THE EFFECT OF ASSIGNING A METADATA OR INDEXING TERM ON DOCUMENT ORDERING

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY (IN PRESS, 2013)

January 11, 2013
Author’s draft of article to be published

Robert M Losee
U. of North Carolina,
Chapel Hill, NC 27599-3360
losee@unc.edu

ABSTRACT

The assignment of indexing terms and metadata to documents, data, and other information representations is considered useful, but the utility of including a single term is seldom discussed. We discuss a simple model of document ordering and then show how assigning index and metadata labels improves or decreases retrieval performance. The Indexing and Metadata Advantage (IMA) factor measures how indexing or assigning a metadata term helps (or hurts) ordering performance. Performance values and the associated IMA expressions are computed, consistent with several different assumptions. The economic value associated with various term assignment decisions is developed. The IMA term advantage model itself is empirically validated with computer software that shows that the analytic results obtained agree completely with the actual performance gains and
losses found when ordering all sets of 14 or fewer documents, while when the formulas in the software are changed to differ from this model, the predictions of the actual performance are erroneous.

INTRODUCTION

Access to libraries and automated information retrieval systems has often been facilitated by the use of assigned topical index terms and metadata by the searcher or searching mechanism. When there are multiple documents that are of relevance to the searcher, the order in which the documents are presented to the searcher may affect the quality of the user experience. How an individual term affects the order in which documents are presented to the user is part of the foundation of the science of knowledge organization and is the focus of this work.

Documents, data, and variables may have a range of characteristics. Some of these characteristics present attributes about the document or variable, such as the creator of a document, the date of creation, the owner, security requirements, etc. Other characteristics may represent topicality or “aboutness,” characteristics that indicate membership in a class, such as a document being about “cats” or a variable being about “grams of mass.” We focus here on topical indexing and metadata, although the discussion below applies equally well to non-topical phrases and terms, such as a document’s author, place of generation, or date of production. These non-topical terms can be applied correctly or incorrectly, and topical indexing terms or phrases can be correct or incorrect, spelled correctly or incorrectly. Topicality has more nuanced issues, such as whether a book about “ostriches” would be of interest to someone searching for a single picture of a “bird.”

The emphasis below is on the assignment of indexing or metadata expressions to facilitate the ordering of documents. Ordering is understood to be the ranking of informative media, referred to here as documents, so that the user may move progressively through the list. Document ordering for retrieval places those documents most likely to be relevant at the beginning and those least likely to be relevant at the end of the list. Given the assumption that a relevant document is worth a constant positive benefit, and that a non-relevant document is worth a constant zero benefit, this ordering can be shown to be Bayesian optimal (Van Rijsbergen, 1979; Bookstein, 1983; Manning, Raghavan, & Schutze, 2008). This list may be studied analytically (Church, 2010; Losee R. M., 1998) and characteristics of ordering robustness (Zhou & Croft, 2008) for the list may be studied and manipulated. This is somewhat different than the ordering desired when browsing, where one may attempt to arrange documents so that each document is close, in terms of subjectivity, to the documents to its left and right on a shelf, or on a display on a screen (Losee R. M., 1992). Browsing along shelves may be seen as searching a two dimensional arrangement. This two dimensional approach may be expanded to a third and further dimensions, but one, two, and three dimensions seem to be adequate for modeling most existing human-computer interface applications. The focus of our document ordering may be seen as presenting the user with a single dimension, the document
expected to be the most probably relevant at the front of the ordered list, with the user moving through the list from the front to the back.

A range of methodologies may be used to study document ordering and the assignment of index terms or metadata. For example, in 2007, Zobel and Hawking published “Does Topic Metadata Help with Web Search” that uses a large quantity of data, showing the circumstances in which metadata was superior to other terminology, and the circumstances in which this data suggests that metadata was inferior to other types of terminology (Zobel & Hawking, 2007). This kind of analysis is often used in information retrieval experiments, such as the TREC studies, where large quantities of data are used with either real or synthetic queries and real or synthetic needs and relevance judgments.

The analysis below is methodologically different than studies of “real” systems or experimental systems, which produce statistically analyzed results. Instead, probabilistic methods may be used to model document ordering (Yip, 2009) and then the effect of properly or improperly adding a metadata term to a relevant or non-relevant document can be examined, showing how the label improves or harms the document ordering. If mathematically correct, these results hold when the assumptions of the models are correct, with the only performance changes being due to the parameters of the models. Using analytic models developed below, we are not concerned with whether queries accurately reflect real-world queries or whether the relevance judgments are accurate. Assumptions of probabilistic retrieval models are often not correct, but are often close enough to correct to support a significant understanding of performance factors. For example, much of statistics assumes that data is normally distributed, although it is usually realized that this is only an approximation, and very little real-world data is exactly normally distributed. Statistical methods are valuable in estimating the parameters of the performance with specific realistic databases; analytic methods developed here are better at showing “why” certain methods are expected to be better than others, as well as showing that, in fact, one method is superior to another, on the average.

