Bibliographic database access using free-text and controlled vocabulary: An evaluation

Jacques Savoy
Institut interfacultaire d'informatique
Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract

This paper evaluates and compares the retrieval effectiveness of various search models, based on either automatic text-word indexing or on manually assigned controlled descriptors. Retrieval is from a relatively large collection of bibliographic material written in French. Moreover, for this French collection we evaluate improvements that result from combining automatic and manual indexing. First, when considering various contexts, this study reveals that the combined indexing strategy always obtains the best retrieval performance. Second, when users wish to conduct exhaustive searches with minimal effort, we demonstrate that manually assigned terms are essential. Third, the evaluations presented in this article study reveal the comparative retrieval performances that result from manual and automatic indexing in a variety of circumstances.

Keywords: Bibliographic database; manual indexing; automatic indexing; evaluation, French test collection.

1. Introduction

During the last decade, electronic bibliographic tools have been accessed by an increasing amount of users, many of whom might be classified as novices. During this same period, the cost of scientific journals has increased exponentially, forcing many university libraries to reduce the number of journal titles or substitute paper versions with those that can be accessed electronically, thus increasing the demand for Internet-based information access. This trend is also due to the ever-increasing availability of various and general reference information services (e.g., the Oxford English Dictionary, Encyclopaedia Britannica, or various statistics covering national or other themes) together with several bibliographic databases (e.g.,

Inspec, Biosis, Springer LINK or the ISI Web of Science). In the latter case, in order to provide effective access, various modern indexing and abstracting services (such as MEDLINE, PsychINFO, Chemical abstracts) make use of some sort of human subject analysis and indexing (Milstead, 1992), often invoking a controlled vocabulary (e.g., Library of Congress Subject Headings).

Manual indexing (Anderson and Pérez-Carballo, 2001a) usually relies on the use of controlled vocabularies in order to achieve greater consistency and to improve manual indexing quality (Svenonious, 1986). The advantage of these authority lists is that they prescribe a uniform and invariable choice of indexing descriptors and thus help normalize orthographic variations (e.g., "database" or "data base"), lexical variants (e.g., "analyzing", "analysis") or examine equivalent terms that are synonymous in meaning. The level of generality may be represented by hierarchical relationships (e.g., "Ford" is a "car"), and related-term relationships (e.g., "see also"). However, while controlled vocabularies may increase consistency among indexers, it is more important to increase indexer-requester consistency, thus leading to an increase in the chance that searchers will be able to locate the information they require (Cooper, 1969).

Working with a relatively large French collection of bibliographic records, this paper compares the retrieval performance of an automatic text-word indexing with an indexing strategy based on manually assigned controlled descriptors. The remainder of this paper is organized as follows. Section 1.1. presents related work describing the usefulness of manual indexing terms for searchers while Section 1.2 reviews previous work that has compared automatic and manual indexing performance. Section 2 describes the Amaryllis corpus and its thesaurus, along with the various search models used in this paper. Section 3 presents our evaluation methodology and compares the retrieval effectiveness of various approaches used to index and retrieve French documents. Finally, a conclusion provides an account of our study's main findings.

1.1. Manually assigned descriptors and searchers

In order to verify whether or not manual indexing might be of use to searchers, various studies have analyzed the search process. When comparing search performance based on either controlled vocabularies or full texts (able to match terms in the text of the article),

Tenopir (1985) found that Boolean full-text searches resulted in better recall yet less precision when compared to controlled-vocabulary searches (31 queries on the Harvard Business Review Online database containing years 1976 to 1984, and a controlled list containing around 3,000 terms). Thus the use of controlled vocabularies seems to improve precision (Tenopir, 1985), (Svenonious, 1986).

In similar research using a subset of the MEDLINE collection known as OHSUMED, Hersh et al. (1994) investigated search performance differences. When submitting queries, searchers either used terms only found in topic descriptions and in title and abstract sections of retrieved scientific papers, or they also considered the MeSH (Medical Subject Headings, a list containing around 19,000 main headings). Overall, performance differences were small and statistically insignificant, illustrating that MeSH descriptors did not offer any real advantages. However, when comparing experienced librarians with physicians, the former demonstrated a statistically significant advantage in recall, thus suggesting that with trained intermediaries, assigned descriptors manually could be worthwhile. This finding partially contradicts Tenopir's conclusion. Blair (2002) also indicated that experienced searchers are important components in successful searches on very large systems.

In another study, Spink & Saracevic (1997) showed that using terms extracted from controlled vocabularies or other thesauri did not seem to be the most productive sources of terms when the goal was to increase the retrieval effectiveness, as compared to cases in which terms were provided by users' search statements or by terms extracted during a relevance feedback process.

Moreover, manual indexing may also serve other purposes. For example, using the OHSUMED test-collection, French et al. (2002) showed that when searching in distributed collections and by selecting an appropriate number of MeSH terms (between one and three) from retrieved documents and by using these terms in an augment query, the selection procedure effectiveness could be improved.

1.2. Manual and automatic indexing

Even though assigning descriptors manually does produce mixed results, they can be useful when indexing a document. First, we must recognize that manual indexing is a current practice for various information services (Milstead, 1992). However, few studies have been

conducted in order to analyze and compare the relative retrieval effectiveness of either manual or automatic indexing approaches used within various information retrieval models.

In an early work, Cleverdon (1967) reported that in the Cranfield II test context (1,400 documents, 221 queries), single-word indexing was more effective than using terms extracted from a controlled vocabulary, however both indexing schemes were done by human beings. In order to analyze the performance of automatic indexing approaches, Salton (1972) compared a Boolean search system (MEDLARS) with a ranked output produced by a vector space model (SMART). Based on 450 documents (a rather small number compared to current evaluation standards), this study showed that an automatic indexing procedure was capable of retrieval performances comparable to manual indexing.

Rajashekar & Croft (1995) used the INSPEC test collection (12,684 documents, 84 queries) to evaluate retrieval effectiveness when combining various query formulations and different document representations. In this case, the authors examined an automatic indexing procedure based on the articles' title and abstract sections, manually assigning terms extracted from the title and abstract, and a third document representation based on manually assigning terms extracted from a controlled vocabulary. This study showed that automatic indexing based on the articles' title and the abstract performed better than any other single indexing schemes. While the controlled vocabulary terms by themselves were not effective representations, their presence as an additional source of evidence on document content might improve retrieval performance. More generally, combining multiple query formulations and/or multiple document representations usually improves the mean average precision.

The objective of this current article is to enlarge upon these previous investigations through examining the performance achieved by ten different retrieval strategies and comparing several document indexing schemes. Moreover, in contrast to some earlier studies, an effort was made to place the user at the center of information retrieval evaluation, with the relevance assessments on the Amaryllis corpus being made by the same person who submitted her/his information needs (Saracevic et al., 1988). Finally, our evaluation was based on a large test collection of bibliographic material written in French and covering various scientific disciplines. As did Blair (2002), we too believe that retrieving information from a small database does not satisfactorily reveal all the underlying search problems faced when handling large document collection.

