A Performance Model of the Length and Number of Subject Headings and Index Phrases

Robert Losee

School of Information and Library Science, University of North Carolina,
Chapel Hill, NC 27599-3360 USA, <losee@ils.unc.edu>

Robert M. Losee is a Professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. He conducts research in the area of organizing and retrieving information, with a special interest in ordering documents so that entering at a given point in the ordered list finds one at the best documents for a given topic.

Losee, Robert. **A Performance Model of the Length and Number of Subject Headings and Index Phrases.** *Knowledge Organization, 31(4).* 245-251. 15 refs.

ABSTRACT: When assigning subject headings or index terms to a document, how many terms or phrases should be used to represent the document? The contribution of an indexing phrase to locating and ordering documents can be compared to the contribution of a full-text query to finding documents. The length and number of phrases needed to equal the contribution of a full-text query is the subject of this paper. The appropriate number of phrases is determined in part by the length of the phrases. We suggest several rules that may be used to determine how many subject headings should be assigned, given index phrase lengths, and provide a general model for this process. A difference between characteristics of indexing "hard" science and "social" science literature is suggested.

1. Introduction

How many subject headings should be used to describe a document, or how many index terms does it take to capture the nature of a topic? One could use a very large number of subject headings in every bibliographic record, with a great deal of redundancy existing between the subject headings, or one could choose a single, best, subject heading. Given the expected increase in retrieval precision associated with using fewer subject headings and the increase in recall associated with using more subject headings, it might be desirable to determine the number of subject headings required to give the same ordering performance as full-text queries, which represent what a user considers to be a full statement of topicality. One may choose to sacrifice precision for greater recall by increasing the number of assigned subject headings or index terms, but this is often undesirable.

Understanding the relationship between the number and size of subject headings or index terms assigned to a document provides a critical link between those who develop bibliographic records and those who use the records to retrieve documents. Intuitively there is a relationship between the amount of information in individual headings and the ability of individual subject headings to contribute to successful retrieval. Below we consider the number of subject headings needed to capture full topicality and how the number can be predicted by knowing the length, or the number of terms, in subject heading phrases.

1.1. Subject Headings and Index Terms

Subject headings and index phrases are groups of 1 to *n* terms that represent subject characteristics of documents. These phrases may be selected from a controlled vocabulary, intellectually produced by a cataloger or indexer based on their perceptions about what captures the topic of the material being processed, or subject headings may be automatically-derived free-text phrases, composed of sequential

terms extracted from natural language text. Some of the important relationships between natural language terms and subject headings or index terms are discussed by Dykstra (1988).

Subject headings and index terms are usually assigned so as to capture the topicality of the document, "no more, no less" (Smiraglia 1990, 82). As shown in Foskett's (1996) classic *The Subject Approach to Information*, there are a wide range of goals and methods for producing and assigning subject-indicating phrases. Our concern here is about the general nature of phrase length and the number of phrases used, and the consequent ability to order documents; we hope to rise above specific systems or classes of subject headings or indexing (e.g. *Library of Congress Subject Headings* or the *ERIC Thesaurus*). Because the phrases produced for experiments below are machine generated, they are not near the quality of controlled vocabulary subject headings (e.g., *LCSH*) and do not meet standards such as coextensivity (Smiraglia 1990).

While index terms or subject headings may be assigned using automated procedures that are well described in the literature, these topical indicators are often assigned by humans for use in libraries (Sauperl 2002) or for web pages (Greenberg 2003). The assignment of headings by humans is implicitly based on decisions made by the cataloger or indexer that it is better to assign a subject indicator than not to assign the indicator. When shopping, if there were no cost for the items in the store, we might try to take home everything in the store, but clearly there is a cost in most real-world stores. Similarly, there is a cost to assigning subject indicators, but catalogers often do not know what the cost is. We attempt to characterize some of these tradeoffs below.

The effects associated with the number of subject headings have been studied in several different ways. Banks (2004) summarizes much of the literature, especially that addressing the relative utility of subject headings as indicated by circulation records. She found, as have others, that there is not a strong relationship between the number of *LCSH* subject headings and circulation figures. She did find that "the optimal number of subject headings on bibliographic records appears to be one or two, which generated the largest percentage of circulation among the books in the tested sample" (Banks 2004, 22). This may be due to the difference in the inherent utility of books that are manually assigned one or two headings, as compared to the difference in breadth and specificity of those that are assigned four or five headings, for example. Studwell (1990) presents arguments for policies relating to the number of *LCSH* headings used in total, as well as the maximum number of headings that should be used for different categories of headings. Using *LCSH* headings, Khosh (1986) found that there is a weak negative correlation between the number of subject headings and the specificity of individual headings, with Khosh (1987) finding that there is no correlation between the number of subject headings and the class notation length.

