Retrieval effectiveness of controlled and uncontrolled index terms in INSPEC database

Heesop Kim
Department of Library and Information Science,
Kyungpook National University,
Daegu, 702-701, KOREA
e-mail: heesop@knu.ac.kr

ABSTRACT

The purpose of this empirical study is to assess the retrieval effectiveness of controlled and uncontrolled index terms in bibliographic database. Two types of index terms were tested in a web-based environment using the operational large-scale INSPEC database. 15 query types used in the study were both controlled terms and uncontrolled index terms derived from inverse document frequency weights. The retrieval effectiveness was evaluated using Precision. The main finding indicates there are statistically significant differences in precision arising from the two types of index terms; the uncontrolled index terms demonstrate better precision than the controlled index terms.

Keywords: Bibliographic Databases; Controlled vocabulary; Uncontrolled vocabulary; INSPEC database; Information retrieval; Index terms; Precision and recall.

INTRODUCTION

Indexes are needed for any information collection, and the function of an index is to give users systematic and effective shortcuts to the information they need. In most cases, without an index, the retrieval of information would be impractical. An index has two general purposes: to minimize the time and effort in finding information, and to maximize the searching success of the user. Both purposes are accomplished by choosing the best words that will match a user's search language (Cleveland and Cleveland 2001). Good indexing is closely related to the searching stage of information retrieval. When a user brings a query to a system, the query needs to be indexed using the same indexing language that was used to index that target document. The goal of indexing is to accurately represent the content of an item with terms that are explicit to the information searchers.

There are two main approaches to indexing: (1) a user-oriented, and (2) a document-oriented indexing. The user-oriented indexing assumes that users can predict the potential needs and approaches of indexers, and target their choice of terms towards the topics they may seek and the search terms they may use. Whereas, the document-oriented indexing assumes that the only thing users can know the content of the documents. Therefore indexers serve users best by indexing the content precisely, so that when the latter search the index, they will find the document that is relevant to their information need (Browne and Jermey 2007).

The quality of indexing is assessed in two ways: effective retrieval, and the agreement with an optimum set of terms such as consistency with expert decision. The retrieval effectiveness includes both recall, i.e. the proportion of relevant items that are retrieved;

and precision, i.e. the proportion of retrieved items that are relevant. Indexing aims to increase the proportion of relevant items that are retrieved.

Factors that influence indexing quality are: the experience of the indexer, including subject knowledge; vocabulary factors, such as the fit of the controlled vocabulary to the documents; document factors such as quality of writing, complexity, and variation between documents; process factors, such as rules and instructions; and environmental factors like noise and lighting. Studies in the 1960s and 70s suggested that free-text searching could provide results that were equal to or better than searches using human indexing based on a controlled vocabulary. Later studies using larger databases with realistic search queries however have challenged these findings. Controlled and uncontrolled (free-text, or natural) vocabulary searches usually provide unique hits, i.e., each search type finds some relevant items that the other does not find, and are therefore complementary. They also work better for different types of searches. Uncontrolled vocabulary searching is needed to find names of individuals (if these are not indexed by the database), while controlled vocabulary searches are particularly useful for broad concept searches where the topic of interest is not explicitly mentioned in the text (Lancaster 2003). Indexing languages can be categorised into a number of fundamental types. An initial breakdown would include assigned-term and derived-term systems. For the assigned-term, an indexer must assign terms or descriptors on the basis of subjective interpretation of the concepts implied in the document, which have to use intellectual effort. The indexer determines the subject matter of the document at hand, and assigns descriptors from a controlled vocabulary, which identifies concepts expressed by the document's author. For the derived-term system, all descriptors are taken from the text itself. Thus, author indexes, title indexes, citation indexes, and natural language indexes are all derived-term system. Whereas all indexing languages with vocabulary control devices such as subject heading lists, thesauri, and classification schemes are assigned-term systems. The derived-term systems are sometimes called natural language or free-text indexing or uncontrolled vocabulary, because the system allows the indexer to select the terms to be used directly from the text being indexed (Cleveland and Cleveland 2001).

