performance of a given set of user queries with respect to a document collection are normally based on a two by two contingency table which distinguishes between the documents retrieved in answer to a given query and those not retrieved, and between items judged to be relevant to the query and those not relevant.

3. DATA COLLECTION

In this study, like many other IR tests, data collection was involved at each stage of experiment from setting up the controlled variables (e.g., data on the database, data on the selection of the sets of queries, data on the search statements, etc.) to get data for and evaluating the dependent variables.

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) NET(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. 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 (the total number of the Target Sets) x 4 (the total number of the query types) x 6 (the total number oi the choice of fields) x 15 (the number of the ELC per a query). The entire searching processes were performed in the Web-based online operational INSPEC database using the WebSPIRS™ - SilverPlatter' s Information Retrieval System for the Web version.

The summary of each query type for this particular Target Set is presented in Table

(a) including the following information: (i) the 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, and (v) the size of the target set (i.e., 'a + c') which was pre-identified. The size of each the set of retrieved documents, i.e., the search result (i.e., 'a

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

(a) Basic information

Target Set ID	#05	
Query Type - CT_ITF	(t_1)	MICROWAVE MEASUREMENT
	(t_2)	PERMITTIVITY MEASUREMENT
	(t_3)	COAXIAL CABLES
	(t_4)	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

+ b'), and the size of retrieved relevant documents set (i.e., 'a'), are presented in Table (b) for each ELC derived from this query. **Table 5** present the sample of the search results for the query type CT_ITF in 'Abstract' field for Target Set #5.

3. 2 Calculating D Values

To calculate the D measure (i.e., the dependent variables in this experiment), the 15 forms of ELC were combinatorially disjoined to generate all possible logical form of search statement, doing so for each triple "Target Set, query, and choice of record fields". For the D measure, the smaller the value, the more effective is the retrieval. 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. This value was defined and transformed the system-missing value as '1' in SPSS^{TM}.

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 Methodology of Statistical Analysis

The analysis is thus of the effects of searches using different types of queries applied to different choices of record field, expressed using the performance measures D. The statistical analyses used SPSS® 10.0 for Windows™, and the results are presented in figures and tables.

Two gradational analysis 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.

As an auxiliary for the descriptive statistics, the Percentiles (i.e., values that divide cases

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

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).

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.

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

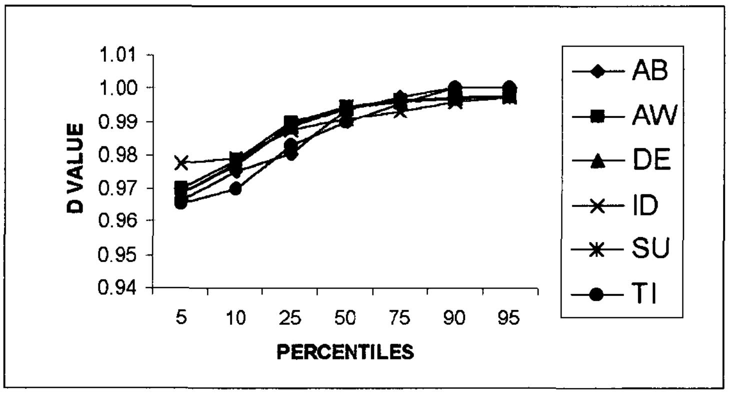


Figure 4. Percentile of D measure for CT__ITF query type - Target Set #5

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			

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

4. 2. 1 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 32,767 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.**

'Title' field search gave the best performance among the six chosen field searches in

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)*

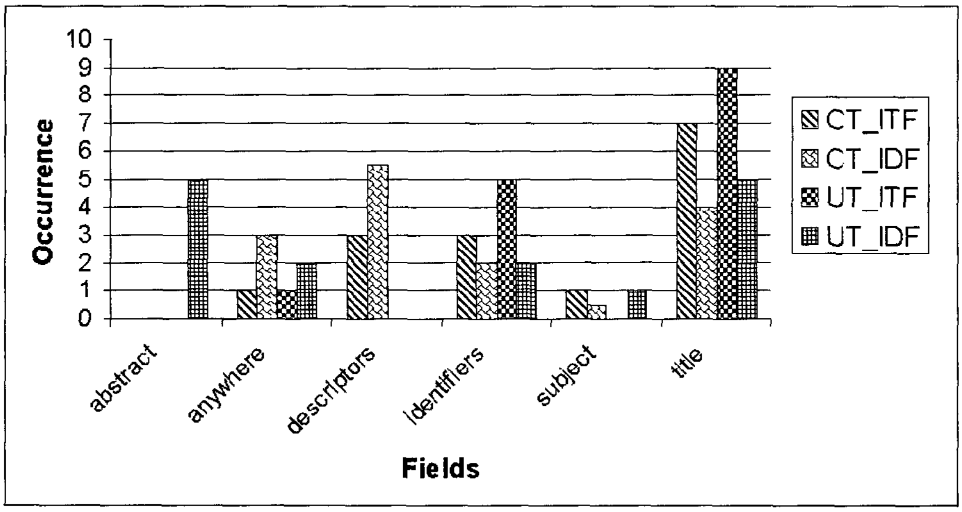


Figure 5. Comparison of the Occurrences of the best D measure Performance from the overall 15 Target Sets

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'.