1.2 Measuring Topic-Matching Performance

The study of the performance of indexing systems has a long history of investigation, with major studies often examining information retrieval systems and databases, with possibly the best being the Cranfield (Cleverdon, 1967) and Salton and Lesk (1968) studies. Retrieval systems are used to order and present documents that have been assigned subject headings or index terms. The resulting ordering of documents, using any of a number of document ranking and indexing algorithms, produces an ordered list whose characteristics may be measured. The performance of the different orderings is then compared, with those subject heading or index term assignment systems producing the better results being treated as the superior representational systems.

The quality of document orderings may be measured in different ways (Losee 1998). Precision and recall are commonly used measures, with precision computed as the proportion of retrieved documents that are relevant, and with recall computed as the proportion of relevant documents in the database that are retrieved. These measures depend upon the presence of relevance judgments, which are usually supplied by the searcher, or, in the case of a standard test database, the producer of the retrieval database. While these relevance judgments are less than perfect, they may be assumed to be indicators of topicality.

An easy to interpret measure that has some desirable statistical characteristics is the Average Search Length (*ASL*) (Losee 1998, 89-90). Measured as the average position of relevant documents in an ordered list of documents, *ASL* has the value 1 when there is a single relevant document and it is at the front of the list, and *ASL* has the value 100 when there is a list of 100 documents and the single relevant document is at the very end of the list. *ASL* is a numeric

value that directly translates into how many documents the user will have to examine when moving through the ordered set of documents until reaching the average position of a relevant document. Interestingly, the *ASL* may be predicted formulaically from parameters of a database and the characteristics of the ranking algorithm, and thus doesn't need to be computed historically from the empirical ordering of documents.

We may normalize the *ASL* by dividing it by the number of documents in the database, producing the Normalized Average Search Length (*NASL*), which ranges from 0 to 1, where 0 is the best possible ordering and 1 the worst. *NASL* may be interpreted as the probability that a document in the ordered list will occur before the average position of a relevant document, where 0 would represent no documents before the average position, and 1 representing certainty that a randomly selected document would occur before the average position of the relevant documents in the list. This measure, as well as a probability derived from it, will serve as the basis for our index term performance analysis.

2. The Number of Subject Headings

The number of subject headings or index terms needed to represent the topicality of a document, query, or topic may be estimated statistically. A major factor in computing this number of subject headings will be the number of terms in a single heading and thus the information carried by a descriptive phrase, with smaller phrases conveying less information, on the average, than more extensive descriptions. Below, we model the numeric aspects of subject headings by examining the relationship between groups of terms, of various sizes, that are extracted from user-supplied queries. These full text queries are provided by users and are the basis for the relevance judgments that link each document and query pair. A full text query provided by the user may be treated as having all the relevance information that the user chooses to provide about a topic. It will often be the case that this query is imperfect, and that better queries can be expressed, but we will focus on what is expressed (although the relationship between performance with perfect queries and performance with the expressed queries will be considered below).

The relationship between the performance with single term queries (extracted from the full query) and the performance with all terms in the query can capture the relationship between the number of small queries that provide the same capability as the full query itself. The relationship can be computed using simple math. If we double the *NASL* performance measure to be in the range of 0 to 2, and delete values above 1 (which only occur with negatively discriminating subject headings), we have the performance measure *W*, the probability that a randomly selected document from the top half of the ordered list of documents will be ordered ahead of the expected position of a relevant document. The value for *W* approaches 0 when there are few documents ahead of the average position of a relevant document in the entire ordered list, and *W* approaches 1 when the average position of a relevant document approaches the middle of the ordered list, possibly due to the random ordering of documents, placing all of the documents in the top half of the ranking ahead of the middle position.

The reason for defining *W* as we have is to allow us to relate the ordering performance with different phrase lengths with the performance with different numbers of phrases. The performance, measured by *W*, with a single term, may be compared to the (better) performance with a full natural language query by noting *how many W values associated with single term queries need to be multiplied together to achieve the superior performance obtained with the full natural language query*, which captures the full topicality that the user chooses to provide. By similarly using queries composed of two, three, and four terms grouped together, we can understand how the size of subject headings may determine the number of subject headings necessary to achieve the level of performance obtained with a full text query.