2. Amaryllis corpus and search models

The Amaryllis corpus was created at INIST (INstitut de l'Information Scientifique et Technique at Vandoeuvre, France, a public organization of around 340 persons), having as its mission to collect, process and distribute the results of technological and scientific research. INIST mainly provides electronic access to two bibliographic databases named FRANCIS (for arts and humanities) and PASCAL (for science, technology and medicine). Overall, the INIST collections include around 26,000 scientific journals (plus various proceedings and Ph.D. theses). The PASCAL database contains around 14.7 million records (76% of them written in English, 9% in French, and as well as other European languages), while the FRANCIS contains 2.5 million records (41% in English, 31% in French, plus some other European languages).

This section describes the overall background for our study and is organized as follows: Section 2.1 contains an overview of the Amaryllis test collection made available during the CLEF (Cross-Language Evaluation Forum) campaign. Section 2.2 describes how we constructed a stopword list and a stemmer for the French language. Finally, Section 2.3 describes the various vector space term weighting schemes used in this paper, together with the Okapi probabilistic model.

2.1. Test collection overview

The corpus used in this paper is in fact a subset of the CLEF 2002 test suite (Peters et al., 2003) called the Amaryllis corpus and is composed of 148,688 scientific bibliographic records written in French. These records consist of a title (delimited by the tag <TI>) and an abstract delimited by the tag <AB>. The title field is not present for all documents; more precisely, only 110,528 documents (74%) have a title field due to the fact that only titles written in French are stored in this test-collection.

Each article contains a set of manually assigned descriptors delimited by the tag <MC>, and the corresponding descriptors written in English are delimited by the tag <KW>. These indexing terms manually assigned by documentary engineers at INIST, who have a good knowledge of the domain to which the indexed article belongs. These assigned descriptors are occurrences or variants of terms extracted from the INIST thesaurus. When the most appropriate terms cannot be found in the INIST thesaurus, the indexer may freely assign them

(although this rarely happens). Table 1 contains examples of these documents, whose general structure corresponds to the examples found in other online bibliographic services.

<DOC>
<DOCNO> AM-000001 </DOCNO>
<AB> Emploi d'un scanneur Bell et Howell pour documents relatifs aux achats (facturation ...) dans la firme britannique Bloor Homes, firme spécialisée dans la construction de logements </AB>
<MC> Scanneur, Document financier, Achat, Facturation, Construction de logement </MC>
<KW> Scanner, Financial document, Purchases, Invoicing, House building </KW>
</DOC>
…
<DOC>
<DOCNO> AM-000004 </DOCNO>
<TI> Les marchés de l'environ créent plus d'emplois que de métiers </TI>
<AB> A mesure que l'observation du marché de l'emploi environnement se développe, les tendances enregistrées depuis quelques années se confirment. Des emplois en augmentation régulière mais des professions et des métiers encore peu nombreux et peu reconnus, une relation formation-emploi difficile à trouver, des métiers écartelés entre faibles et hautes qualifications: les décalages du marché de l'emploi environnement sont encore importants. Il n'en reste pas moins que les préoccupations d'environnement semblent avoir trouvé leur place sur le marché de l'emploi: plutôt que "vague verte", l'environnement s'inscrit dans la durée </AB>
<MC> Protection environnement, Emploi, Marché travail </MC>
<KW> Environmental protection, Employment, Labour market </KW></DOC>
…

Table 1: Amaryllis corpus: Examples of two bibliographic records

INIST created and maintained a thesaurus that was made up available during the CLEF evaluation campaign (a part of which can be found in Table 2). It contains 173,946 entries delimited by the tags <RECORD> … </RECORD>. Each entry contains a French word or expression (delimited by the tag <TERMFR>) and its translation into English (marked by the tag <TRADEENG>). We found 36 entries to be without any real interest (e.g., "1910-1920" in Table 2). In 45,300 entries, the English translation is identical to the French expression (e.g., "Aquitaine" in Table 2). Moreover there are 28,387 multiple entries for a given term. For example, in Table 2 are there are two entries for the expression "Bureau poste," translated as "Post offices" or "Post office." By removing non-pertinent and multiple entries, we obtain a set of 145,523 (173,946 - 36 - 28,387) unique entries.

In addition to the English translation(s) for all entries, the INIST thesaurus contains three different term relationships, namely 26,154 SYNOFRE (French synonym for 18.0% of

the entries), 28,801 AUTOP (automatic expansion, available for 19.8% of the entries) and 1,937 VAUSSI ("See also", given for 1.3% of the entries). The AUTOP association is used to add automatically term(s). An example would be the term "Aquitaine" for which the term "France" is automatically added. Due to a relatively small number of links between terms, this thesaurus can be viewed as a specialized bilingual dictionary or as an authority list having a rather flat structure, as compared to the MeSH thesaurus which included more links for each of its entries (http://www.nlm.nih.gov/mesh/meshhome.html). However, before providing access to this tool, INIST has removed some of the relationships included in their original thesaurus.

<RECORD> <TERMFR> Analyse de poste <TRADENG> Station Analysis … <RECORD> <TERMFR> Bureau poste <TRADENG> Post offices <RECORD> <TERMFR> Bureau poste <TRADENG> Post office … <RECORD> <TERMFR> Isolation poste électrique <TRADENG> Substation insulation … <RECORD> <TERMFR> Caserne pompier <TRADENG> Fire houses <SYNOFRE> Poste incendie … <RECORD> <TERMFR> Habitacle aéronef <TRADENG> Cockpits (aircraft) <SYNOFRE> Poste pilotage … <RECORD> <TERMFR> 1910-1920 <TRADENG> 1910-1920 …	<RECORD> <TERMFR> La Poste <TRADENG> Postal services … <RECORD> <TERMFR> Poste conduite <TRADENG> Operation platform <SYNOFRE> Cabine conduite … <RECORD> <TERMFR> POSTE DE TRAVAIL <TRADENG> WORK STATION <RECORD> <TERMFR> Poste de travail <TRADENG> Work Station <RECORD> <TERMFR> Poste de travail <TRADENG> workstations <SYNOFRE> Poste travail … <RECORD> <TERMFR> Aquitaine <TRADENG> Aquitaine <AUTOP> France … <RECORD> <TERMFR> Carbonate sodium <TRADENG> sodium carbonate <SYNOFRE> Na2CO3 …

Table 2: Samples of various entries in the INIST thesaurus

Moreover, the available INIST thesaurus does not correspond to a standard ISO thesaurus that conforms to ISO recommendations (ISO 1986; ISO 1985) regarding contents, display and methods of construction and maintenance (entries form, abbreviations, vocabulary control, indexing terms, compound terms, basic relationships). For example, the INIST

thesaurus contains orphan terms, descriptors that are not related to any other descriptors (excepted to their English translations). As described previously, the INIST thesaurus is based mainly on the translation relationship while the synonymy, the hierarchy (broader term, narrower term) or the association (related term) relationships play only a secondary role. Each year, an ad hoc committee decides to include new terms (together with their translations and relationships with other terms) in the INIST thesaurus.