If one were to drive 50 miles per hour for two hours, one would expect to have travelled 100 miles. Experimentation is not needed to determine this; the basic laws of physics that we all learned early in our schooling are used in these computations. Most readers of the first sentence in this paragraph were able to determine the distance travelled in their heads, without resorting to the use of external computational devices. Simple analytic models are highly useful to professionals. However, determining exactly how far a particular car can travel given one liter of gasoline is peculiar to that car, and the average distance that the cars in a neighborhood travel is determined by statistical methods. The science of engine performance is determined through the laws of chemistry. Different methods are appropriate for different situations and their different data and assumptions. Below, we address a problem that is very amenable to analysis using analytic and probabilistic methods.

We propose here a method of determining the expected improvement associated with assigning a single metadata or index term. While this is a simplifying situation, the understanding of simple situations usually leads to a better understanding of more complex situations. When the assumptions of this model are approximated or met precisely, one can use this to more effectively index, retrieve or manage information systems. We consider examples of how adding a term numerically improves
results, and consider how this method may be used to economically analyze the assignment of metadata or index terms.

**INDEXING AND ASSIGNING METADATA**

The purpose of assigning a subject indicating term or metadata to a document is to make it easier to locate that document when a user searches using that term as the query. We assume here that a “term” is a set of characters, possibly with a space and including one or more “words,” representing one phrase or entity. Assigning a term to a document should result in moving the document either toward the front of the list of documents to be examined, leaving it at the same position, or moving it toward the back of the list. In cases where terms are assigned to documents incorrectly, the performance of locating documents may be expected to worsen, but if terms are assigned correctly, the order of the documents is expected to improve.

Indexing is the selection of labels from a set of existing set of possible labels. This set of possible labels may be all terms in a natural language, or a limited set of terms from a thesaurus, producing a controlled vocabulary. While machines may generate indexing representations that often represent a complex position in a machine developed term or topical space, most indexing systems use combinations of terms from a natural language, often consisting of a few terms (Foskett, 1996; Willis & Losee, 2013). Metadata may be used, like indexing, to represent topical aspects of documents, but metadata often expands beyond traditional indexing domains to representing data of all sorts, as well as capturing non-topical aspects of data, such as authorship, security status, date of generation of the data, etc. (Greenberg, 2010).

The manual assignment of index terms is often performed by people with a combination of subject expertise and professional knowledge of the library and information sciences (Chan, 2007; Cleveland & Cleveland, 2000). These assigned indexing and metadata terms serve a variety of functions (Bruce & Hillman, 2004; Moen, Stewart, & McClure, 1998; Stvilia & Gasser, 2008), although for the work here we are only concerned with the probabilistic nature of how such terms are used to label or not label documents about, or not about, the topic of the term and the query. Term characteristics, such as term frequency, correlate with usefulness, and measures such as Inverse Document Frequency weighting serve as effective term weighting and indexing indicators (Losee R. M., 1988). Terms and phrases may be assigned using algorithms, often implemented in software, that weight terms or assign terms in a manner that is felt by some to perform as well as intellectually assigned terms (Cleverdon, 1977; Salton & Lesk, 1968). Metadata about documents or data may be produced by software that is often used in the document production process or more directly by the human author or data producer (Greenberg, Spurgin, & Crystal, 2006).
CRITERIA FOR ASSIGNING INDEX TERMS OR METADATA

The assignment of topical terms has long been a goal of librarians, indexers, and metadata professionals. Cutter suggested over a century ago that catalogs should allow topical access to collections (Cutter, 1904), and modern library and information science scholars suggest that topical access should be available for electronic collections of documents, as well as other types of data collections (Greenberg, 2010). More recent developments of the Functional Requirements for Bibliographic Standards (IFLA, 2009) and the Dublin Core (DCMI, 2012) have provided structures that support topical access, as well as the less topical characteristics that also might be of use in retrieving or managing collections of data.

Topical terms have been extracted from text using several different methods. Based on information theoretic considerations, those terms that exhibit high signal to noise ratios may be selected as the best index terms. Inverse Document Frequency (IDF) weighting methods for automatic indexing have been suggested, weighting terms based on the relative rarity of terms in documents (Sparck Jones, 1972). Luhn suggested that mid-range term frequencies were a sign of good topicality for index terms, omitting very common terms and omitting very rare terms (Luhn, 1958). Research continues in the development of automatic term indexing and weighting as a part of developing document ranking methods for traditional retrieval systems as well as for search engines.