When we compute the *W* associated with multiple short subject headings, we multiply the individual *W* values to produce the *W* associated with the group of headings. Multiplication of probabilities such as these is appropriate when the *W* values (one per subject heading) are independent, that is, when the features carry no statistical dependence and there is no overlap between representations. For example, the chance of tossing a coin and getting heads followed by tails is the independent probability of tossing a coin and getting heads, 1/2, times the probability of tossing a coin and getting tails, 1/2, with1/2 times 1/2 equaling 1/4. Most subject headings are chosen so as to be relatively independent, but obvious relationships exist between some topics, such as *eating*, and other related topics, such as *obesity*.

We assume here that successive terms taken from a query may be viewed as a phrase. When the terms used are limited to adjectives and nouns, these bear a resemblance to realistic index phrases. We ignore the order in which the terms occurs in a phrase, taking a *bag-of-words* approach, where the *set* of terms is considered, but the order is considered unimportant.

Below, we compare the empirical ordering performance using a short phrase composed of successive nouns and adjectives as a retrieval query. This will allow us to consider the relationship between the individual phrases and the full information need expressed through the query (with associated relevance judgments). Clearly a single book can serve the information needs associated with more than one question, and a book may be broader (or narrower) than a query, just as a query may be broader (or narrower) than a book. However, we examine here the ability of small phrases to discriminate adequately between relevant and non-relevant documents over the set of documents and ask "how many phrases would it take to order documents as well as the ordering obtained with a full natural language statement of topicality (a query);" the question of full-topical representation vs. partial topical representation is the important issue here.

We will also consider the performance upper bounds, the best performance possible for a query, using just the words composing the query. This ranking "cheats" in that it finds the best possible ordering, knowing in advance what the user finds useful and what is not considered useful, but treats documents with identical sets of terms (that are in the query) as having equal ranking. Natural language is not always effective at conveying topicality clearly and unambiguously, and documents may be better ordered (in hindsight) so as to achieve the best performance available (vis-à-vis the query features), which is usually better than that obtained using a "standard" information retrieval matching procedure. It is unclear that the average performance for topics will ever be better than that expressed by the terms in query, other than by improving the query; it is clearly the case that the numbers developed assume that the queries and relevance judgments are reasonably high quality.

The matching method used here is the CLMF (Coordination Level Matching-Term Frequency) method, which assigns as the document weight the number of term occurrences that occur in the query and in the document. This weight is consistent with the popular TF-IDF weight used in most search engines when each term is treated as having the same weight. We chose the CLMF method because we did not want to complicate our discussion by considering the relative rarity or specificity of a term when examining the relationship between the number of subject headings and their length.

3. Data Analysis & Results

We perform this analysis on a research version of a digital library and information retrieval test system called Nyltiac (located at http://Nyltiac.com) that can process and retrieve properly formatted textual data from any discipline. Beginning with the standard ADI (American Documentation Institute) test database (Baeza-Yates and Ribeiro-Neto 1999), we were able to assign part-of-speech tags (Brill 1992) for titles and abstracts to produce a part-of-speech tagged version of the ADI database. Terms that are similar may be brought together through the use of a suffix-stripping algorithm (Porter 1980), bringing together *cat* and *cats*, for example, by removing the final *s* in *cats*. This database uses language somewhat typical of the social sciences, with one query, for example, being "How can actually pertinent data, as opposed to references or entire articles themselves, be retrieved automatically in response to information requests?" Using Nyltiac, we were able to measure the ordering performance (W) using the nouns and adjectives in queries and documents. Mathematica™ was used for post-ordering data analysis.

We can see the expected number of subject headings or index phrases that would occur in a document if the phrases were used instead of the query in the next to the last column in Table 1. The varying number of queries represents the number of queries produced (and associated relevance values) when all possible one, two, etc. sequential word phrases are extracted from the original queries. If a query has 4 nouns or adjectives, it can produce 4 separate one word queries, with the relevance judgments from the original query being assigned to all of the 4 new queries. Similarly, we could obtain 3 two-term queries (first and second terms, second and third terms, third and forth terms). The *NASL* shown in the middle column of Table 1 represents the ordering performance, with smaller numbers being better than larger numbers.