As with the TREC (Text REtrieval Conference) model, each topic was divided into three sections, namely a brief title, a one-sentence description and a narrative part identifying concepts related to the main request topic (see Table 3). Within this test collection are 25 queries written by INIST librarians, specialists in the various topic domains. Relevance assessments corresponding to each request were also made at INIST by the same person who wrote the topic statement. Of course, we would prefer having more queries in order to ground our conclusions on more solid foundations.

<TOP>
<NUM> 001 </NUM>
<F-TITLE> Impact sur l'environnement des moteurs diesel </F-TITLE>
<F-DESC> Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution </F-DESC>
<F-NARR> Concentration et toxicité des polluants. Mécanisme de formation des polluants. Réduction de la pollution. Choix du carburant. Réglage de la combustion. Traitement des gaz d'échappement. Législation et réglementation </F-NARR> </TOP>

Table 3: Example of an Amaryllis request

In order to provide an overview of the Amaryllis test collection, in Table 4 below we report certain statistics on the main characteristics of the bibliographic records and requests. As this table indicates, the test collection contains 413,262 unique indexing terms and regarding the number of relevant items per request, it shows that the mean (80.72) is greater than the median (67) and that the standard deviation is relatively large (46.0675), thus indicating that this test collection varies greatly relative to the number of pertinent articles per query.

To better depict the size of the various sections contained in a typical Amaryllis document, we included various statistics regarding the number of indexing terms compared to the whole document (under the label "all"), or those appearing in the manually assigned sections (under the label "MC & KW"), or only in the title and abstract sections (under the label

"TI & AB"). Thus a typical document is composed, in mean, of 104.617 indexing terms, while its title and abstract contains, in mean, 73.413 words. Finally, the manually assigned terms (both in French and English) number on average 31.205 terms or around 15 per language.

	Amaryllis
Size (in MB)	195 MB
# of documents	148,688
Number of queries	25
Number of relevant items	2,018
Mean rel. items / request	80.72
Standard deviation	46.0675
Median	67
Maximum	180 (Query #25)
Minimum	18 (Query #23)

	Number of indexing terms per document		
	all	MC & KW	TI & AB
# of unique terms	413,262	134,721	380,970
Mean	104.617	31.205	73.413
Standard deviation	54.089	14.675	48.65
Median	91	28	58
Maximum	496	166	435
Minimum	6	2	1

Table 4: Test collection statistics

2.2. Stopword lists and stemming procedures for the French language

In order to index and retrieve French documents, we needed to define a general stopword list for this language, made up of many words considered of no use during retrieval, but very frequently found in document content. These stopword lists were developed for two main reasons: Firstly, we hoped that each query and a document match would be based only on pertinent indexing terms. Thus, retrieving a document just because it contains words like "be", "your" and "the" (or their French equivalents) in the corresponding request does not constitute an intelligent search strategy. These function words represent noise, and may actually reduce retrieval performance because they do not discriminate between relevant and non-relevant articles. Secondly, by using them we would reduce the size of the inverted file, hopefully within the range of 30% to 50%. The stopword list used for this experiment can be found at http://www.unine.ch/info/clef/ and represents an enhanced version of one that we previously developed for this language (Savoy, 1999).

After removing high frequency words, an indexing procedure then applies the stemming algorithm, attempting to conflate word variants into the same stem or root. In developing this procedure for the French language, our first attempt (Savoy, 1999) removed only inflectional suffixes such that singular and plural word forms or feminine and masculine forms would conflate to the same root. More sophisticated schemes have already been proposed for the English language for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship"), such as the stemmer developed by Lovins (1968) based on a list of over 260 suffixes or Porter's stemmer (1980) that looks for about 60 suffixes. In order to develop a French stemmer able to remove certain derivational suffixes, our solution (available at http://www.unine.ch/info/clef/) will consider a limited list of 26 derivational suffixes. Thus compared, to the two English stemmers cited previously, our approach can be qualified as a "light" stemming procedure, given that the French language involves a more complex morphology than does the English language (Savoy, 1999), (Sproat, 1992).

2.3. Indexing and searching strategies

In order to obtain a broader view of the relative merit of different retrieval models, and also to compare the retrieval performance of manual and automatic indexing procedures based on various environments, we have implemented ten search models. The notation for these retrieval models and their corresponding weighting formulas are found in Appendix 1. In the simplest case, we adopted a binary indexing scheme within which each document (or request) is represented by a set of keywords, without any weights. To measure the similarity between documents and requests, we count the number of common terms, computed according to the inner product (retrieval model denoted "doc=bnn, query=bnn" or "bnn-bnn"). For document and query indexing however binary logical restrictions are often too limiting. In order to weight the presence of each indexing term in a document surrogate (or in a query), we may take term occurrence frequency into account (denoted tf) thus allowing for better term distinction and increased indexing flexibility (retrieval model notation: "doc=nnn, query=nnn" or "nnn-nnn").

Those terms however that do occur in the collection very frequently are not considered very helpful in distinguishing between relevant and non-relevant items. Thus we might count their frequency in the collection (denoted df), or more precisely the inverse document

frequency (denoted by idf = ln(n/df), with n indicating the number of documents in the collection), thus assigning more weight to sparse words and less weight to more frequent ones. Moreover, a cosine normalization could prove beneficial and each indexing weight would vary within the range of 0 to 1 (weighting scheme "doc=ntc, query=ntc").

Other variants could also be created, especially if we consider the occurrence of a given term in a document as a rare event. Thus, it may be a good practice to give more importance to the first occurrence of this word as compared to any successive or repeating occurrences. Therefore, the term frequency component would be computed as $0.5 + 0.5 \cdot [\text{tf} / \max \text{tf in a document}]$ (the term weighting scheme is denoted "doc=atn"). Moreover, we should consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document. To account for this, we integrate document length within the weighting scheme, leading to more complex formulae; for example the IR model denoted by "doc=Lnu" (Buckley et al., 1996), "doc=dtu" (Singhal et al., 1999). Finally, we also conducted various experiments using the Okapi probabilistic model (Robertson et al., 2000).

3. Evaluation

This section presents an evaluation of our experiments, and is organized as follows: Section 3.1 describes our evaluation methodology and compares the relative performance of ten retrieval models that access the French corpus in response to short, medium-size or long queries. Rather than all sections included in each document, Section 3.2 evaluates the mean average precision obtained from an automatic indexing procedure based only on the title and abstract sections of scientific articles and also when using only the document representation based on terms extracted from a controlled vocabulary list developed by human beings. Section 3.3 investigates enhancements attainable by incorporating a pseudo-relevance feedback (or blind query expansion) procedure. Finally, by comparing manual and automatic indexing procedures, Section 3.4 evaluates the best retrieval model using the expected search length (Cooper, 1968) in order to better assess users' efforts, depending on whether they are more interested in precision or in recall.