Rowley (1994) reviews the issue of controlled and uncontrolled languages in a perspective on information retrieval practice and research used in searching the databases of the online hosts, in-house IR systems, online public access catalogues, and databases stored on CD-ROM. The article divides the history of the debates into four ears and concludes that a mixture of controlled and uncontrolled indexing languages used to search a wide variety of different kinds of databases as suggested same in the field of soil mechanics (Muddamalle 1998). Svenonius (1986) reviews on the issue of uncontrolled versus controlled vocabularies and suggested that some areas for research would clarify outside the context of large-scale or case-study retrieval experiment to contribute to a rational basis for the design of retrieval tools such as thesauri. Lancaster (2003) identifies that controlled vocabulary and uncontrolled vocabulary is one of the factors that have the most effect in determining consistency in indexing with the number of terms assigned.

The issue of controlled and uncontrolled vocabulary continues to be an area of debate and study in information retrieval (Aitchison and Tracy 1969; Aitchison et al. 1970; Bhattacharyya 1974; Henzler 1978; Calkins 1980; Carrow and Nugent 1981; Fugmann 1982; Svenonius 1986; Fidel 1991; Muddamalle 1998; Savoy 2005; White 2013). However, little has been done in the context of large-scale operational experiment. The difficulties of evaluating an information retrieval system are also well known (e.g., Sparck Jones 1981; Borlund 2003; Buckley and Voorhees 2000; Kekalainen and Jarvelin 2002; Ruthven 2005;

Jarvelin 2009; Belkin, Cole and Liu 2009) although it is perhaps a fair criticism of much or even most empirical work in this field that it had tended to focus on static test collections. Another difficulty is how to define the relevant documents in large-scale test collections. This was the motivation for the present study. Therefore the following specific research objective was recognised: to discover whether differences in retrieval effectiveness arising from controlled vocabulary and uncontrolled vocabulary are significant in a large-scale bibliographic database. The following sections detail the methodologies employed to achieve the research objective, and discuss the result from the research experiment.

THE EXPERIMENT

Hypothesis

The research objective: "to discover whether differences in retrieval effectiveness arising from controlled vocabulary and uncontrolled vocabulary are significant in a large-scale bibliographic database" is reformulated more precisely as the following null hypothesis (H_0) and alternative hypothesis (H_1):

> H_0: No difference in retrieval effectiveness exists between controlled and uncontrolled index terms (i.e., H_0: $\mu_1 = \mu_2$),
>
> H_1: Differences in retrieval effectiveness exists between controlled and uncontrolled index terms (i.e., H_1: $\mu_1 \neq \mu_2$)

where:

> μ_1 is the mean value of Precision measure for the 'Controlled index terms',
>
> μ_2 is the mean value of Precision measure for the 'Uncontrolled index terms'.

To test the hypothesis-pair defined above, the experiment was conducted under the following environments.

Query types – Controlled and Uncontrolled Index Terms

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, or a search expression (the information retrieval 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 is centred on two considerations: (a) previous survey results showed that most searchers use less than 3 terms per query in average (Wallace 1993; Kim 1998) and (b) the reality of the 'combinatorial explosion' as a restraint on the analysis of data in this experiment using 'logical variety'.

In the case of INSPEC, controlled terms can be defined as the descriptor field which contains standard (or preferred) term from the INSPEC thesaurus. Each record has at least one term (and usually several) assigned to it. The 2011 edition of the INSPEC thesaurus contains approximately 9,400 preferred terms. Whereas, uncontrolled terms can be defined as the identifier field which 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 descriptor field.

Differences in query type (i.e. in the 'type' of term defining the query) can affect the performance result (Soergel 1985; Lancaster 2003). The two most contrast characters were chosen: (a) controlled terms (CT) – contains standard (or preferred) term from the INSPEC thesaurus. Each record have at least one term assigned to it; and (b) uncontrolled terms (UT) - contains free language words and phrases assigned by the human INSPEC index experts. At the same time, we adopted a well-known weighting technique to choose terms for the queries: inverse document frequency (IDF). 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 one different types of weighting technique (i.e., IDF) generated queries of two distinct character: (i) CT_IDF - a query type made up of the controlled terms derived from the inverse document frequency weights; and (ii) UT_IDF - a query type made up of the uncontrolled terms derived from the inverse document frequency weights.

In this 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:.