Rules for indexing are developed, but they do not directly address the question of how index terms or metadata contribute to improvements in accessing documents and data. The goal of a science is to describe, predict, and understand phenomena. Below we will suggest methods to describe and predict how index terms and metadata improve document ordering. Learning this measure describing how ordering is improved or harmed with index or metadata assignments can lead one to a fuller understanding of how indexing and metadata assignments actually perform what they are supposed to do.

A science has a methodology, a set of tools that are frequently used to advance the knowledge of the field. Combined with existing facts and measures or units that are the study of the discipline, one can study the assignment of terms as subject labels and how the presence or absence of labels changes the ordering of media for possible presentation and use by a user. Basic units considered below will include a searcher, a document, an expressed information need (a query), document relevance, a search length, a cost to a searcher, terms (index terms or metadata), and term assignments.

ORDERING DOCUMENTS

Documents are assumed to be of varying degrees of relevance to a potential user. To simplify, we assume here that relevance is binary, that is, a document is relevant or non-relevant to the user, or that there is a range of underlying relevance values, with those above a cutoff being labeled as “relevant” and those at or below this point being labeled as “non-relevant.” We also assume that
there is only a single index term being used, with this term being assigned or not assigned to a given document.

While the assumptions that terms are binary and that there is only a single term being considered are relatively straightforward, a third, more subtle assumption, is that the ordering with the single term is such that index term or metadata feature is a positive relevance discriminator that is more likely to occur in a relevant document than a non-relevant document. If the term is a negative relevance discriminator, the negation of this term would be used instead, since the negation would have a positive relevance discrimination value.

The assignment of indexing terms should make relevant documents easier to locate. This may be accomplished by ordering documents so as to place relevant documents near the front of the ordered list and to place the non-relevant documents further back on the list. One might view the science of indexing and metadata as the study of the placement of an individual document in a list of documents, given the index terms or metadata assigned to the document. The expected position of a relevant document (EPRD) in the ordered list of documents is computed as

$$EPRD = D[Q A + (1-Q)(1-A)] + \frac{1}{2}. \tag{1}$$

Here $D$ represents the number of documents, $Q$ represents the probability of optimal ranking, and $A$ is the normalized expected position of a relevant document. The average position of a relevant document, which is also the average search length ($ASL$), is the number of documents that would be examined up to the average position of a relevant document (Church, 2010; Losee R. M., 1998). An $ASL$ or $EPRD$ of 10 would mean that one would have to look at 10 documents, on the average, if one were trying to examine a randomly selected relevant document in the ordered list of documents or data. This model assumes that one wants to examine all the relevant documents, and that the average position of all the relevant documents is an adequate surrogate for the positions of all the relevant documents. If there appear to be too many documents, one might raise the relevance standards so that fewer documents are labeled relevant, or other measures might be useful (Church, 2010). Comparing different positions, as computed by EPRD, is easiest if the values of retrieving different numbers of relevant documents are linearly related.

The variable $A$ is the expected position of a relevant document, where the position in the set of documents is scaled from 0 to 1, with a low value for $A$, near 0, representing all relevant documents at the beginning of the list, and a high $A$, approaching 1, representing all relevant documents at the end of the list. Here, $A = (1-p+t)/2$, where $p$ is the probability that a relevant document is assigned the index term in question, and $t$ is the probability that a document (regardless of relevance) is assigned the index term. As an example, consider the hypothetical example where $t = 1$, when all documents have the term, and $p=0$, when no relevant documents have the term. In this hypothetical situation, $A=1$. In the best case hypothetical position, if $p=1$ and $t=0$, then $A=0$. Note that this scenario is impossible, since if all relevant documents have the term, then some of the set of other documents must have the term, so $t$ could not be 0. Note a similar problem with the earlier example. When $p = t$, for example, when $p = t = .3$, then $A=1/2$, indicating that the expected position of a relevant document
is about half way through the ordered list of documents. This is because when \( p=t \), the ordering is essentially random because the term does not discriminate between relevant and non-relevant documents or data.

The EPRD may be used *retrospectively* to measure the results of a search that has already been conducted. In this case, the parameters described in this article are computed directly from the ordered documents. In *predictive* situations, in which one attempts to predict the performance of future orderings, the parameters for future orderings may be estimated based upon other considerations (Bookstein, 1983; Church, 2010; Losee R. M., 1988; Manning, Raghavan, & Schutze, 2008; Zhou & Croft, 2008). However, when using a simulation such as the one described below, we choose to use the exact parameter values, the values that would be computed in many retrospective situations, when measuring performance. These are the “correct” values and contain no error in describing this particular dataset. For the purpose of this study, to avoid issues about what parameter estimation technique is best, in all cases we use the correct (retrospective) values for each parameter.