3.1. Evaluation of various search models

As a retrieval effectiveness indicator, we adopted the non-interpolated average precision (computed on the basis of 1,000 retrieved items per request), thus allowing both precision and recall to use a single number, an evaluation procedure applied during the TREC or CLEF evaluation campaign (Voorhees and Harman, 2000; Braschler and Peters, 2002). This mean average precision is computed based on the following consideration. For a given query q, we may compute the precision achieved after retrieving r documents, denoted by $Prec_r(q)$, as follows:

$$Prec_r(q) = \frac{D_r^{rel}(q)}{D_r(q)}$$

in which $D_r(q)$ is the set of retrieved documents for query q containing the first r records, and $D_r^{rel}(q)$ is the set of pertinent items included in this r first retrieved documents. To define the non-interpolated average precision, the system computes this precision value after each relevant document found in the answer list, and based on this set of precision values, an average is computed for the query q. Of course, instead of using a single query q, the system performance is computed according to a set of queries (25 in our case) and, we compute the mean over all queries average precision to obtain the mean average precision or non-interpolated average precision.

To determine whether or not a given search strategy is better than another, a decision rule is required. To achieve this, we could have applied statistical inference methods such as Wilcoxon's signed rank test or the Sign test (Salton and McGill 1983, Section 5.2), (Hull, 1993) or the hypothesis test based on bootstrap methodology (Savoy, 1997). In this paper we will base our statistical validation on the bootstrap approach because this methodology does not require that the underlying distribution of the observed data be a normal one. As stated in (Salton and McGill, 1983) and demonstrated in (Savoy, 1997), the mean average precision distribution is not always a normal one and this fact may invalidate the underlying statistical test.

In our statistical testing, the null hypothesis H_0 states that both retrieval schemes produce similar mean average precision. Such a null hypothesis will be accepted if two retrieval schemes return statistically similar means, and will otherwise be rejected. Thus, as

shown in the tables appearing in this paper, we have underlined statistically significant differences based on a two-sided non-parametric bootstrap test, and based on those having a significance level fixed at 5%. However, a decision to accept H_0 is not the equivalent of the null hypothesis H_0 being true, rather it represents the fact that "H_0 has not been shown to be false," resulting in insufficient evidence against H_0.

Moreover, in the current study, we have a relatively small number of observations (25 queries). Thus, even when faced with two retrieval schemes having different retrieval performances (H_0 is false), our statistical test cannot detect this difference in retrieval effectiveness, due to the sample size being too small.

Our evaluation results are based on queries using only the Title (T), the Title and Descriptive (TD) sections or the Title, Descriptive, and Narrative sections (TDN), as reported in Table 6. In these evaluations, we considered all sections of the document: title, abstract, and controlled vocabulary descriptors assigned by INIST's indexers. The resulting performance can thus be viewed as the best retrieval effectiveness, one that can be obtained with this corpus with respect to the given retrieval model.

	Mean average precision (% change)		
Query Model \ mean # of terms	T 3.7 terms	TD 15.6 terms	TDN 20.8 terms
Okapi-npn (baseline)	**37.27**	**46.44**	**54.17**
doc=Lnu, query=ltc	34.79 (-6.7%)	43.07 (-7.3%)	49.87 (-7.9%)
doc=atn, query=ntc	35.01 (-6.1%)	42.19 (-9.2%)	51.44 (-5.0%)
doc=dtu, query=dtc	31.82 (-14.6%)	39.09 (-15.8%)	47.97 (-11.4%)
doc=ltn, query=ntc	31.78 (-14.7%)	39.60 (-14.7%)	47.50 (-12.3%)
doc=lnc, query=ltc	26.84 (-28.0%)	37.30 (-19.7%)	46.09 (-14.9%)
doc=ltc, query=ltc	25.85 (-30.6%)	33.59 (-27.7%)	42.47 (-21.6%)
doc=ntc, query=ntc	21.55 (-42.2%)	28.62 (-38.4%)	33.89 (-37.4%)
doc=bnn, query=bnn	21.03 (-43.6%)	20.17 (-56.6%)	24.72 (-54.4%)
doc=nnn, query=nnn	8.99 (-75.9%)	13.59 (-70.7%)	15.94 (-70.6%)

Table 6: Mean average precision of various indexing and searching strategies (using different parts of the queries)

We can clearly see from the mean average precision depicted in Table 6 that the Okapi probabilistic model is in first position. It always produces the best mean average precision and this performance is used as the baseline from which the percentage of change is computed. In second position is the vector-space model "doc=Lnu, query=ltc" and in the third "doc=atn, query=ntc". The traditional tf-idf weighting scheme ("doc=ntc, query=ntc") does not exhibit

very satisfactory results, and the simple term-frequency weighting scheme ("doc=nnn, query=nnn") or the simple coordinate match ("doc=bnn, query=bnn") show poor retrieval performance. Using the Okapi as a baseline, this table also indicates how all differences in mean average precision are usually statistically significant (percentage of change underlined in Table 6 and computed according to the bootstrap statistical testing method, with a significance level of 5%).

On the other hand, when the query contained more search terms, the resulting retrieval performance increases. For example, using the Title and Descriptive (TD) sections (with a mean of 15.6 search terms), improvement is +24.6% compared to short queries with a mean of 3.7 search terms) for the Okapi model (from 37.27 to 46.44), or of +45.3% when all sections (TDN) of the topic description (37.27 vs. 54.17) are taken into account. When computing the mean improvement over our ten search models, we found an average retrieval enhancement of +26.7% compared to short queries (T) with TD requests, or +53.2% when compared to short requests (T) with TDN query formulation.

3.2. Evaluation of manual vs. automatic indexing

The Amaryllis corpus does however show another interesting feature. The sections <TI> and <AB> are used to delimited the title and the abstract respectively of each French scientific article written by the author(s) while the sections <MC> or <KW> include the controlled manually assigned terms extracted from the INIST thesaurus.

Based on the Amaryllis corpus we are therefore able to evaluate whether manually assigned descriptors resulted in better retrieval performance as compared to the scheme based on automatic indexing. To tackle this question, we evaluated the Amaryllis collection using all sections (denoted "all" in Tables 7 and 8), using only the manually assigned terms (performance listed under the label "MC & KW") or using only titles and abstracts from bibliographic records (under the label "TI & AB").

Based on short (Table 7) or medium-size (Table 8) queries, the mean average precision for the combined indexing strategy is better than both the single manual or automatic indexing schemes and these differences are usually statistically significant (difference computed with a significance level of 5% and underlined in these tables). A single exception to this rule is obtained when using the simple coordinate match (model denoted "doc=bnn, query=bnn") for

which the manual indexing scheme performs better than the combined approach (22.71 vs. 21.03 in Table 7). However, this difference is not statistically significant.