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

An Operational Large-Scale INSPEC Database

The operational INSPEC database was used in this experiment. At the time of testing, the database covered publications from 1969 to February 1999. In 2012, the Institution of Engineering and Technology (IET) announced that INSPEC contains over 10 million bibliographic records and is growing at the rate of 600,000 records each year, and that each year over 4,000 scientific and technical journals and some 2,000 conferences publications and other publications were being scanned (The Institution of Engineering and Technology 2012). Figure 1 shows an example of INSPEC bibliographic record.

Relevant Documents

According to Pao (1989) most information retrieval experimental studies have adopted one of the following three methods to define the relevant documents: (a) prepared answer sets – prior to any searching, answer sets are prepared for a number of queries to be searched in the experimental data set; (b) pooled relevant retrieved sets – real requests are collected and each is searched by a searcher; and (c) cited references as relevant sets – this scheme succeeds in bypassing the process of subjective relevance assessment and the arbitrary establishment of a total relevance set in the test collections. Since cited documents are assumed to be relevant to the citing document, this provides a set of ready-made known relevant documents for the query in the information retrieval experiment.

ACCESSION NUMBER: 8408376

TITLE: A reappraisal of the chemical composition of the Orion nebula based on Very Large Telescope echelle spectrophotometry
AUTHOR(S): Esteban, C.; Peimbert, M.; Garcia-Rojas, J.; Ruiz, M. T.; Peimbert, A.; Rodriguez, M.
AUTHOR AFFILIATION: Inst. de Astrofisica de Canarias, La Laguna, Spain
JOURNAL: Monthly Notices of the Royal Astronomical Society vol.355, no.1, p.229-47, 89 refs.
PUBLISHER: Blackwell Science for R. Astron. Soc.
PUBLICATION DATE: 21 Nov. 2004
PUBLISHED IN: United Kingdom
DOI: 10.1111/j.1365-2966.2004.08313.x
CODEN: MNRAA4
ISSN: 0035-8711
SICI: 0035-8711(20041121)355:1L.229:RCCO;1-7
LANGUAGE: English
DOCUMENT TYPE: Journal Paper
TREATMENT: Bibliography; Experimental

ABSTRACT: We present Very Large Telescope (VLT) UVES echelle spectrophotometry of the Orion nebula in the 3100-10 400 Angstroms range. We have measured the intensity of 555 emission lines, many of them corresponding to permitted lines of different heavy-element ions. This is the largest set of spectral emission lines ever obtained for a Galactic or extragalactic H II region. We have derived He+, C2+, O+, O2+ and Ne2+ abundances from pure recombination lines. This is the first time that O+ and Ne2+ abundances have been obtained from these kinds of lines in the nebula. We have also derived abundances from collisionally excited lines for a large number of ions of different elements. In all cases, ionic abundances obtained from recombination lines are larger than those derived from collisionally excited lines. We have obtained remarkably consistent independent estimations of the temperature fluctuation parameter, t2, from different methods, which are also similar to other estimates from the literature. This result strongly suggests that moderate temperature fluctuations (t2 between 0.02 and 0.03) are present in the Orion nebula. We have compared the chemical composition of the nebula with those of the Sun and other representative objects. The heavy-element abundances in the Orion nebula are only slightly higher than the solar ones, a difference that can be explained by the chemical evolution of the solar neighbourhood.

CLASSIFICATION CODES: A9840H H II regions, emission nebulae ; A9580J Photographic region astronomical observations
THESAURUS TERMS: astronomical photometry; astronomical spectra; element relative abundance; H II regions; spectrophotometry
FREE TERMS: Orion nebula chemical composition; Very Large Telescope echelle spectrophotometry; VLT UVES; emission line intensity; permitted line; heavy-element ion; Galactic region; extragalactic H II region; ion abundance; recombination line; collisionally excited line; temperature fluctuation; 3100 to 10400 Å; He; C; O; Ne
CHEMICAL INDEXING: He el; C el; O el; Ne el
NUMERICAL INDEXING: wavelength 3.1E-07 to 1.04E-06 m
ASTRONOMICAL OBJECT INDEXING: M42

