



PERGAMON

Information Processing and Management 37 (2001) 255–277

**INFORMATION
PROCESSING
&
MANAGEMENT**

www.elsevier.com/locate/infoproman

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

James D. Anderson ^{*}, José Pérez-Carballo

*School of Communication, Information, and Library Studies, Rutgers The State University of New Jersey,
4 Huntington Street, New Brunswick, NJ 08901-1071, USA*

Received 4 February 2000; accepted 2 May 2000

Abstract

Does human intellectual indexing have a continuing role to play in the face of increasingly sophisticated automatic indexing techniques? In this two-part essay, a computer scientist and long-time TREC participant (Pérez-Carballo) and a practitioner and teacher of human cataloging and indexing (Anderson) pursue this question by reviewing the opinions and research of leading experts on both sides of this divide. We conclude that human analysis should be used on a much more selective basis, and we offer suggestions on how these two types indexing might be allocated to best advantage. Part I of the essay critiques the comparative research, then explores the nature of human analysis of messages or texts and efforts to formulate rules to make human practice more rigorous and predictable. We find that research comparing human versus automatic approaches has done little to change strongly held beliefs, in large part because many associated variables have not been isolated or controlled.

Part II focuses on current methods in automatic indexing, its gradual adoption by major indexing and abstracting services, and ways for allocating human and machine approaches. Overall, we conclude that both approaches to indexing have been found to be effective by researchers and searchers, each with particular advantages and disadvantages. However, automatic indexing has the over-arching advantage of decreasing cost, as human indexing becomes ever more expensive. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Automatic indexing; Human indexing; Indexing and abstracting services; Allocation of indexing effort

^{*} Corresponding author. Tel.: +1-732-932-7501; fax: +1-732-932-6916.
E-mail address: jda@scils.rutgers.edu (J.D. Anderson).

1. Automatic indexing

We now turn from human indexing (addressed in Part I of this essay) to machine indexing, or the analysis of text by means of computer algorithms. Here the focus is on automatic methods used behind the scenes with little or no input from individual searchers, with the exception of relevance feedback. Searching options and techniques, such as methods for creating effective search statements, adding weights to terms, specifying proximity requirements, using truncation, wild cards, or combining terms with boolean or role operators, are considered to be part of searching syntax and are outside the scope of this essay.

Operational automatic indexing has thus far been restricted, for the most part, to language text. Research into ways to apply automatic techniques to image, sound, and other types of text is still in its infancy, compared to the half century of work on automatic indexing of language text. Altavista, the well-known web search engine (<http://www.altavista.com>), began to offer a search option in late 1998 that attempts to locate images that are “visually similar” to a retrieved image. “Visually similar” is not the same as “conceptually similar”, so that the results often appear to be based on color and patterns, rather than on particular objects or activities portrayed.

Throughout the history of automatic indexing, two major theoretical models have emerged: the vector-space model and the probabilistic model. Sparck Jones, Walker, and Robertson (2000) have provided a thorough review of the development, versions, results, and current status of the probabilistic model. In comparing this model to others, they conclude that “by far the best-developed non-probabilistic view of IR is the vector-space model (VSM), most famously embodied in the SMART system (Salton, 1975; Salton & McGill, 1983a). In some respects the basic logic of the VSM is common to many other approaches, including our own [i.e., the probabilistic model] . . . In practice the difference [between these two models] has become somewhat blurred. Each approach has borrowed ideas from the other, and to some extent the original motivations have become disguised by the process. . . . This mutual learning is reflected in the results of successive round[s] of TREC. . . . It may be argued that the performance differences that do appear have more to do with choices of the device set used, and detailed matters of implementation, than with foundational differences of approach” (part 2, pp. 829–830).

The focus of our discussion will be on the automatic indexing of language texts. The various tactics and strategies are emphasized, rather than the underlying theoretical models. Useful background and further detail can be found in Croft (1989), Harman (1994), Korfhage (1997), Kowalski (1997), Salton (1989) and Sparck Jones and Willett (1997).