The \( Q \) component of ordering represents the quality of the ordering algorithm, more precisely, the probability of optimal ranking. The optimal ranking is the best ranking that can be achieved, given the indexing and metadata term assignments. Optimal ranking is not understood here to be placing all the relevant documents first, followed by all the non-relevant documents, as this can only be accomplished with the use of an all-knowing oracle. Instead, the optimal ranking is placing those documents first that have the terms most likely to be found in relevant documents, followed by those documents that do not have the search term. When there is only one term being considered, there are two possible ordering with binary features: placing the documents with the feature first, followed by documents without the feature, and the ranking of documents without the feature first, followed by documents with the feature. This is a function of the search engine or ranking algorithm (Church, 2010; Losee R. M., 1998), not a function of the intrinsic difficulty associated with locating specific documents given specific queries, which is what \( A \) captures (Losee & Paris, 1999).

In Equation 1, there are two major components. The first, the \( Q \) times \( A \) component, represents the probability that the ranking is optimal times the normalized ranking provided by \( A \), where, because the term is a positive discriminator, this is the best ordering that can be obtained for documents, given this single term. This is added to the second component, \((1-Q) \) times \((1-A)\), which represents the probability that the ranking is non-optimal (which is worst case ranking when the single feature is binary) times the expected position of a relevant document in a worst case ordering.

To simplify the discussion, one might assume that the quality of the ranking is 1, that is, the quality of the ranking algorithm is perfect. Assuming this, the expected position of a relevant document may be estimated as \((D \ A) +1/2\). This may be viewed as taking the \( A \) normalized position and multiplying it by the number of documents.

The above work may be parameterized to address the numbers of documents of various types. We denote the number of relevant documents with the term as \( r \), the number of relevant documents as \( R \),
the number of non-relevant documents with the term as \( n \), and the number of non-relevant documents as \( N \). Given these parameters, we may compute \( p = r/R \) and \( t = (r+n)/(R+N) \). If there were 10 relevant documents \( (R) \), 5 with the term \( (r) \), then \( p = 5/10 \). If we add to this 100 non-relevant documents \( (N) \), 20 of them with the term \( (n) \), then \( t = (5+20)/(10+100) \).

Assuming optimal ranking \( (Q=1) \), the expected position of a relevant document becomes:

\[
EPRD = (R+N)[(1 - (r/R) + (r+n)/(R+N))/2] + 1/2
\]

where \( p = (r/R) \) and \( t = (r+n)/(R+N) \).

**SHOULD A DOCUMENT BE INDEXED BY A TERM?**

The quality of indexing may be studied by noting the retrospective retrieval performance produced by experiments or quasi-experiments. This Cranfield paradigm has been widely used to evaluate indexing, ranking, and interfaces (Cleverdon, 1977). In addition to gathering data showing what performance results from changes in the retrieval system characteristics, one can determine analytically the performance or expected performance that will be obtained with specific changes in systems. Such analytic methods can be compared to existing retrospective results, or can be used to predict future expected performance.

One simple analytic form can provide an answer to the question, “when should a document or data item have a specific indexing or metadata label?” A simple answer to this question is that the document should be given the label only when it improves ranking performance.

Ranking performance with labels can be measured using the expected position of a relevant document \( (EPRD) \), as given in Equations 1 or 2. Using these, one might compute whether a document should be indexed as occurring only when the EPRD with an indexing term applied is better (lower) than the EPRD without the indexing term being applied.

A weakness of this approach to an indexing rule is that it is dependent upon a specific single number measure of retrieval performance. Single number performance measures have the advantage that they are easily compared, with the more desired or less desired value being obvious, whereas multiple different measures, such as precision-recall, are difficult to compare. While different performance measures address different concerns (Losee R. M., 2000), often there is a large amount of agreement between the measures. However, if one uses a measure that is analytically tractable, such as EPRD, more sophisticated techniques can be developed, such as the decision rule that indexes based on improving the EPRD. By knowing certain parameters, the values for \( R, N, r, \) and \( n \), both with and without indexing, one can determine whether performance is improved by an indexing operation. Furthermore, we consider below exactly how much improvement occurs with different indexing actions.
FOUR CATEGORIES OF INDEXING

We can now consider more formally the consequences of various indexing or metadata assignment actions. To do this, one can study the improvement in $EPRD$ given actions, such as a term being assigned or not assigned, or different states of nature, the document being relevant or not relevant. We compute the Indexing and Metadata Advantage (IMA) value, the difference in $EPRD$, so that a positive value represents improvement, movement toward the beginning of the ordered list of documents (a lower $EPRD$), and a negative IMA value represents the amount of movement toward the bad end of the list, away from the searcher’s starting position. We are considering only a single term being assigned or not assigned, and treat the query as having the single term present or absent. For this section, we also assume that $Q = 1$.