Figure 1: INSPEC Sample Record

In this study, we adopted the third method to define the relevant documents., A 'target set' was defined as the set of relevant documents that is available in our test collection (i.e., the INSPEC database). In essence, the cognitive behaviour of end-users (the authors of review papers) was used as an operational definition of 'relevance'. Although there is controversy, this had the advantages over conventional approaches to defining 'relevant documents' by third party judges of: (a) resting the definition on a definite, given information need (namely the need to identify sources most appropriate to the information that the author of the review paper was reviewing); and (b) 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 also used previously by Heine (1984) in the context of retrieval from Medline, and by Waffenschmidt, Hausner and Kaiser (2010) in their evaluation of German CCMED database. Bradshaw (2003) proposed an indexing technique that joins measures of relevance and impact in a single retrieval metric, and indicated that reference directed indexing improved relevance. More discussions on the concepts of relevance can be seen in other papers, e.g., Harter (1992), Schamber (1994), Borlund (2003), and Saracevic (2007).

For this experiment, 15 'review papers' published between 1997 and 1998 recorded by INSPEC were generated in this manner using a restricted random sampling technique. 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 Science Citation Index (SCI) through the UK Bath Information and Data Services (BIDS) that serves to obtain a 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 was identified and 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. 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.

Search Statements – Elementary Logical Conjunctions (ELCs)

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 our experiment this structure was based on: (a) Boolean connectives 'AND' and 'NOT'; (b) specifications of one search fields to be searched against, i.e. (in our case), 'Anywhere' (i.e., all fields); (c) a range by publication dates of the coverage of the specific base-document involved; and (d) 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 and wildcard.

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 Elementary Logical Conjunctions (ELCs) of each query's terms. For example, the four search terms that made up each query defined such ELCs as: 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 '$\neg$' denotes 'AND NOT'. Since each query term can be either 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). However, the all-negated ELC (e.g., E_ab16 shaded row in Table 1) was an exception since 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).

Search Processes

All ELCs (except E_ab16) were presented to the INSPEC database for the 'Anywhere (AW)' search fields once a specific query type had been chosen. 'Anywhere' does not refer to a field, but this 'words anywhere' option searches against the free text fields in the INSPEC, including the other fields.

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, 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).

An Operational IR System – WebSPIRS™

In applying experimental-derived search statement to the INSPEC, an operational Web-based SilverPlatter®'s Information Retrieval System (WebSPIRS™) was used which was designed and maintained by SilverPlatter®. A snapshot of WebSPIRS™ is shown in Figure 2.

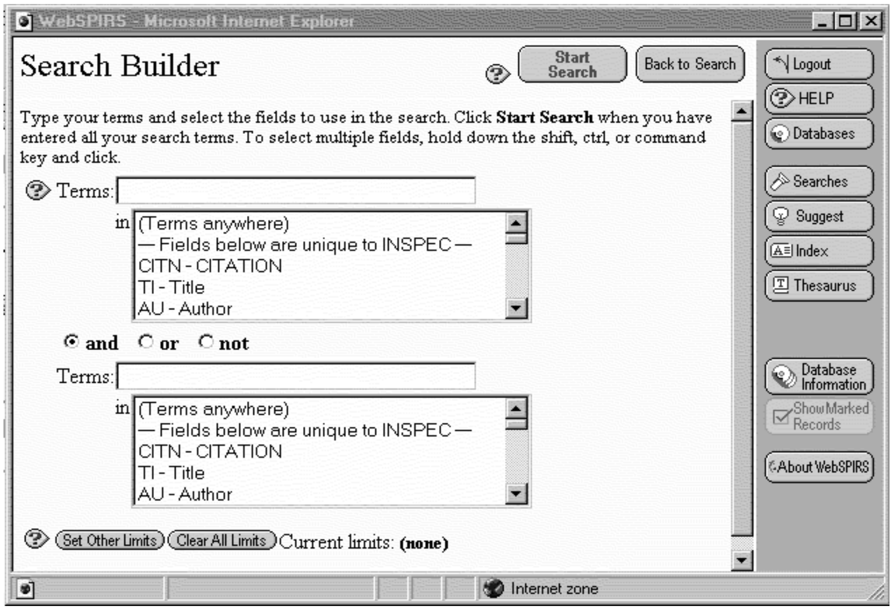

Figure 2: WebSPIRS™ Search Screen

Performance Measures - Precision

Retrieval performance is traditionally and most often measured by precision (P) and recall (R) although these two measures are continuously the foci of controversial discussions. The definition of the Precision is "the ratio between the number of relevant documents retrieved and the total number of documents retrieved from a system" (Chu 2003, p.70) where we adopt this measure in our experiment. The 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 (referred to as a 'query') and those not retrieved, and between documents judged relevant to the information need and those not relevant (Table 1).