	Mean average precision (% change)			
Query Indexing sections Model	T all baseline	T MC & KW vs. all	T TI & AB vs. all	TI & AB vs. MC & KW
doc=Okapi, query=npn	**37.27**	**29.56** (-20.7%)	**23.73** (-36.3%)	(-19.7%)
doc=Lnu, query=ltc	34.79	25.81 (-25.8%)	22.74 (-34.6%)	(-11.9%)
doc=atn, query=ntc	35.01	29.11 (-16.9%)	23.32 (-33.4%)	(-19.9%)
doc=dtu, query=dtc	31.82	28.51 (-10.4%)	23.89 (-24.9%)	(-16.2%)
doc=ltn, query=ntc	31.78	26.40 (-16.9%)	20.42 (-35.7%)	(-22.7%)
doc=lnc, query=ltc	26.84	21.66 (-19.3%)	16.77 (-37.5%)	(-22.6%)
doc=ltc, query=ltc	25.85	20.90 (-19.1%)	17.42 (-32.6%)	(-16.7%)
doc=ntc, query=ntc	21.55	17.58 (-18.4%)	16.04 (-25.6%)	(-8.8%)
doc=bnn, query=bnn	21.03	22.71 (+8.0%)	11.29 (-46.3%)	(-50.3%)
doc=nnn, query=nnn	8.99	8.63 (-14.1%)	5.12 (-43.0%)	(-40.7%)
Mean difference		-14.4%	-35.0%	-22.9%

Table 7: Mean average precision when comparing manual and automatic indexing procedures (Title only)

	Mean average precision (% change)			
Query Indexing sections Model	TD all baseline	TD MC & KW vs. all	TD TI & AB vs. all	TI & AB vs. MC & KW
doc=Okapi, query=npn	**46.44**	**37.23** (-19.8%)	**29.97** (-35.5%)	(-19.5%)
doc=Lnu, query=ltc	43.07	32.17 (-25.3%)	28.22 (-34.5%)	(-12.3%)
doc=atn, query=ntc	42.19	35.76 (-15.2%)	28.16 (-33.3%)	(-21.3%)
doc=dtu, query=dtc	39.09	32.29 (-17.4%)	27.23 (-30.3%)	(-15.7%)
doc=ltn, query=ntc	39.60	32.90 (-16.9%)	24.58 (-37.9%)	(-25.3%)
doc=lnc, query=ltc	37.30	29.29 (-21.5%)	26.12 (-30.0%)	(-10.8%)
doc=ltc, query=ltc	33.59	26.62 (-20.8%)	24.44 (-27.2%)	(-8.2%)
doc=ntc, query=ntc	28.62	24.16 (-15.6%)	21.55 (-24.7%)	(-10.8%)
doc=bnn, query=bnn	20.17	19.80 (-1.8%)	11.71 (-41.9%)	(-40.9%)
doc=nnn, query=nnn	13.59	11.00 (-19.1%)	7.39 (-45.6%)	(-32.8%)
Mean difference		-19.1%	-45.6%	-32.8%

Table 8: Mean average precision when comparing manual and automatic indexing procedures (Title and Desc queries)

When comparing retrieval effectiveness for manual (label "MC & KW") and automatic (label "TI & AB") indexing schemes (as shown in the last column of Table 7 and 8), we can see that for all search models, manually assigned descriptors result in better mean average precision than do automatic indexing procedures. However, these differences are usually not statistically significant (except for the four underlined observations). This statistical finding

seems a priori counter-intuitive. For example in Table 7, the mean average precision for the Okapi model is 29.56 when using manually assigned descriptors and only 23.76 for an indexing process based on the documents' title and abstract sections. The difference between these two runs is 19.7% (last column in Table 7) and thus the manual approach is favored. However, a query-by-query analysis reveals that the manual indexing run improved retrieval effectiveness for 15 queries out of a total of 25. For 10 requests however the automatic indexing procedure depicted a better retrieval performance. Thus, in order to find a statistically significant difference between two retrieval schemes, the performance difference between individual requests should favor one given retrieval model for a large number of queries and the difference must be significant (e.g., an improvement of 0.1% cannot be viewed as significant).

Users usually enter very short queries however and are more interested in the precision revealed by the first 5, 10 or 20 retrieved items listed on the first results page (Spink et al., 2001). In order to obtain a more precise picture within this context, in Table 9 we reported precision results for 5, 10 or 20 documents retrieved using the Okapi probabilistic model. This table shows that the manual indexing scheme (labeled "MC & KW") obviously results in better performance when compared to the automatic indexing approach (labeled "TI & AB"), relative to the precision achieved after 5, 10 or 20 documents. These differences are however not statistically significant (bootstrap testing with a significance level of 5%). We achieved the best performance from using both indexing approaches (performance depicted under the label "all"), resulting in differences that are usually statistically significant.

	Precision (% change)		
Query Precision \ Indexing sections	T all (baseline)	T MC & KW	T TI & AB
Mean average precision	37.27	29.56	23.73
Precision after 5	68.0%	59.2% (-12.9%)	58.4% (-14.1%)
Precision after 10	66.0%	54.8% (-17.0%)	52.8% (-20.0%)
Precision after 20	60.0%	48.6% (-19.0%)	45.4% (-24.3%)

Table 9: Precision after 5, 10 or 20 retrieved documents (Okapi search model)

Considering the expense of manual indexing, Table 9 shows that the enhancement is disappointing. When compared to the precision after 10 documents, manual indexing shows a precision of 54.8% as compared to 52.8% for the automatic approach. Strictly speaking, this comparison is correct. However, if an institution such as INIST decides to manually index

each scientific article, it might also adopt an indexing strategy that takes into account both manually assigned descriptors and automatic indexing procedures based on the articles' title and abstract sections. Thus, comparing the performance under the "all" column in Table 9 with the precision shown under the "TI & AB" column seems to be a more reasonable approach. In this case, including a manual indexing procedure improves precision after 10 documents from 52.8% to 66.0%. Thus, in mean, we obtain 1.3 more relevant documents after 10 retrieved items when including a manually based indexing procedure. Does this improvement matter? From Lantz's study (1981), we knew that when an IR system provides 6.2 relevant citations, only 3.5 documents would finally be consulted. However, this mean value varies from one discipline to another; for every 100 relevant document retrieved, medical scientists tend to read more articles (around 42) while engineers consult only a small number of papers (8). Biological, physical and social scientists form an indistinguishable group, tending to read 27 articles on average. Since this study was conducted in London, we may consider that cultural differences could invalidate, or at least attenuate, this finding. Nonetheless, manual indexing can be viewed as more important for recall-oriented users such as lawyers or medical researchers than for precision-oriented users.

We must recognize however that, to the best of our knowledge, little research has been done regarding the impact of additional relevant documents and to what extent they will meet user information needs. The interactive track at TREC (Hersh and Over, 2001; Over, 2001; Hersh, 2003) presented an interesting set of studies on various aspects of human-machine interactions, and also some more specific experiments pertaining to cross-lingual information retrieval systems were presented in (Gonzalo and Oard, 2003).