2. In the beginning was the word

Automatic indexing of language text has traditionally begun with individual words. The first step is to decide what constitutes a word. The usual definition is one or more characters separated by spaces or punctuation¹. The hard part in this definition (and the part in which systems vary,

¹ Defining words on the basis of spaces and punctuation works for most alphabetic or syllabic writing systems, but in Chinese, for example, where each character represents a morpheme or syllable rather than a “word” as generally understood, the determination of word boundaries is much more difficult and is similar to attempts to identify multi-word phrases in English. Several participants in the Text REtrieval Conferences (TREC), especially TREC 5 and TREC 6, have worked on this problem in relation to the indexing and retrieval of Chinese text (TREC, 1992–1999).

leading to very different results) is how to deal with punctuation. When are marks of punctuation part of a word (as in 501(c)(3) or A.L.A.)? Some researchers are experimenting with procedures that bypass words all together, by simply matching sequences of characters – for example, all sequences of three, four, or five characters, without regard to spaces or punctuation (Mayfield & McNamee, 1998; Huffman, 1995; Cavnar, 1994).

The most troublesome punctuation marks are the hyphen, the slash, and occasionally the apostrophe, the comma, and the period or full stop. Parentheses and underscores can also be significant. Our discussion will focus on the English language, so we will ignore the problem of diacritical marks or accents, but these are also problematic, because in some languages, they can distinguish between very different words, such as “ano” (anus) and “año” (year) in Spanish.

Hyphens are tricky because they can be used to connect two separate words, as in “full-text”. For example, we use “full text” when text is used as a noun, but we link these two words with a hyphen when they are used together as an adjective, as in “full-text database”. Hyphens are also used to create compound words, such as “data-base”. In English, new compound words often lose their hyphens over time, so that “data base” became “data-base,” which evolved into today’s “database”. Some automatic indexing algorithms treat the hyphen as a space, so that the characters before and after the hyphen become separate words (“on-line” becomes “on” and “line”!). Some systems ignore the hyphen, treating it as nothing, so that “MS-DOS” becomes “MSDOS” and “full-text” becomes “fulltext”. One solution is to use all possible combinations, so that “on-line” would become “on” and “line”, “online”, and “on-line”.

The same approach can be used with the slash, which is sometimes used as an integral part of a word or acronym, as in “OS/2”, but perhaps more often is used to combine two different, often contrasting or complementary words, as “high/low” or “lesbian/gay”. If all possible combinations are used, then we would have “lesbian”, “gay”, “lesbiangay”, and “lesbian/gay”. Slashes are used in universal resource locators (URLs) for world-wide web documents, so they should not be ignored.

Like slashes, underscores and full stops have become common elements within email addresses and URLs. Full stops are also common in initialisms. If character sequences containing full stops (without following spaces) are treated in all possible ways, then “A.L.A.” would also be treated as “ALA” and “A”, “L”, and “A”.

Parentheses can also appear within various types of words, but even more commonly in section headings, such as “501(c)(3)”, the provision in the US Internal Revenue Code that provides for tax-deductible contributions to many non-profit organizations. As a result, many organizations are called “501(c)(3)” organizations, and that sequence of characters can be an important searchable word. A possible solution is to treat parentheses as parts of words if there is no space to the left of a left parenthesis or to the right of a right parenthesis.

If the apostrophe is ignored, “can’t” would become “cant” and “I’ll” would become “Ill”. Here too, perhaps the best solution is to include all possibilities. In syntactic analysis, the difference between “its” and “it’s” may be important.

“501(c)(3)” also illustrates the importance of including numbers as words. And if numbers are included as words, then internal commas and full stops become important. It might be O.K. to treat “1,998” as the same as “1998” (possibly confusing an integer number with a date), but there is a big difference between “.103” and “103”.

However, including numbers as indexable words may have an important negative impact as well. Harman (1994, p. 250) points out that there are an unlimited number of unique numbers, so that including them can cause the number of unique words to “explode”, slowing down the whole indexing process. The role of numbers will vary among different types of texts.

One possible solution would be to index only numbers that are combined with alphabetic letters, such as our “501(c)(3)” example. But that would make it impossible to search for tax forms like “1040” or “1099” or a movie like “2001”! Despite possible problems, the best solution appears to be to include all numbers and number–letter combinations within the definition of words to be indexed. In many automatic indexing systems, words that occur only once are eliminated from the index, and because many numbers occur only once, they will be eliminated in any case. But common numbers, such as “1999”, “2000” and “2001” would be retained.