Table 1: 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) Performance measures – Precision

Symbol	Formula	Explanation
P	$$\frac{a}{a+b} \quad (2)$$	The ratio between the number of relevant documents retrieved and the total number of documents retrieved from a system.

DATA COLLECTION

Searching

The form of a search statement ELC 13 for the CT_IDF query in Target Set #10 is, for example, as follows:

Find: (SOLID-STATE PHASE TRANSFORMATIONS *in* **AW**) **NOT** (HIGH-TEMPERATURE SUPERCONDUCTORS ***in*** **AW**) **NOT** (ELECTRON DIFFRACTION EXAMINATION OF MATERIALS ***in*** **AW**) **NOT** (TYPE II SUPERCONDUCTORS ***in*** **AW**) **AND** (**PY**=1969-1997)

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 900 searches for one query type. That is 15 (number of the Target Sets) x 4 (number of the query) x 15 (number of ELCs per query).

Table 2: ELC Searches – for CT_IDF Query Type in Target Set #10

(a) Basic Information

Target Set ID	#10	
Search Terms Type - CT_IDF	(t_1) (t_2) (t_3) (t_4)	SOLID-STATE PHASE TRANSFORMATIONS HIGH-TEMPERATURE SUPERCONDUCTORS ELECTRON DIFFRACTION EXAMINATION OF MATERIALS TYPE II SUPERCONDUCTORS
Publication Date Limit	1986 – 1997	
Total number of relevant documents (a + c)	69	

(b) Search Results

Label	ELCs	No of Retrieved Relevant Doc. (a)	No of Retrieved Doc. (a + b)
E_aw01	t_1 AND t_2 AND t_3 AND t_4	0	0
E_aw02	t_1 AND t_2 AND t_3 AND $\neg t_4$	0	0
E_aw03	t_1 AND t_2 AND $\neg t_3$ AND t_4	0	0
E_aw04	t_1 AND $\neg t_2$ AND t_3 AND t_4	0	0
E_aw05	$\neg t_1$ AND t_2 AND t_3 AND t_4	0	2
E_aw06	$\neg t_1$ AND $\neg t_2$ AND t_3 AND t_4	0	73
E_aw07	$\neg t_1$ AND t_2 AND $\neg t_3$ AND t_4	0	228
E_aw08	$\neg t_1$ AND t_2 AND t_3 AND $\neg t_4$	0	18
E_aw09	t_1 AND $\neg t_2$ AND $\neg t_3$ AND t_4	0	0
E_aw10	t_1 AND $\neg t_2$ AND t_3 AND $\neg t_4$	0	0
E_aw11	t_1 AND t_2 AND $\neg t_3$ AND $\neg t_4$	0	0
E_aw12	t_1 AND $\neg t_2$ AND $\neg t_3$ AND $\neg t_4$	0	22
E_aw13	$\neg t_1$ AND $\neg t_2$ AND $\neg t_3$ AND t_4	20	8636
E_aw14	$\neg t_1$ AND $\neg t_2$ AND t_3 AND $\neg t_4$	10	4047
E_aw15	$\neg t_1$ AND t_2 AND $\neg t_3$ AND $\neg t_4$	1	3170
E_aw16	$\neg t_1$ AND $\neg t_2$ AND $\neg t_3$ AND $\neg t_4$	Excepted	Excepted

The summary of each query type for this particular Target Set is presented in Table 2(a) including the following information: (i) a Target Set reference number, (ii) the four search terms used, (iii) the range of publication dates particular to this Target Set, and (iv) 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 2(b) for each ELC derived from this query. Table 2 and 3 presents the sample of the search results for the query type of controlled index terms (i.e., CT_IDF) and uncontrolled index terms (i.e., UT_IDF) for Target Set #10, respectively.