Manual indexing does however include other advantages. For example, the Westlaw company (an online legal research service) manually indexes various court decisions. This improves online searching and also provides their users with concise statements covering entire cases, clarifying them or linking them to other particular and pertinent cases that also implicate or apply a given legal concept. Moreover, the manually versions of various legal documents may also be published, not only for manual search purposes but also to provide information for various digests and annotated legal data services.

3.3. Blind query expansion

It has been observed that pseudo-relevance feedback (blind-query expansion) can be a useful technique for enhancing retrieval effectiveness through automatically developing enhanced query formulations. In this study, we adopted Rocchio's approach (Buckley et al., 1996) whereby the system was allowed to add m terms extracted from the k highest ranked documents retrieved with the original query. The new request was derived from the following formula:

$$Q' = \alpha \cdot Q + \beta \cdot \frac{1}{k} \cdot \sum_{j=1}^{k} w_{ij}$$

in which Q' denotes the new query built for the previous query Q, and w_{ij} denotes the indexing term weight attached to the term T_j in the document D_i. In our evaluation, we fixed $\alpha = 0.75$, $\beta = 0.75$.

We used the Okapi probabilistic model in this evaluation and enlarged the query from 10 to 40 terms taken from the 5 or 10 best-ranked articles. The results depicted in Table 10 indicate that for the Amaryllis corpus the optimal parameter setting seems to be around 30 terms and these values are very similar to those found by other studies done during the last CLEF evaluation campaign (Peters et al., 2003), based on other languages.

	Mean average precision (% change)		
Query Model \ Indexing sections	T all	T MC & KW	T TI & AB
Okapi-npn (baseline)	37.27	29.56	23.73
5 docs / 10 best terms	41.64 (+11.7%)	33.03 (+11.7%)	24.94 (+5.1%)
5 docs / 20 best terms	41.71 (+11.9%)	33.28 (+12.6%)	25.09 (+5.7%)
5 docs / 30 best terms	42.53 (+14.1%)	32.82 (+11.0%)	25.30 (+6.6%)
5 docs / 40 best terms	42.35 (+13.6%)	32.99 (+11.6%)	25.09 (+5.7%)
10 docs / 10 best terms	43.50 (+16.7%)	33.13 (+12.1%)	25.47 (+7.3%)
10 docs / 20 best terms	43.38 (+16.4%)	**33.33** (+12.8%)	25.70 (+8.3%)
10 docs / 30 best terms	**44.01** (+18.1%)	33.18 (+12.2%)	**25.87** (+9.0%)
10 docs / 40 best terms	43.92 (+17.8%)	32.97 (+11.5%)	25.85 (+8.9%)

Table 10: Mean average precision using blind-query expansion (Okapi model)

Using the bootstrap testing approach, Table 10 shows that differences in mean average precision are always statistically significant when using the combined or manual indexing strategies. When the system uses only the title and the abstract sections of the bibliographic

records, improvements in mean average precision are only significant for the best three query expansion parameter settings.

If we compute the precision after 5, 10 or 20 documents using the best query expansion setting, Table 11 shows how Rocchio's blind query expansion improves precision compared to Table 9 which shows corpus indexing using all sections or when the indexing is limited to the articles' title and abstract sections. However, for manual indexing, even if the mean average precision increases from 29.56 to 33.33, the precision after 5 documents decreases from 59.2% (Table 9) to 58.4% (Table 11). Thus, even though query expansion usually improves the overall performance, in some cases it may actually reduce early precision.

	Precision (% change)		
Query Precision \ Indexing sections	T all (baseline)	T MC & KW	T TI & AB
Mean average precision	44.01	33.33	25.87
Precision after 5	74.4%	58.4% (-21.5%)	59.2% (-20.4%)
Precision after 10	70.8%	54.4% (-23.2%)	52.8% (-25.4%)
Precision after 20	62.4%	50.8% (-18.6%)	47.6% (-23.7%)

Table 11: Precision after 5, 10 or 20 retrieved documents using blind query expansion

When using the performance resulting from indexing all documents sections as a baseline as shown in Table 11, differences in the precision after 5, 10 or 20 is around 20% and always statistically significant. From this table we cannot detect statistically significant differences when comparing manually assigned descriptors (labeled "MC & KW") with free-text indexing schemes based on the bibliographic records' title and the abstract sections (labeled "TI & AB"). Since most users prefer a precision-oriented search system where only a relatively few items are retrieved, in the next section we will investigate and more precisely evaluate user effort that is required to reach a given number (or percentage) of pertinent items.

3.4. Expected search length

As a retrieval effectiveness measure, Cooper (1968) suggested computing the expected search length, defined as the mean number of non-relevant items that users must scan in order to satisfy their information need. Thus this retrieval effectiveness approach really measures users' efforts to discard non-relevant items. Users' needs may be viewed according to various types, such as "only one relevant document is wanted," or more generally "k pertinent articles are required." When applying this retrieval measure in our context, we have developed a

retrieval model that provides users with a ranked list of retrieved items instead of a weak (or set oriented) ordering, as described in Cooper's paper (1968). Our current situation is thus simpler.

When evaluating our best search model using the expected search length as reported in Tables 12 and 13, we did not compute the mean but rather the median and the 3rd quartile. These latter two values are more robust location statistics and they are less influenced by outliners or extreme observations that may dramatically change a mean value (Savoy, 1997). For example, when indexing documents based on all sections and without considering a blind query expansion approach, the first relevant item for Query #3 can be found in position 99, while for 19 requests, the first retrieved item is pertinent. The resulting mean search length is therefore 4.24, and this performance value will be 0.33 if Query #3 is ignored. Clearly, 4.24 does not seem to reliably indicate the fact that for 19 requests out of 25, the first retrieved item is relevant. Based on such considerations, we prefer adopting the median and the 3rd quartile (depicted in parenthesis in Tables 12 and 13).

The median values shown in Table 12 indicate that for 50% of requests, the first retrieved item is always pertinent, no matter which indexing scheme is used. As a second indicator, figures in parenthesis show the 3rd quartile, indicating the number of non-relevant records to scan for 75% of the queries. With the combined indexing strategy, the 3rd quartile also has a value of 0. Thus, for this indexing strategy, the first retrieved item is also relevant for 75% of the requests. Based on the title and the abstract sections (last column of Table 12), for 75% of the submitted requests, on average users must discard one non-relevant item before finding a relevant document. Similar findings hold when indexing the test collection using only the manually assigned descriptors (under the label "MC & KW"). On the other hand, when accounting for blind query expansions (bottom part of Table 12), the conclusions drawn are similar.