A similar issue is whether to include single character words. The most common single character words in English are “I” and “a”, and in many indexing procedures, these words end up on a stop list. Not to include single character words will make it hard to find “vitamin C”! If single characters are included, but “a” is on the stop list, how can “vitamin A” be found? Some of these problems can be eliminated through a careful sequencing of analysis steps. For example, if phrases are identified first, then “vitamin A” would be considered as a phrase rather than two separate words. On the other hand, if stemming is performed first, some words may be reduced to forms that are on a stop list.

All these examples are evidence that defining a word for automatic indexing (and machine matching) is not a simple and straightforward task. These problems have been discussed by Brooks (1998), who gives many more examples and cites relevant literature.

Upper- and lower-case letters can also be problematic. Upper- and lower-case letters are usually merged for calculating term frequency, but retaining upper-case letters can be useful for identifying proper nouns such as names of persons, organizations, and countries. But certainly, searchers should not have to worry about whether particular terms may have upper or lower-case letters.

3. Simple keyword indexing

The simplest automatic indexing is based on providing access to every occurrence of every word. In displayed indexes (browsable by humans), this kind of indexing can be presented to users in formats known as permuted, keyword-in-context (KWIC), keyword-out-of-context (KWOC), keyword-along-side-of-context (KWAC) and similar syntactic designs. For machine matching, this is straightforward free-text, full-text indexing and retrieval, the kind common in many word processing programs. It will find every occurrence of every character or combination of characters, even the most insignificant of words (from the topical stand point), such as “an” or “the”.

4. Negative vocabulary control

The first attempt to improve simple keyword indexing is often to use a stop list of insignificant words, designed to eliminate indexing and retrieval of words like “an” and “the”. Eliminating

stop words can reduce the size of the index significantly, and speed up processing. Francis, Kučera, and Mackie (1982) suggest that the 10 most frequently used words in English can account for 20–30% of the words in a text.

The obvious words for a stop list are prepositions, articles, and conjunctions. Beyond that, expanding the stop list can be problematic, because one person's meaningless word is another person's essential word. "Aspect", for example, might be a good candidate for a stop list, but then "aspect" is an important attribute of verbs in the Russian language, so it could be an important keyword in the context of Russian grammar or linguistics. Some electronic databases have dropped "a" as a stop word because of topics like "vitamin A".

Some automatic indexing systems use stop lists of several hundred words, but to be on the safe side, some IR databases have reduced their stop words to as few as eight: "and", "an", "by", "from", "of", "or", "the", and "with". According to Harman (1994, p. 252) the MEDical Literature Analysis and Retrieval System (MEDLARS) of the U.S. National Library of Medicine has even fewer stop words.

Using a stop list can be considered negative vocabulary control. The usual approach to vocabulary control is to list the terms that *may* be used (preferred terms). In this case, exactly the opposite tack is used: the terms that may *not* be used are listed.

5. Counting words

It quickly became apparent in the early days of automatic indexing that the simple occurrence of a word did little to indicate the theme, meaning or purpose of a language text. So programs began counting words in texts, hoping that the term frequency would better indicate what is important in a message. Of course, the most meaningless words – the function words like conjunctions, articles, and prepositions – are also the most frequently occurring words, and these, as already suggested, are not very helpful. But if these words have been eliminated by placing them on a stop list, then counting words can provide a criterion for ranking likely candidate texts in response to a keyword search.

6. Comparative counting and weighting

Next came the realization that sometimes term frequency (TF) within documents does not help much in distinguishing one text from another within a single collection or IR database. Take librarianship, for example. The word "library" will probably occur in most if not all texts in a collection or IR database on librarianship, so the mere fact that it occurs in a text does not tell us very much. So how about comparing the frequency count in single texts with the overall occurrence for the same words in an entire collection or IR database? This way we can identify words that are *unusually* frequent in a particular text – words that occur frequently in some texts but do not occur frequently across the entire collection. This relative frequency could be more useful in finding useful documents than simple word frequency within documents. "Inverse document frequency (IDF)" is the measure used to indicate the frequency of terms across documents in the collection. The fewer the documents that have a term, the higher the IDF score for that term. The

IDF score can be combined with term frequency (TF) within particular documents to help identify useful documents.

Term statistics such as IDF and TF have been combined into more or less complex mathematical formulas in order to compute a weight for each of the terms that appear in a document. This weight is then used to compute a score for the whole document in relation to a particular query or search statement. The essential point of these formulas is to assign weights to terms in a document according to how good these terms are for distinguishing among documents. For example, as noted, terms that appear in almost all documents in the IR database are not as good as terms that appear frequently within some documents but not in the whole corpus.