Table 3: ELC Searches – for UT_IDF Query Type in Target Set #10

(a) Basic Information

Target Set ID	#10	
Search Terms Type - UT_IDF	(t_1)	YBA2CU3O7
	(t_2)	PLANAR DEFECTS
	(t_3)	HIGH-RESOLUTION ELECTRON MICROSCOPY
	(t_4)	HIGH TEMPERATURE SUPERCONDUCTOR
Publication Date Limit	1986 – 1997	
Total number of relevant documents (a + c)	69	

(b) Search Results

Label	ELCs	No of Retrieved Relevant Doc. (a)	No of Retrieved Doc. (a + b)
E_aw01	t_1 AND t_2 AND t_3 AND t_4	0	0
E_aw02	t_1 AND t_2 AND t_3 AND $\neg t_4$	0	0
E_aw03	t_1 AND t_2 AND $\neg t_3$ AND t_4	0	0
E_aw04	t_1 AND $\neg t_2$ AND t_3 AND t_4	0	0
E_aw05	$\neg t_1$ AND t_2 AND t_3 AND t_4	1	11
E_aw06	$\neg t_1$ AND $\neg t_2$ AND t_3 AND t_4	3	112
E_aw07	$\neg t_1$ AND t_2 AND $\neg t_3$ AND t_4	2	43
E_aw08	$\neg t_1$ AND t_2 AND t_3 AND $\neg t_4$	2	50
E_aw09	t_1 AND $\neg t_2$ AND $\neg t_3$ AND t_4	0	14
E_aw10	t_1 AND $\neg t_2$ AND t_3 AND $\neg t_4$	0	1
E_aw11	t_1 AND t_2 AND $\neg t_3$ AND $\neg t_4$	0	0
E_aw12	t_1 AND $\neg t_2$ AND $\neg t_3$ AND $\neg t_4$	0	20
E_aw13	$\neg t_1$ AND $\neg t_2$ AND $\neg t_3$ AND t_4	32	20156
E_aw14	$\neg t_1$ AND $\neg t_2$ AND t_3 AND $\neg t_4$	4	2278
E_aw15	$\neg t_1$ AND t_2 AND $\neg t_3$ AND $\neg t_4$	1	457
E_aw16	$\neg t_1$ AND $\neg t_2$ AND $\neg t_3$ AND $\neg t_4$	Excepted	Excepted

RESULTS

Statistical Analysis

Statistical analyses were conducted to explore and describe the retrieval effectiveness for controlled and uncontrolled index terms in INSPEC database. Several statistics summary were examined using the descriptive procedure in the Statistical Product and Service Solutions (SPSS) tool. This is a principal procedure 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.

As shown in Table 4 and 5, the descriptive statistics are presented including parameters such as: (a) N (i.e., number of cases – 32,767); (b) Mean value (i.e., the arithmetic averages); (c) Standard Deviation (i.e., a measure of how much observations vary from the mean, expressed in the same units as the data); (d) Standard Error (i.e., a measure of variability); (e) 95% confidence interval for the mean with lower bound and upper bound; (f) Minimum (i.e., the smallest value); and (g) Maximum (i.e., the largest value). Both Tables 4 and 5 present an example of the descriptive statistics of Precision for controlled and uncontrolled index terms in Target Set #10.

Table 4: Descriptive Statistics of Precision for Controlled Index Terms – Target Set #10

N	MEAN	STD. DEVIATION	STD. ERROR	95% CONFIDENCE INTERVAL FOR MEAN		MIN	MAX
				Lower Bound	Upper Bound		
32767	1.55E-03	8.6466E-04	4.7767E-06	1.5477E-03	1.5664E-03	.000	.002

Table 5: Descriptive Statistics of Precision for Uncontrolled Index Terms – Target Set #10

N	MEAN	STD. DEVIATION	STD. ERROR	95% CONFIDENCE INTERVAL FOR MEAN		MIN	MAX
				Lower Bound	Upper Bound		
32767	6.7188E-03	1.0748E-02	5.9375E-05	6.6025E-03	6.8352E-03	.000	.091

To compare between the two index terms, all descriptive statistics were accumulated in one table and depicted in a graph. Table 6 presents all the 15 sets of the mean value of Precision for Controlled and Uncontrolled Index Terms. It reveals that uncontrolled index terms outperformed in all the 15 tests without any exception in this experiment. It is obvious that using uncontrolled index terms in INSPEC database give high precision in retrieval effectiveness.