RELEVANT DOCUMENT ASSIGNED THE TERM

When a relevant document in our collection is assigned the index term in question, with the document not having been assigned the term previously, the value of $r$ increases by 1, $R$ stays the same, and the $EPRD$ (using Equation 2) decreases from $(R+N)[(1 – (r/R) + (r+n)/(R+N))/2] + 1/2$ to $(R+N)[(1 – ((r+1)/R) + ((r+1)+n)/(R+N))/2] + 1/2$. Note that the parameters are for the initial state and not for the final state, which will often have 1 added or subtracted from the initial value. If we subtract the second from the first, the difference between them becomes $+(N/R)/2$. This positive Indexing and Metadata Advantage is the performance improvement, measured in documents, the decrease in the $EPRD$, that is obtained by making a particular indexing or metadata assignment decision.

Consider an example where there are 10 relevant documents and there are 90 non-relevant documents. If there are 5 relevant documents with the index term, and this is increased to 6, the $EPRD$ is decreased by $(90)/10/2 = 9/2$. Consider a similar situation where there are 900 non-relevant documents and 100 relevant documents. In this situation where the term in question is added to a relevant document, the improvement, or decrease in $EPRD$, is $(900/100)/2 = 9/2$. We can conclude from the formula and this example that when a database is scaled in size so that both the numbers of non-relevant and relevant documents increase at the same rate, the IMA remains the same.

An informal explanation of the IMA of $(N/R)/2$ is as follows. There are always $R+N$ documents. When most of the documents in a database are not relevant to a given query, such as in a large database or a large library, one may approximate the number of documents in the database as $N$. We may split these documents into $R$ groups. Given that randomly distributed relevant documents are expected to be located throughout an ordered list of documents, one can measure the size of these segments and divide by two, moving $N/R$ halves toward the front of the ordered dataset.

RELEVANT DOCUMENT UNASSIGNED THE TERM

When a relevant document with the term is unassigned the index term in question, the value of $r$ decreases by 1, $R$ stays the same, and the $EPRD$ increases from $(R+N)[(1 – (r/R) + (r+n)/(R+N))/2] + 1/2$ to $(R+N)[(1 – ((r-1)/R) + ((r-1)+n)/(R+N))/2] + 1/2$, producing an IMA of $-(N/R)/2$. Thus, unlike assigning a relevant term to a relevant document, where the document moves toward the front of the ordered
list of documents, removing such a term moves that document away from the front of the ordered list, with the amount of movement toward the back of the list \((-\frac{N}{R})/2\) being exactly the same amount or IMA as assigning a relevant term to a relevant document, but moving in the opposite direction.

**NON-RELEVANT DOCUMENT CORRECTLY NOT ASSIGNED THE TERM**

Using similar formulas as with the above examples, correctly removing the term from a non-relevant document will clearly improve performance. In this case, using the formulas above, where we subtract the modified EPRD from the original EPRD, \(n\) is decreased by 1, resulting in the \(EPRD\) changing by an IMA of \(+1/2\).

**NON-RELEVANT DOCUMENT INCORRECTLY ASSIGNED THE TERM**

When a non-relevant document is incorrectly assigned the term in question, the \(EPRD\) decreases by an IMA of \(-1/2\), moving the \(EPRD\) \(1/2\) of a position toward the back of the ordered list of documents.

Using these methods, the Indexing and Metadata Advantage factors can be computed for the set of possible actions and relevance states. These formulas are given in Table 1a.

**OTHER ASSUMPTIONS FOR MODELS OF IMPROVEMENT**

Above, we assumed optimal ranking, a \(Q\) of 1, so that the only factors that determine the quality of the indexing are the parameters data about the data or documents and the assigned relevance judgments. If we assume that \(Q\) is less than 1, then the probability of the optimal ranking becomes a factor in determining the \(EPRD\) and the \(IMA\). Table 1a shows the IMA formulas that apply in the general case where \(Q\) may be any value ranging from 1 down to 0.5.

Other single number measures of performance may be used besides EPRD. For example, one may also compute the percent of performance improvement (Losee R. M., 2007) provided by the indexing choice by computing the percent of performance improvement (between randomness and the best possible performance) before and after the term is added or removed. While the computation of this percent is possible and has been performed by the author, it adds little to illustrating the strengths and weaknesses of adding or deleting metadata that isn’t illustrated by using the methods described above. This may be the case for many other single number measures other than the EPRD.