Some of these formulas do not have a very strong theoretical basis, but are created through trial and error. Others have a much more solid foundation based on probability theory, for instance. Quite a bit of effort has gone into refining these formulas because very significant performance increases can be obtained.

Weighting schemes have been a major component of papers presented at TREC (annual Text REtrieval Conferences), organized by the U.S. National Institutes for Standards and Technology (TREC, 1992–1999). The proceedings of these conferences can be found online at <http://www.trec.nist.gov/>. Gerard Salton was long an advocate for term weighting approaches and was himself a pioneer in developing techniques for weighting. He and Christopher Buckley summarize the results of the previous 20 years in their 1988 paper “Term-weighting approaches in automatic text retrieval” (Salton & Buckley, 1988), which was reprinted by Sparck Jones and Willett (1997) in *Readings in information retrieval*.

7. Improving the count: stemming

There are sets of related words that are derived from a common root and appear in a variety of forms, depending on particular functions in a sentence or variations in meaning. Thus we have “index”, “indexes”, “indexer”, “indexable”. We also have variants, such as “indices” as another form for the word “indexes”. Stemming was developed to automatically remove certain common suffixes, or word endings (and sometimes prefixes, like “re” or “re-” as in “re-indexing”) in order to increase the count for important words, and also in order to find word occurrences when the word form in the text does not match the word form in the search statement.

The purpose of stemming is to conflate or merge many different words into a single form. The hope is that stemming will increase recall of relevant documents, without too great a cost in decreased precision. But automatic stemming tends to produce errors. The study of word forms or internal structure is called “morphology” in linguistics. Morphological analysis of language is not an easy task, and the more simple or routine stemming procedures do not attempt morphological analysis. Krovetz (1993) has pointed to the kinds of errors that can result from ignoring morphology and the relationship between form and meaning. His examples include the erroneous merging of “organization” and “organ”, “doing” and “doe”, “policy” and “police”, “past” and “paste”, and “arm” and “army” (p. 193).

The simplest stemming is limited to removing the “s” used to make words plural. Of course, all “s”s could simply be removed, regardless of their function, but a more careful “s” stemming

algorithm attempts to distinguish among different types of concluding “s”s, as in “business”, “businesses”, “tomatoes”, “mathematics,” etc.

A much more complicated stemmer (named for its creator Lovins) goes after more than 260 possible suffixes, while a popular middle-of-the road stemmer (called Porter after its creator) settles on 60 or so suffixes. Harman (1994, p. 253) illustrates their differences with the search statement “panels subjected to aerodynamic heating”. The Porter stemmer would not only reduce to the same root “aerodynamic” and “aerodynamics” (as would a simple “s” stemmer), but also “aerodynamically”. It would also combine “heating” and “heated,” as well as “subjected,” “subject,” “subjective,” and “subjects”. The more comprehensive Lovins stemmer would also deal with “aerodynamicist,” as well as “heat,” “heats,” and “heater”.

As noted by Krovetz (1993), stemming may produce unwanted results. After all, an “indexer” is not the same thing as an “index”; “indexing” is a process as opposed to the entity “index”; and “indexable” is an attribute. According to Harman (1994, p. 253), “research has shown that *on the average* results were not improved by using a stemmer”. Her 1991 paper, “How effective is suffixing”, focuses on this question (Harman, 1991). Other researches, however, show improved results in a variety of studies (Hull, 1996; Paice, 1996), especially when more attention is paid to morphological analysis (Krovetz, 1993). Many factors can affect performance, such as “the length of queries, the length of documents, the distribution of the different variants of the word forms. . . , and the way queries are presented”. Results can be better, for example, if phrases are identified in queries, because phrases help to eliminate some errors (Krovetz, 1999). But no matter what the results, searchers have grown to expect at least simple “s” stemming so that they do not have to worry about singular and plural forms.

8. Words versus phrases

Sometimes single words just are not sufficient to name or describe a topic. “Information science”, for example, is not the same as “information” and “science”. Nor is “birth control” simply “birth” and “control” (it is actually control, or prevention, of conception, not of birth!). One of the most oft-cited examples is “venetian blinds”, which if not treated as a phrase, could be retrieved by a query about blind Venetians.