Types	Queries	NASL	Estimated Number of Subject Headings Queries	Upper Bounds
1 Term	253	.4468	3.82	9.70
2 Terms	218	.4029	1.99	5.05
3 Terms	183	.3691	1.42	3.60
4 Terms	148	.3340	1.07	2.70
Full Query	35	.3252	1.00	(NASL= 0.1679) 1.00

Table 1. *Performance with ADI database with titles and abstracts for 82 documents and 35 full text queries. CLMF weighting was used. Nouns and adjectives only.*

The second to the last column in Table 1 shows the expected number of subject headings decreasing at a decreasing rate as the size of the phrases grows. Using a linear regression to predict the number of subject headings from the logarithm of the phrase length for the data with both titles and abstracts combined, we are able to predict the number of phrases needed to capture the same amount of document ordering power as a full text query as:

$$\text{number of phrases} = -2.3 \log(\text{phrase length}) + 3.9.$$

The adjusted R^2 value for this is .94, suggesting that the phrase length variable is an excellent predictor of the number of phrases needed, although the large size is due in part to the small sample size of phrase sizes 1 through 4, making it relatively easy to fit the data. If one wishes not to use the logarithm (which addresses the "curve" in the data), a weaker fit (adjusted R^2 = .59) is produced with the linear equation:

$$\text{number of phrases} = -0.4 \text{ phrase length} + 3.6.$$

As the number of terms used in a subject heading increases, the number of subject headings needed to carry as much information as the full query decreases. The second rule, although less accurate, is probably easier to understand and apply with the simple prose rule: Begin with 3.6 phrases and then take away 4/10 of a phrase for each term included in each phrase.

The last column in Table 1 predicts the number of phrases that should be used if one wishes to achieve the best possible retrieval performance obtainable using only the query terms. One can produce no better than this value, given the aforementioned constraints, and having more independent phrases than the number given would be wasteful in any event. We note that for 3 (and 4) term phrases, the expected number of phrases (when computed vis-à-vis the upper bounds) is 3.6 (and 2.7, respectively), which is similar to what one often finds in existing catalog records. The lower amount of information available when using shorter subject headings necessitates using more of them than are needed with larger subject headings. We can predict the number of subject headings needed in this upper bounds or best case situation as:

$$\text{number of phrases} = -2.3 \log(\text{phrase length}) + 6.04$$

which has an adjusted R^2 of .94.

One may examine the relationship between the lengths of subject headings by considering the relationships between subject headings that differ by a length of 1 term. For the ADI dataset, it takes 1.92 subject headings of length 1 to provide the ordering of a single phrase of length 2, 1.41 subject headings of length 2 to provide the ordering of a single phrase of length 3, and 1.33 subject headings of length 3 to provide the ordering of a single phrase of length 4. As Table 1 shows in the next to the last column, it takes 1.07 subject headings of length 4 to provide the ordering of fully expressed query. As we see in Table 1, the larger phrases have more ordering capability than smaller phrases.

Table 2 shows data from a different discipline. The CF7479 standard retrieval database (Baeza-Yates and Ribeiro-Neto, 1999, p. 94-96) consists of 1239 document abstracts and titles with the first 50 queries in the full database, all on the topic of Cystic Fibrosis (CF), as well as the associated relevance

Types	Queries	NASL	Estimated Number of Subject Headings Queries
1 Term	329	.4122	2.43
2 Terms	279	.3575	1.40
3 Terms	230	.3204	1.05
4 Terms	182	.2918	0.87
Full Query	50	.3128	1.00

Table 2. *Performance with CF Database with titles and abstracts for 1239 documents and 50 queries. CLMF measure was used. Nouns and adjectives only. Terms were stemmed.*

judgments. A sample query from the database, showing the type of language used, is "Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respi-

ratory tract in CF?" The CF queries are full scientific questions, and the terminology is much more precise (and less ambiguous) than that used in the ADI database.

Note that the extracted phrases provide much better ordering (*NASL*) than do the comparable ADI phrases. The regression for this dataset (adjusted $R^2 = .91$) is:

$$\text{number of phrases} = -2.8 \log(\text{phrase length}) + 3.3.$$

The comparable ADI regression has a slope of -2.3; this difference may be due to the different ambiguity and precision levels found in the discipline-specific sub-languages in the different databases. The CF database contains more precise and informative terms and phrases constructed from the nouns and adjectives in the queries than does the ADI database, and thus the phrases of a given length in Table 2 are smaller than the comparable phrases in Table 1. Clearly, one needs fewer subject-bearing phrases when the phrases contain more precise language, such as one may find in the harder sciences and medicine.