In some cases, metadata may not be binary, and one may wish to capture data values that are, for example, normally distributed or Poisson distributed (Losee R. M., 1998). Let us assume here that \(Q=1\) and that the average values assigned to the metadata are normally distributed with a mean of 0 and a standard deviation (SD) of 1. If the relevant material has an average value of 1, then the \(A\) value becomes the sum of all the values of the normal distribution with mean 0 and SD of 1 from 1 (the average for relevant documents) up to infinity. This is sometimes referred to as a survival function. In this case, the \(A\) value is numerically 0.16. If we have 100 documents, the \(EPRD\) is
approximately 16. If the data were changed or measured again so that the relevant documents had an average of 2, while the average for all data stayed the same, the EPRD falls to 2.5. Similar techniques are available for features with other distributions. If users were to assign tags or vocabulary features to a document, with users assigning them independently, a Poisson model might be appropriate for this situation, allowing the EPRD of such a term to be computed. The formula associated with computing $A$ for the Poisson distribution, as well as other distributions, is given in Losee (1998, Chapter 6).

We can see here a different model of indexing or assigning metadata. With binary indexing or assigning of terms, people often describe the choice of assigning a term or not. The alternative model is very mechanistic, where data is produced in a certain way, taking on a range of values. When normally distributed values are collected, the data values and the collection mechanism in part determines the Indexing and Metadata Advantage, just as an automatic indexing or metadata generation algorithm has various qualities and in part determines the IMA.

The economic cost associated with the EPRD can be approximated if we assume that there is a fixed cost associated with each non-relevant document the user chooses to examine (Cooper, 1968). Assume that the number of non-relevant documents examined is proportional to the EPRD minus the number of relevant documents one would expect to find before examining the EPRD. If the number of relevant documents, $R$, is much smaller than the number of non-relevant documents, then the $R$ component may be dropped if we only need to approximate the cost associated with the EPRD. If we compute the IMA as the difference of the EPRD values, the $R$ constants would be expected to largely cancel out, approximating the IMA as discussed above. Thus, the Indexing and Metadata Advantage may serve as an approximation of the cost improvement of assigning an indexing or metadata term. Managing retrieval and indexing choices with a goal of minimizing the EPRD is clearly one method of approximating the decrease or minimization of the expected cost of retrieval.

**AN EXAMPLE**

An example of the application of this type of technique is provided in Table 1. The top part, Table 1a, shows the formulas for IMA, the increase or decrease in performance due to adding or removing an index or metadata phrase or term. The first line of the table, shows the IMA formulas when $Q=1$, that is, when the ranking algorithm is doing the best possible. This may not be perfect ranking, as the ranking algorithm will not usually know what the user considers relevance; instead, it will rank documents based on the features assigned.

The second row of Table 1a shows the IMA for cases where the ranking is not optimal, and thus $Q$ is less than 1. Note that this second line in Table 1a is similar in some respects to the first line, where $Q=1$.

Table 1b shows sample numeric values for IMA that represent the improvement associated with retrieving a set of documents where there are 20 relevant and 100 non-relevant documents, and 12 of
the relevant documents have the term in question, and 4 of the non-relevant documents have the
term in question. Considering the data in the Table, when Q=1, the first row in the numeric example
shows that, for a realistic situation where there are many more non-relevant documents than relevant
documents, terms added correctly or incorrectly to relevant documents have a much bigger impact
than terms added to non-relevant documents. When Q=0.8, the IMA values are all smaller, as are the
costs.

Empirical Validation of the Model

The performance of the model may be empirically validated a number of ways. One is to prove the
basic formulas a number of different ways, for example, algebraically, geometrically, using number
theory, etc. Other methods of validation have exhaustively tried all the possible situations that
might occur, or those most likely to show the desired characteristics. For example, the four color
problem suggests that for any map drawn on a two dimensional surface, such as a sheet of paper or a
computer screen, four colors are enough for coloring a map, so that each area has one color and there
will be no adjacent areas of the same color (Appel & Haken, 1977). By exhaustively trying all the
possible color combinations, consistent with certain graph theoretic considerations, Appel and Haken
were able to show that 4 colors would always be sufficient. No other formal proof was available at
the time that Appel and Haken conducted their work. It was clearly valuable to know that
exhaustively trying all possibilities resulted in no cases in which more than 4 colors were needed. It
raised a further question, whether research such as this constituted a “proof.” We assume that our
own empirical validation work provides added support for the correctness of the model, but do not
expect the reader to treat any validating software to constitute a proof, but rather support for an
argument.

Many other conjectures in mathematics have been studied by trying out large numbers of possible
values to see if conflicts with the conjectures can be found. These computer analyses can study many
questions, such as whether the digits in the mathematical constant Pi are randomly distributed, or
many questions about the distribution of prime numbers, the size and locations of gaps between
prime numbers, and so forth. These studies may begin at a starting point, usually the beginning of a
sequence, such as 0 or 1, with software then examining all the items of interest from the starting point
up to a certain finishing point, or one may randomly select many examples from a wide range of
possible examples.