The names of persons and organizations also suffer when reduced to individual words. What is the significance of single words like “Smith” or “National”? The single term “Smith” is not very helpful in identifying a particular individual, nor is “National” (when, for example, “the National Information Standard Organization” is reduced to single words)? So automatic indexing researchers have worked on methods and algorithms for identifying phrases in text, including proper nouns, in an attempt to keep such phrases together. There are also procedures for identifying proper names per se: names of persons, organization, countries, brand names – all of which are very important terms in certain kinds of searches.

Analyzing or parsing the grammatical structure of text can help distinguish between “junior college” versus “a junior in college.” Sophisticated procedures can also combine variations such as “college junior” and “junior in college” into a single form. But such careful analysis procedures tend to be expensive and time-consuming, and the pay-off in terms of results versus investment is still an open question in a variety of contexts. Research in this area is very much a work in

progress. Some recent preliminary results are described in reports on TREC experiments (Callan, Croft, & Broglio, 1995; Strzalkowski, Lin, & Pérez-Carballo, 1997). Pérez-Carballo and Strzalkowski (2000), Sparck Jones (1999), Strzalkowski, Pérez-Carballo, and Marinescu (1996) and Strzalkowski, Lin, Wang, and Pérez-Carballo (1999) provide overviews of several natural language processing techniques in information retrieval, including phrase identification.

The identification of some word phrases can be accomplished in electronic searches by simply specifying that certain words must be contiguous or in close proximity. But for displayed indexes, where the indexing must be done in advance and displayed in meaningful headings, some sort of phrase identification is helpful. Automatic phrase identification is also potentially useful for searching systems that accept natural language queries in the form of sentences or paragraphs, where the user is not expected to indicate phrases. Phrase identification techniques can be applied against both the query and the indexable matter of messages (abstracts or full texts).

One approach that has been tried in some automatic indexing situations could be called a “try everything” method. Over time a list of word phrases important to the subject area is compiled. Then all word pairs and word triplets (possibly quadruplets as well, and so on) are matched against this list. Any such combination that matches is kept as a word phrase in the indexing system. Take, for example, the first sentence of this paragraph. The following word pairs would be matched:

```
one approach
approach that
that has
has been
been tried
tried in
in some
some automatic
automatic indexing
indexing situations
situations could
could be
be called
called a
a “try”
“try everything”
“everything” method
```

In this case, only “automatic indexing” would probably be matched, and therefore kept as a single term of two words. In an earlier sentence, “National Information Standards Organization” might have been found and identified as a meaningful phrase if all four-word sequences were matched against a list of important phrases, including organizational names. A similar approach does not rely on a pre-existing list of phrases, but matches potential phrases against large collections of texts, and word combinations that frequently occur together are accepted as potential phrases.

More sophisticated approaches involve parsing of texts to identify parts of speech and syntactic structures. It is this type of analysis that seeks to distinguish between “college junior” and “junior college”, and also to match “college junior” to “junior in college” (Strzalkowski et al., 1999). This kind of approach allows the system to conflate different grammatical variants into a single “normal” form.

An example of using a variety of techniques to identify phrases and to select potentially more useful ones in texts is “Keyphind” (key phrase find) system created by Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998). Theirs is a hybrid approach, relying on both syntactic analysis of lexical patterns and statistical analysis of frequency counts. Important attributes for identifying key phrases include location of its first occurrence in a text; frequency within text; and frequency in a collection of documents. Using these techniques, the 12 potentially most important phrases for each document are selected. Some stemming is performed to merge singulars and plurals and alternative spellings.

Phrases are especially useful for browsable displayed indexes. Gutwin et al. (1998) demonstrate this with examples from a portion of the New Zealand Digital Library (Witten, Nevill-Manning, McNab, & Cunningham, 1998).

In one example a user types in a query term “text”, and in response, the phrase browser (Keyphind) lists key phrases containing the word “text”, ranked by the number of documents containing the phrase, in an initial window:

Phrase	# Docs
text editor	12
text compression	11
text retrieval	10
full text	8
program text	8
text database	7
text generation	7
structured text	6
input text	5
plain text	5
text element	5
text entry	5
text line	5
text processing	5
text widget	5
text window	5
free text	4
text editing	4
text file	4
text search	4
block of text	3

handwritten text	3
natural language text	3
parallel text	3

Certainly these phrases are more indicative of possible topics discussed in documents than just the word “text”. By browsing these phrases that actually occur in documents, a user can survey what is available and pick the phrase closest to her/his desire, guided also by the information about the number of documents that are available containing the phrase.