Table 6: Mean value of Precision for Controlled vs. Uncontrolled Index Terms

Target Set no.	Mean of Controlled Index Terms	Mean of Uncontrolled Index Terms
1	0.00484	0.13067
2	0.00288	0.02719
3	0.00153	0.01426
4	0.00288	0.24467
5	0.01024	0.02420
6	0.00580	0.05544
7	0.02387	0.06290
8	0.00153	0.00267
9	0.00532	0.14496
10	0.00155	0.00672
11	0.00223	0.36329
12	0.01380	0.03868
13	0.00104	0.00295
14	0.00272	0.20756
15	0.01035	0.02509

Test of Hypothesis

The statistical hypothesis test, t-test at a level of significance 0.05 was carried out to test between the hypotheses:

> H_0: No difference in retrieval effectiveness exists between controlled and uncontrolled index terms (i.e., H_0: $\mu_1 = \mu_2$),
>
> H_1: Differences in retrieval effectiveness exists between controlled and uncontrolled index terms (i.e., H_1: $\mu_1 \neq \mu_2$)

where:

> μ_1 is the mean value of Precision measure for the 'Controlled index terms',
>
> μ_2 is the mean value of Precision measure for the 'Uncontrolled index terms'.

Table 7 presents the results of t-test which indicated 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. In other words, statistically there exists significant difference between controlled and uncontrolled index terms in their retrieval effectiveness in INSPEC database.

Table 7: T-test Result of Precision for Controlled and Uncontrolled Index Terms

	Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval	
					Lower	Upper
Controlled Term	3.739	14	.002	.0060388133	.002574763	.009502863
Uncontrolled Term	3.243	14	.006	.0900834733	.030506931	.149660015

Note that the t-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate to compare the means of two groups, and especially appropriate as the analysis for the two-group randomized experimental design.

CONCLUSIONS AND FUTURE WORK

This paper reports the retrieval effectiveness of controlled and uncontrolled index terms in an operational INSPEC database. We describe strategies of the experimental design employed in this study in the form of hypothesis, search statement and process, and test collection (e.g., Query, Database, Relevant document, IR system, and Relevance judgement). Two main conclusions have been drawn, principally expressed in terms of the result of this large-scale operational bibliographic database environment.

Firstly, the uncontrolled index terms reveal better Precision than the controlled index terms in their retrieval effectiveness in INSPEC database. For high precision, it is recommended to search with identifier which contains free language words and phrases assigned by the human INSPEC index experts.

Secondly, the two types of index terms, i.e., controlled and uncontrolled index terms, are statistically significant in the result of t-test. Therefore, these two index terms have explicitly different characters in nature as some previous studies (e.g., Rowley 1994; Lancaster 2003; White 2013) reported.

Future studies may consider investigating the combination of two index terms rather than isolation of one particular type of index terms, although it would involve a more complicated experimental design. For example, Muddamalle (1998) and Savoy (2005) conclude that the best performance could be achieved by the two in combination. There are also many intrinsic and situational variables taken into account in the effectiveness of index terms, for example, the nature of the subject discipline, the nature of size and levels of the index term, and the nature of the retrieval system (Svenonius 1986; Boyce and McLain 1989).

Finally, while there are definite advantages to both types of index terms, it is clear that appropriate enhancements, for example, limited use of indexing and/or development of searching aids, are likely to improve their effectiveness (Lancaster 2003).

ACKNOWLEDGEMENT

The author would like to appreciate the anonymous referees for their constructive comments. This research was supported by Kyungpook National University Research Fund, 2010.

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).

Belkin, N., Cole, M. and Liu, J. 2009. A Model for evaluation of interactive information retrieval. *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*: 7-8.

Bhattacharyya, K. 1974. The effectiveness of natural language in science indexing and retrieval. *The Journal of Documentation*, Vol. 30, no.3: 235-254.

Borlund, P. 2003. The concept of relevance in IR. *Journal of the American Society for Information Science*, Vol. 54, no. 10: 913-925.