Below, we will follow the approach of starting at the beginning and examining all possible orderings
and the associated indexing options that occur for less than a cutoff number of documents.
Traditional experimental techniques using sample databases may show the performance with some
realistic documents and queries. The goal here is different: we attempt to lead to an understanding of
the effect on ordering of one term in one document, and no more. Using a retrieval test database
such as TREC to study the IMA model would require that documents be ranked by the presence of
the single term in question, with the terms being added or removed iteratively in all documents, with
the resulting ordering’s empirical performance being compared with the predicted ordering’s performance. This could be done for all terms. In this situation, one is no longer using the experimental database itself, but permutations of the database.

A simulation, using the exhaustive evaluation of all sets of all possible documents within a particular size range, with the documents consistent with the stated assumptions, including the assumption of a single term, is assumed by the author to be superior to using a traditional experimental retrieval testing techniques, manipulated as suggested above, because all possible orderings are considered in the simulation’s exhaustive evaluation, rather than just a relative few that occur naturally in an experimental database, along with the permutations simulated and associated with the experimental database. Using all additions and deletions of possible terms with an experimental database is essentially the same as our simulation model, where we iterate through all the possible numbers of documents up to a certain number of documents.

Empirical validation software has been developed to exhaustively evaluate the term additions and deletions described above for all simulated document orderings of any size. These were tested for all simulated orderings of 14 and fewer documents, but other values can be tested. While we are not making a formal proof, we are testing to see if the predictions offered by using the formulas in Table 1 always produce the same document ordering performance results that are empirically found to occur. The software shows 100% agreement between predictions (where $Q=1$) and the actual rankings. For example, in the case of evaluating all orderings of documents, from 1 to 14 documents, the IMA software found that 2328410 predictions were accurate, and 0 were wrong. Manually changing the formulas in the software results in differences being produced between actual ranking performance and the predicted performance, so that far less than 100% of the predictions are accurate. The software, available at [http://ils.unc.edu/~losee/ima](http://ils.unc.edu/~losee/ima) is a gawk (gnu awk) program that will run on most unix or linux systems. This software may be downloaded and executed to provide support for the validity of this model. Options and documentation are given in the code. Note that the code is optimized for readability, not for execution speed. When considering all sets of documents up to 14 documents, there are many different valid orderings, but some of these may be duplicates. Note that when actual and predicted performance figures are compared, because of rounding errors that occur in computations, performance numbers are rounded to their nearest thousandth before being compared. In this simulation, predictions either match or do not match the empirical ordering so that there are no issues of statistical significance in the simulation.

The major objection to using this form of argumentation is that it may miss examples that contradict the theorem being tested. For example, if assigning metadata labels acts differently at some point above the cutoff of 14 documents used here, say, for 20 documents, the empirical validation software and conclusion would be invalid. However, we believe that if these predictive formulas are shown to be valid for all orderings of 8 or fewer documents, they are likely to be valid for all numbers of documents. The software makes several assumptions, described earlier. These include that terms are binary, that relevance is binary, that there is a single term, and that term presence denotes that the term is a positive discriminator.
A Decision Rule for Assigning Index or Metadata Terms

Criteria for assigning indexing or metadata terms can be developed based upon the use of Indexing and Metadata Advantage measures. Each IMA represents a cost or benefit of making a particular decision, such as choosing to label or not label, in the context of a given state of nature, that is, the document or data is relevant or is not relevant. By computing the expected IMA associated with assigning a label and comparing it to the expected IMA associated with not assigning a label, one can rationally choose to assign an index or metadata term if the expected IMA associated with assigning the label is greater (more beneficial) than the expected IMA associated with the act of not assigning the label. The expected IMA or cost of assigning labels is the probability of assigning the label in question to a relevant document multiplied by the IMA associated with labeling a relevant document with the term in question, added to the probability of assigning the label to a non-relevant document multiplied by the IMA associated with assigning the label to a non-relevant document. A similar expected cost may be computed for not labeling a document. By comparing these two values, one may index or assign metadata terms when the expected IMA associated with labeling is superior to the expected IMA associated with not labeling.

The probabilities used in computing the expected IMA require the accumulation of data on what was relevant or non-relevant for each term. Beginning with the data, a trained indexer may develop accurate estimates of the probabilities to be used in making these estimates (Cooper, 1978).