	Expected search length		
Query Model \ Indexing sections	T all	T MC & KW	T TI & AB
Without blind query expansion			
To find first relevant item	0 (0)	0 (1)	0 (1)
To find 2 relevant items	0 (1)	1 (2)	1 (3)
To find 3 relevant items	1 (2)	1 (6)	1 (4)
To find 5 relevant items	2 (5)	2 (40)	3 (10)
To find 10 relevant items	3 (10)	5 (76)	10 (33.5)
With blind query expansion			
To find first relevant item	0 (0)	0 (2)	0 (1)
To find 2 relevant items	0 (1)	0 (9)	0 (2)
To find 3 relevant items	0 (1)	1 (16)	1 (5)
To find 5 relevant items	1 (4)	2 (24)	2 (12)
To find 10 relevant items	1 (13)	7 (82)	8 (22.75)

Table 12: Median (3rd quartile) number of items to discard in order to find k relevant records (Okapi model, with and without blind query expansion)

When users want to find a greater number of pertinent articles from a large collection, they must anticipate scanning a large (or huge (Blair, 2002)) number of retrieved items. For example, a lawyer preparing to defend a client wants to find around 75% of all relevant documents (Blair and Maron, 1985). In order to apply the expected search length when faced with recall-oriented users, in Table 13 we reported the median (and 3rd quartile) number of non-relevant items to be discarded in order to retrieve a given percentage of relevant articles, their percentages varying from 10% to 75%. Clearly however it is difficult to estimate a priori the number of relevant items a lawyer will need for a given legal precedent search. Moreover, after retrieving a given amount of pertinent documents, users cannot distinguish between the situation where additional pertinent articles do exist and the situation where the desired documents no longer exist. In addition, for a given database, there is also a difference between the objective percentage of relevant documents already extracted and the corresponding subjective percentage estimated by the user. As mentioned by Blair & Maron, (1985, p. 293) "This meant that, on average, STAIRS could be used to retrieve only 20 percent of the relevant documents, whereas the lawyers using the system believed they were retrieving a much higher percentage (i.e., over 75 percent)." In a similar vein, recent work done by Sormunen (2001) seems to indicate that low precision in high recall searching is unavoidable when using a Boolean search system, but this precision level may potentially be improved by considering better best-match searching.

From data shown in Table 13, we can see that the best solution is always combining both manual and automatic indexing. In our experiment, in order to obtain 75% of the pertinent articles using a blind query expansion (bottom part of Table 13), the median number of non-relevant items to be discarded is 253 and the 3rd quartile 502. Clearly, a recall-oriented search implies a much greater effort on the part of users, even with the best solution. This strategy is obviously better than other approaches requiring greater user effort (2,243 articles must be discarded with manual indexing or 3,499 with only automatic indexing).

	Expected search length		
Query Model \ Indexing sections	T all	T MC & KW	T TI & AB
Without blind query expansion			
To find 10% of relevant items	2 (5)	4 (55)	5 (20)
To find 20% of relevant items	8 (15)	19 (107)	28 (90)
To find 1/3 of relevant items	25 (54)	41 (122)	93 (295)
To find 50% of relevant items	85 (164)	141 (574)	313 (981)
To find 75% of relevant items	379 (843)	865 (4,964)	2,438 (6,253)
With blind query expansion			
To find 10% of relevant items	1 (5)	3 (51)	4 (28)
To find 20% of relevant items	4 (13)	8 (82)	19 (83)
To find 1/3 of relevant items	20 (35)	22 (178)	59 (274)
To find 50% of relevant items	51 (134)	141 (534)	202 (796)
To find 75% of relevant items	253 (502)	756 (2,243)	1,416 (3,499)

Table 13: Median (3rd quartile) number of items to discard in order to find a given percentage of relevant records (Okapi model, with and without blind query expansion)

Aside from the extreme case, Table 13 demonstrates that manual indexing (labeled "MC & KW") results in better retrieval effectiveness than does automatic indexing (labeled "TI & AB") when users wish more than a third of the relevant records. For lower percentages, automatic indexing tends to involve less user effort.

Our experiment thus confirms that for conducting an exhaustive search, manual indexing is important. As stated by Bates (1998, p. 1196) "However, as the "core" terms will probably retrieve a relatively small percentage of the relevant records (certainly under the half, in most cases), they must nonetheless tolerate sifting through lots of irrelevant ones. It is the purpose of human indexing and classification to *improve* this situation, to pull more records into that core than would otherwise appear here …"

4. Conclusion

Using the relatively large Amaryllis corpus, we compared the retrieval effectiveness of human controlled vocabulary based indexing to that of an automatic indexing using ten retrieval models. Using the title and abstract sections in French bibliographic records, Tables 7 and 8 show how manually assigned descriptors that were mainly extracted from an authority list performed better than did an automatic indexing scheme. However, the difference between these indexing schemes is usually not statistically significant. This main conclusion was confirmed using ten different indexing and search procedures.

The best mean average precision is however always obtained when both manually assigned descriptors and automatic text-word indexing schemes are combined (see Tables 7 and 8). As additional findings, this study has shown that:

- the Okapi probabilistic model provides the best retrieval effectiveness when considering different query formulations (Table 6) or when indexing is based on different sections of bibliographic records (Tables 7 and 8);
- results found using a French corpus corroborate with previous studies based on a collection of Web pages written in English (Savoy and Picard, 2001). Thus, the French language does not reveal any specific difficulties when known indexing and search strategies are applied;
- when users incorporate more search keywords, the resulting retrieval effectiveness increases between 24% and 45% (see Table 6). Thus, helping users find more search terms is a valid concern for man-machine interfaces;
- applying a blind query expansion may enhance the retrieval effectiveness by about 10% (see Table 10);
- when comparing manually assigned descriptors with an automatic indexing approach, based on title and abstract sections, retrieval performance favors the manual approach although a statistically significant difference between these two approaches cannot always be found (see Tables 7 and 8);
- when comparing the precision obtained after 5, 10 or 20 retrieved documents (see Tables 9 or 11), there is no real difference between these two indexing strategies;

- when an exhaustive search is required in order to retrieve 50 to 75% of relevant records, manually assigned descriptors prove to be an attractive approach, specially when used in combination with an automatic indexing scheme (see Table 13).

Of course, these findings still need to be confirmed through submitting more queries and other test collections containing manually assigned descriptors for each stored document. While our study demonstrates that combining both indexing approaches proves to be the best retrieval performance, Anderson & Pérez-Carballo (2001a) have shown that in the "machine vs. human indexing" debate there are various, often hidden, key variables. These are important in terms of exhaustive searches (humans tend to be more selective, able to make clearer distinctions between relevant and peripheral information while computers tend to take each term occurrence into account), specificity (machines tend to index document based on words as they appear in the text while humans tend to use more generic terminology), size of document units (human indexing focuses on larger units such as complete chapters or complete monographs while automatic indexing tends to work at the paragraph level).