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,* Vol. 40, no. 4: 273-276.

Bradshaw, S. 2003. Reference directed indexing: Redeeming relevance for subject search in citation indexes. *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, Vol. 2769: 499-510.

Browne, G. and Jermey, J. 2007. *The indexing companion*. Sydney: Cambridge University Press.

Buckley, C. and Voorhees, E. M. 2000. Evaluating evaluation measure stability. SIGIR 2000 *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l: 33-40.

Calkins, M.L. 1980. Free text or controlled vocabulary? A case history step-by-step analysis… plus other aspects of search strategy. *Database,* Vol. 3, no. 2: 53-67.

Carrow, D. and Nugent, J. 1981. Comparison of free-text and index search abilities in an operating information system. Information Management in the 1980s: *Proceedings of the American Society for Information Science 40th Annual Meeting*. September 26 October 1. 1977. White Plains. NY: Knowledge Industry Publication: 131-138.

Chu, H. 2003. *Information representation and retrieval in the digital age*. New Jersey: Information Today.

Cleveland, D.B. and Cleveland, A.D. 2001. *Introduction to indexing and abstracting*. Colorado: Libraries Unlimited.

Fidel, R. 1991. Searchers' selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science,* Vol.42, no. 7: 501-514.

Fugmann, R. 1982. The complementarity of natural and indexing languages. *International Classification,* Vol.9, no. 3: 140-144.

Harter, S. P. 1992. Psychological reference and information science. *Journal of the American Society for Information Science*, Vol. 43, no. 9: 602-615.

Heine, M.H. 1984. Information retrieval from classical databases from a signal-detection standpoint: A review. *Information Technology: Research and Development*, Vol. 3, no. 2: 95-112.

Henzler, R.G. 1978. Free or controlled vocabularies: Some statistical user-oriented evaluations of biomedical information systems. *International Classification,* Vol. 5, no. 1: 21-26.

The Institution of Engineering and Technology 2012. IET InspecDirect: User guide. Available at http://inspecdirect-service.theiet.org/private/home.aspx.

Jarvelin, K. 2009. Explaining user performance in information retrieval: Challenges to IR evaluation. *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval,* ICTIR 2009: 289-296.

Kekalainen, J. and Jarvelin, K. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, Vol. 53, no. 13: 1120-1129.

Kim, H. 1998. *User differences in interactive web-based OPAC evaluation*. Unpublished M.Phil thesis, Department of Information Studies, University of Sheffield, UK.

Lancaster, F.W. 2003. *Indexing and abstracting in theory and practice*, 3rd Edition. London: Library Association Publishing.

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,* Vol. 49, no.10: 881-887.

Pao, M.L. 1989. *Concepts of information retrieval.* Colorado: Libraries Unlimited.

Robertson, S.E. and Sparck Jones, K. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, Vol.27, no.3: 129-146.

Rowley, J. 1994. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, Vol. 20, no. 2: 108-118.

Ruthven, I. 2005. Integrating approaches to relevance, In: A. Spink and C. Cole, eds. *New Directions in Cognitive Information Retrieval*, Dordrecht: Springer.

Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science,* Vol. 58, no. 13: 1915-1933.

Savoy, J. 2005. Bibliographic Database Access using Free-text and Controlled Vocabulary: An Evaluation. *Information Processing and Management*, Vol.41, no. 4: 873-890.

Schamber, L. 1994. Relevance and information behavior. *Annual Review of Information Science and Technology*, Vol. 29 : 3-48.

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.

Svenonius, E. 1986. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science,* Vol.37, no.5: 331-340.

Swets, J. A. 1963. Information retrieval systems. *Science*, Vol. 141, no. 3577: 245-250.

Waffenschmidt, S., Hausner, E., Kaiser, T. 2010. An Evaluation of Searching the German CCMED Database for the Production of Systematic Reviews. *Health Information and Libraries Journal,* Vol. 27, no. 4: 262-267.

Wallace, P. 1993. How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. *RQ*, Vol.33, no.3: 239-252.

White, H. 2013. Examining scientific vocabulary: mapping controlled vocabularies with free text keywords*. Cataloging & Classification Quarterly*, Vol.51, no.6: 655-674.