The estimation of the number of subject headings above assumed that the subject headings are statistically independent, an assumption that is often violated by the use of related subject headings. If we add additional subject headings that overlap, they carry less new information than independent subject headings, requiring more subject headings (or longer subject headings) proportional to the degree of overlap.

4. Recommendations & Conclusions

The number of assigned subject headings needed to provide the ordering provided by a full text query is clearly related to the length of the phrases used. The data analyzed here suggest that those assigning index terms or subject headings to documents need to assign several headings when the headings are small, but can assign fewer headings when the headings themselves are larger and more descriptive.

Clearly, we need to represent a topic using enough subject headings to allow for the ordering performance of documents at least at the level of that provided by a full statement of information need or topicality. One should use as few phrases as possible, because we know that as more subject headings are used, the precision declines. Using fewer subject headings than the number needed to produce the ordering associated with a natural language topical statement may result in certain topics being omitted, producing uneven topical representation.

Because of the nature of this study, we are unable to say whether it is better to have too few or too many subject headings than the numbers in Table 1. Common sense suggests that using too many subject headings, distributed evenly around the topical area, is safer than using too few.

The results obtained here are dependent upon specific databases. Clearly, using numerous larger databases would give a better perspective of the length vs. number tradeoffs in indexing and assigning subject headings in a variety of subject areas and using differing representational philosophies. Using the above data, we can recommend the following approximate rules:

1. Simple rule: Use at least 1 independent subject heading if the subject headings are of length 3 or 4, while at least 2 subject headings should be used if the subject headings are of length 2.
2. Complex rule: Using the regression in the paper, it is suggested that one use as the number of subject headings at least the quantity:
   three and a half minus (2 and a half times the logarithm of the phrase length).
3. In the case of mixed lengths, we may use a point system. We give "points" for each length of a subject heading so that the points add up to at least 3.8. A single term heading is assigned 1 point, a 2 term heading is assigned 1.9 points, a 3 term heading is assigned 2.7 points, and a 4 term heading is assigned 3.59 points.

These numbers are derived from ordering of documents in the ADI database, and probably reflect the kinds of rules one should use in assigning topical terms or phrases to documents in the Social Sciences. Given the data in Table 2 and the increased precision of the language in medicine, the phrases should probably be significantly smaller or less frequent when indexing for medical or hard science documents. While the use of the ADI database to study this problem only provides an approximation to what would be found in other situations, we feel that this presentation of the method used, as well as the numbers obtained, provides a unique insight into the assignment of subject headings and index terms.

Acknowledgements

The author wishes to thank Jane Greenberg and Richard Smiraglia for their helpful comments on an earlier version of this article.

References

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern information retrieval*. Harlow, England: Addison Wesley.

Banks, J. 2004. Does the number of subject headings on a bibliographic record affect circulation intensity? *Technical services quarterly* 21n3. 17-24.

Brill, E. 1992. A simple rule-based part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing.* Morristown, NJ,: Association for Computational Linguistics, pp. 152-55.

Cleverdon, C. W. 1967. The Cranfield tests on index language devices. *Aslib proceedings* 19: 173-92.

Dykstra, M. 1988. LC subject headings disguised as a thesaurus. *Library journal* 113: 42-46.

Foskett, A. C. 1996. *The subject approach to information, 5th ed*. London: Library Association Publishers.

Greenberg, J. 2003. Metadata and the World Wide Web. *Encyclopedia of library and information science*. New York: Marcel Dekker, pp. 1876-88.

Khosh-Khui, S.A. 1986. Effects of subject specificity: part 1: specificity of L C subject headings and depth of subject analysis in monographic records. *Technical services quarterly* 4n2: 59-67.

Khosh-Khui, S. A. 1987. Effects of subject specificity: part 2: relationship of L C subject headings specificity and class notation length. *Technical services quarterly* 4n3: 33-39.

Losee, R. M. 1998. *Text retrieval and filtering: analytic models of performance*. Boston: Kluwer.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14n3: 130-137.

Salton, G. and Lesk, M. E. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computing Machinery* 15: 8-36.

Sauperl, A. 2002. *Subject determination during the cataloging process*. Lanham, MD: Scarecrow Press.

Smiraglia, R. P. 1990. Subject access to archival materials using *LCSH*. *Cataloging and classification quarterly* 11n3-4: 63-90.

Studwell, W. E. 1990. Subject suggestions 6: some concerns relating to quantity of subjects. *Cataloging and classification quarterly* 10n4: 99-104.