The Use of IMA by Information Professionals

Information professionals can benefit greatly from an understanding of how metadata and indexing terms contribute to the improvement of access to the documents or data. Indexers with a greater understanding of the variables associated with their own work are likely to produce superior indexing and metadata tags. Knowing the linkage between the process and a specific output often results in an understanding of how to produce useful labels, and possibly the best indexing and metadata, avoiding many mistakes. An understanding in science is often beneficial to the scientists, as well as to the users of the systems that they develop and support. For example, accurate knowledge about how the human body works allows medical practitioners to best treat patients. Those believing in bloodletting as a treatment for many symptoms harmed many medieval patients, whereas modern understandings of the body have led to improved treatments with improved survival. Knowing the precise ordering effect of an index term or metadata assignment decision helps improve the chances that the decision will increase the chances of the desired goal occurring.

Being able to simplify the IMA model, such as assuming that \( Q=1 \), may simplify the analysis of the problem and result in much simpler considerations under appropriate circumstances. The choice of the appropriate model can result in increasing accuracy for users and decreasing the cost of assigning metadata and index terms.
Searchers and subject specialists can benefit from knowledge of how indexers assign subject metadata by considering whether their own knowledge about the particulars of their situation is different from the probable situation perceived by the indexer. Knowing that the term being used for searching is a new use of the term probably not known to a general indexer, or that the term itself is ambiguous, may help the searcher determine the likely quality of the indexing, suggesting that the term would be the best to search with or perhaps that another term would be more likely to produce the best results.

Those who manage indexing services or retrieval systems may benefit from knowing the relationships between the decisions made and the expected outcomes produced based upon those choices. For example, given a choice between a good retrieval system and a more expensive but excellent retrieval system, what will be the average impact on future searchers? Would most managers spend 50% more for a system if it reduced the EPRD for users by 2 documents, on the average? How much time will be saved for the searchers? What rules and practices should be supported by managers?

Those who develop vocabularies and metadata systems can benefit from considering the consequences of their vocabulary inclusion decisions. One simple decision to be made is the number of metadata or vocabulary terms to use in the system. By increasing the range of values for the parameters \( r \) and \( n \) and by spreading the choice of vocabulary terms across a larger set of terms, greater or lesser specificity or exhaustivity can be achieved.

**CONCLUSION AND SUMMARY**

This research provides a model that improves our scientific understanding of indexing, metadata, and the performance associated with their application. This approach can be used retrospectively, to analyze existing data and the past effect of assigning metadata. Such scientific principles can also be used to predict accurately the effect of future actions. These principles explain how assigning some metadata helps and why assigning poor metadata hurts. By formalizing the increase or decrease in ordering performance when an index term is applied, as relevant documents move toward or away from the searcher in an ordered list of documents or other metadata carriers, information professionals can better understand the utility of indexing and metadata.

The gain associated with assigning a term to a document or deciding not to assign a term is that the document moves from one group of documents to another. Documents move from the set of relevant or non-relevant documents, with or without the term, to another set. One can clearly expand this model, to move beyond binary relevance, for example, or beyond binary metadata term assignment. Other extensions exist, such as assuming a continuous value for data or term assignment weights or multiple terms (Losee R. M., 1998). Some of these are simple, and others complex enough so as to add little to an understanding of the effects of metadata assignment. This is certainly an area for future research and development. Future work may address the use of multiple terms in a document.
An advantage of using such models of indexing and metadata assignment is that one can analytically determine the maximum, average, and the minimum performance obtained with label assignments. This has potential economic utility in deciding whether further improvement or harm is possible, as well as determining whether the benefit of moving one or more documents a certain number of documents toward the searcher would be worth the cost. Using these principles, one can develop predictions, often taking the form of a decision rule, suggesting that in certain situations a term should be assigned, such as when the term is expected to improve system use more than the cost associated with assigning the term, or a rule may be used to suggest not assigning a term.
<table>
<thead>
<tr>
<th>TABLE 1: Indexing and Metadata Advantage (IMA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE 1a</td>
</tr>
<tr>
<td>IMA Changes in Values for Expected Position of Relevant Document</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Positive IMA values represent improvements (decreased EPRD)</th>
<th>Assignment of term for relevant document</th>
<th>Relevant document not assigned term</th>
<th>Non-relevant document removing assigned term</th>
<th>Non-relevant document incorrectly assigned term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Ranking, Q = 1</td>
<td>(N/R)/2</td>
<td>-(N/R)/2</td>
<td>+1/2</td>
<td>-1/2</td>
</tr>
<tr>
<td>Q &lt; 1</td>
<td>(1/2)(2NQ - N)/R</td>
<td>-(1/2)(2NQ - N)/R</td>
<td>Q - 1/2</td>
<td>-Q + 1/2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 1b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing and Metadata Advantage with N=100, R=20, n=4, r=12</td>
</tr>
</tbody>
</table>

| Q=1               | 5/2       | -5/2     | +1/2     | -1/2     |
| Q=0.8             | 3/2       | -3/2     | +3/10    | -3/10    |

**WORKS CITED**