4. 2. 2 AVOVA analysis

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

- H_0: No differences exist between the mean-D values for the six choices of field(s).
- 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 H_0 was rejected and the alternative hypothesis H_1 accepted. The result was thus significant beyond the 95% level.

5. CONCLUSIONS AND FURTHER RESEARCH

This study has, since it is primarily methodological and empirical, produced a large mass of data sets and it has not been easy to distinguish that which is significant from that which is of little novelty and/or value, but obtaining such a large quantity of data was seen 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.

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, which we have argued for, and also the typicality or otherwise of the INSPEC database as a bibliographical database.

Firstly, the overall D measure results of the comparative evaluation with the performance of the chosen six field searches may confirm in descending order as follows: TI 〉 ID 〉 DE 〉 AW 〉 AB 〉 SU for data aggregated for all query types.

Secondly, for data separated for each query type, this ranking became TI 〉 DE 〉 ID 〉 AW = SU 〉 AB for CT__ITF query type; DE 〉 TI 〉 AW 〉 ID 〉 SU 〉 AB for CT__IDF query type; TI 〉 ID 〉 AW 〉AB = DE = SU for UT__ITF query type; and AB = TI 〉AW = ID 〉 SU 〉 DE for UT__IDF query type.

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

Firstly, 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.

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.

Bibliography

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).

Beaulieu, M., Robertson, S.E. and Rasmussen, E.M. 1996. "Evaluating Interactive Systems in TREC," *Journal of the American Society for Information Science* **47**(1): 85-94.

Blair, D. and Marson, M.E. 1990. "Full-Text Information Retrieval: Further Analysis and Clarification," *Information Processing & Management* **26**(2): 437-447.

Borlund, P. 2000. "Experimental Components for the Evaluation of Interactive Information Retrieval Systems ," *Journal of Documentation* **56**(1): 71-90.

Borlund, P. and Ingwersen, P. 1997. "The Development of a Method for the Evaluation of Interactive Information Retrieval Systems," *Journal of Documentation* **53**(3): 225-250.

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.

Cleverdon, C.W. 1968. *The Methodology of Evaluation of Operational Information Retrieval Systems based on a Test of MEDLARS.* Cranfield: College of Aeronautics.

Ellis, D. 1984. "Theory and Explanation in Information Retrieval Research," *Journal of Information Science* **8**(1): 25-38.

Ellis, D. 1996. "The Dilemma of Measurement in Information Retrieval Research," Journal of the American Society for Information Science **47**(1): 23-36.

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. 1999. "Measuring the Effects of AND, OR and NOT operators in Document Retrieval Systems using Directed Line Segments," *Workshop on Logical and Uncertainty Models for Information Systems of the Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, London, 5-6 July, 1999. 22pp.

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

Heine, M.H. 1973b. "The Inverse Relationship of Precision and Recall in terms of the Swets' Model," *Journal of Documentation* 29(1): 81-84.

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.

Kim, H. et al. 1999. "Correlations between Users' Characteristics and Preferred Features of Web-based OPAC Evaluation", *ETRI Journal* **21**(4): 83-93.

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.

Pors, N.O. 2000. "Information Retrieval, Experimental Models and Statistical Analysis," *Journal of Documentation* **56**(1): 55-70.

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. and Beaulieu, M. 1997. "Research and Evaluation in Information Retrieval," *Journal of Documentation* **53**(1): 51-57.

Robertson, S.E., Walker, S. and Beaulieu, M. 2000. "Experimentation as a Way of Life: OKAPI at TREC," *Information Processing & Management* **36**(1): 95-108.

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. 1972. "The "Generality" Effect and the Retrieval Evaluation for Large Collection," *Journal of the American Society for Information Science* **23**(1): 11-22.

Salton, G. 1992. "The State of Retrieval System Evaluation," *Information Processing & Management* **28**(4): 441-449.

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.

Saracevic, T. 1995. "Evaluation of Evaluation in Information Retrieval," In Edited by

E. A. Fox, P. Ingwesen and R. Fidel. *SIGIR '95: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 18th Annual International Conference on Research and Development in Information Retrieval* **18**. 1995 July 9-13, Seattle, WA. New York, NY: ACM Press: 138-146.

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.

Sparck Jones, K. Ed. 1981. *Information Retrieval Experiment.* London: Butterworths.

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, J. M. 1981. "The Pragmatics of Information Retrieval Experimentation," In Edited by K. Sparck Jones. *Information Retrieval Experiment.* London: Butterworths: 59-102.

Tague-Sutcliffe, J. M. 1992. "The Pragmatics of Information Retrieval Experimentation, Revisited," *Information Processing & Management* **28**(4): 467-490.

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.