From the point of view of bibliographic database developers, a single automatic indexing procedure that is clearly faster and cheaper might be adopted. Such an approach is capable of interesting but not optimal retrieval performance levels for those users who want high precision after retrieving only a few documents (see Tables 9, 11 and 12). In addition to this, and depending on the underlying costs of human indexing and clientele needs, our study has provided a general view of retrieval effectiveness concerning manual and/or automatic indexing strategies for those users who require more relevant documents or who need to conduct exhaustive searches.

Finally, it could prove worthwhile to stop viewing and treating every document as equally important. In fact, the "80-20 rule" may also apply in large document collections or IR databases where around 20% of articles will provide the expected answer to 80% of needs. Along these same lines, Anderson & Pérez-Carballo (2001b) suggested developing methods that could predict which documents might be more important, and for those documents a human analysis could be applied.

Acknowledgments

This research was supported in part by the SNSF (Swiss National Science Foundation) under grants 21-58 813.99 and 21-66 742.01.

Appendix 1. Weighting schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a document D_i, we might use various approaches such as shown in Table A.1, where n indicates the number of documents in the collection, t the number of indexing terms, tf_{ij} the term occurrence frequency of term T_j in document D_i, df_j the number of documents in which the term T_j appears, idf_j the inverse document frequency (computed as $idf_j = \ln(n/df_j)$), the document length (the number of unique indexing terms) of D_i is denoted by nt_i, l_i indicates the number of indexing terms of D_i, and for avdl, b, k_1, pivot and slope are constants. For our experiments, these constants were assigned the following values, avdl=200, b=0.5, k_1=1.5, pivot=30 and slope=0.2. For the Okapi weighting scheme, K represents the ratio between the length of D_i measured by dl_i (sum of tf_{ij}) and the collection mean was noted by avdl.

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5+0.5 \cdot tf_{ij} / \max tf_{i.}]$
Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$	npn	$w_{ij} = tf_{ij} \cdot \ln\left(\left(n - df_j\right) \Big/ df_j\right)$
lnc	$w_{ij} = \dfrac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} \left(\ln(tf_{ik}) + 1\right)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}}$

dtc	$w_{ij} = \left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln\left(\ln(tf_{ik}) + 1\right) + 1\right) \cdot idf_k\right)^2}$
ltc	$w_{ij} = \left(\ln(tf_{ij}) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}$
dtu	$w_{ij} = \dfrac{\left(1 + \ln\left(1 + \ln(tf_{ij})\right)\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$
Lnu	$w_{ij} = \dfrac{\left(\ln(tf_{ij}) + 1\right) \Big/ \left(\ln\left(dl_i \Big/ nt_i\right) + 1\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$

Table A.1: Weighting schemes

References

Anderson, J.D., & Pérez-Carballo, J. (2001a). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management*, 37(2), 231-254.

Anderson, J.D., & Pérez-Carballo, J. (2001b). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing & Management*, 37(2), 231-254.

Bates, M.J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13), 1185-1205.

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299.

Blair, D.C. (2002). The challenge of commercial document retrieval, Part 1: Major issues, and a framework based on search exhaustively, determinacy of representation and document collection size. *Information Processing & Management*, 38(2), 273-291.

Braschler, M., & Peters, C. (2002). CLEF methodology and metrics. In Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (Ed.), *Evaluation of Cross-Language Information Retrieval Systems*, LNCS #2406, (pp. 394-404). Berlin: Springer-Verlag.

Buckley, C, Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In Harman, D.K. (Ed.), *Proceedings of TREC-4* (pp. 25-48). Gaithersburg, MD: NIST Publication #500-236, 1996.

Cleverdon, C.W. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, 19, 173-192

Cooper, W.S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *American Documentation*, 19(1), 30-41.

Cooper, W.S. (1969). Is interindexer consistency a hobgoblin?. *American Documentation*, 20(3), 268-278.

French, J.C., Powell, A.L., Gey, F., & Perelman, N. (2002). Exploiting manual indexing to improve collection selection and retrieval effectiveness. *Information Retrieval*, 5(4), 323-351.

Gonzalo, J. & Oard, D.W. (2003). The CLEF 2002 interactive track. In Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (Ed.), *Advances in Cross-Language Information Retrieval*, LNCS #2785, (pp. 372-382). Berlin: Springer-Verlag.

Hersh, W., Buckley, C., Leone, T.J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In Croft, W.B. & van Rijsbergen, C.J. (Ed.), *Proceedings of the 17th International Conference of the ACM-SIGIR'94*, (pp. 192-201). London, UK: Springer-Verlag.

Hersh, W. & Over, P. (2001). Interactivity at the text retrieval conference (TREC). *Information Processing & Management*, 37(3), 365-367.

Hersh, W. (2003). TREC-2002 interactive track report. In Voorhees, E.M. & Buckland, L.P. (Ed.), *Proceedings TREC-2002* (pp. 40-45). Gaithersburg, MD: NIST Publication #500-251.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Korfhage, R., Rasmussen, E. & Willett, P. (Ed.), *Proceedings of the 16th International Conference of the ACM-SIGIR'93*, (pp. 329-338). New York, NY: The ACM Press.

ISO (1985) Guidelines for the establishment and development of multilingual thesauri. ISO (5964:1985).

ISO (1986) Guidelines for the establishment and development of monolingual thesauri. ISO (2788:1986). See also ANSI standard at http://www.techstreet.com/cgi-bin/detail?product_id=52601

Lantz, R.E. (1981). The relationship between documents read and relevant references retrieved as effectiveness measures for information retrieval systems. *Journal of Documentation*, 37, 134-145.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

Milstead, J.L. (1992). Methodologies for subject analysis in bibliographic databases. *Information Processing & Management*, 28(3), 407-431.

Over, P. (2001). The TREC interactive track: An annotated bibliography. *Information Processing & Management*, 37(3), 369-381.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Ed.) (2003). *Evaluation of cross-language information retrieval systems*, LNCS #2406. Berlin: Springer-Verlag.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Rajashekar, T.B., & Croft, W.B. (1995). Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science*, 46(4), 272-283.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, 23(2), 75-84.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill.

Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3), 161-176.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944-952.

Savoy, J., & Picard, J. (2001). Retrieval effectiveness on the web. *Information Processing & Management*, 37(4), 543-569.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D., & Pereira, F. (1999). AT&T at TREC-7. In Voorhees, E. M. & Harman, D.K. (Ed.), *Proceedings TREC-7* (pp. 239-251). Gaithersburg, MD: NIST Publication #500-242.

Sormunen, E. (2001). Extensions of the STAIRS study – Empirical evidence for the hypothesized ineffectiveness of Boolean queries in large full-text databases. *Information Retrieval*, 4(3-4), 257-273.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science*, 48(8), 741-761.

Spink, A., Wolfram, D., Jansen, M.B.J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Sproat, R. (1992). *Morphology and computation*. Cambridge, MA: The MIT Press.

Svenonious, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5), 331-340.

Tenopir, C. (1985). Full text database retrieval performance. *Online Review*, 9(2), 149-164.

Voorhees, E.M. & Harman, D. (2000). Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing & Management*, 36(1), 3-35.