If the user now selects the phrase “text retrieval”, brief surrogates for the 10 documents containing this phrase are displayed at the bottom of the browser. However, in a second window beside the initial phrase window, other phrases that co-occur with “text retrieval” are displayed according to frequency. These phrases can be used to further refine the search, or they may suggest other paths of inquiry to the user. Here are the key phrases that co-occurred with “text retrieval”:

Occurring with “text retrieval”	# Docs
information retrieval	4
full text	2
natural language processing	2
retrieval system	2
document retrieval	2
signature file	2
engineering sector	1
signature length	1
office information system	1
search string	1
signature size	1
machine learning	1
scheme-independent retrieval	1
term clustering	1
feature selection	1
navigation tool	1
office system	1
relational database system	1
feature set	1
recognition error	1
recent development	1
information retrieval system	1
natural language indexing	1
feature extraction	1

Now, if the user selects the co-occurring phrase “natural language processing”, surrogates for the two documents containing both “text retrieval” and “natural language processing” will

be displayed in the lower window. By selecting one of these surrogates, the full text can be retrieved.

9. Clustering

All indexing is based on classing or clustering items based on similarities in characteristics. “Classing” simply means to create or define classes of items and/or to assign items to classes. “Clustering” means to create or identify groupings or clusters of items. When an indexer assigns the term “dogs” to a document, or when a computer finds the term “dogs” in a document, a class is created and this document is associated with all other documents that have been indexed with the term “dogs”. The common meaning of “classification”, which often has had less to do with creating classes than with the arrangement of classes in a logical and meaningful order, is an important issue for the creation of browsable displayed indexes – a topic outside the scope of this essay. The term “clustering” is used more often when the classing, or gathering together, is done through automatic, or algorithmic, means. The term “classing” usually implies human judgment.

Some search interfaces provide users with related terms based on clustering – the formation of clusters based on the co-occurrence of terms in documents or the co-occurrence of index terms assigned to them. Similar clusters can be generated on the basis of any document characteristic – other examples include authors’ names, journals in which articles are published, and reference citations.

In machine matching searches, classes or clusters of documents are created based on the terms in a search statement. Another example of clustering is the popular feature in modern IR systems that allows a user to request other documents like one he or she already knows about. In this case, the search program creates a cluster of documents that are similar with the original document and with each other. This similarity can be computed in any number of ways based on a range of characteristics, such as authors, affiliations, index terms, abstract terms, other text terms and reference citations. A number of documents can be identified and can be ranked by their degree of similarity.

Automatic clustering techniques are attempts to compute degrees of association among either terms (term-clustering) or documents (document clustering). Document clustering may be used to organize the document files (static clustering) or on the fly in order to present sets of retrieved documents to the user (dynamic clustering). Clustering has been used in attempts to improve the efficiency and the efficacy of retrieval systems. Work on clustering has been reported since the early days of IR research (Stiles, 1961) to the present (Hearst, 1999). Early work suggested that clustering techniques could improve retrieval performance. More recent work on clustering has failed to show significant improvements in retrieval effectiveness. On the other hand, modern applications of clustering seem to be useful to support user interaction with IR systems. For example, Hearst (1999) describes “scatter-gather” techniques based on dynamic document clustering to facilitate the evaluation of retrieved documents. Documents are first “scattered” into clusters based on document similarities. The documents “gathered” into selected initial clusters can again be scattered and gathered into new clusters to further refine groupings.

The idea of storing documents in clusters such that “similar” documents are “near” each other has been proposed and explored since Salton and colleagues’ early work in the 1960s. This idea is supported by the cluster hypothesis: similar documents tend to be relevant to the same queries. In Salton and McGill (1983a,b) the document clustering techniques used in the 1983 implementation of SMART are described. Clustering results in a file organization that may reduce file accesses and improve the efficiency (speed) of the system but not necessarily its retrieval efficacy. Clustering research through the mid-1980s was described and assessed by Willett (1988) in “Recent trends in hierarchic document clustering: a critical review”.

The intimate details of how the files are organized inside the computer is better left to programmers. Such details may change depending on changes in the hardware and lower level operating system details without any noticeable effect on any of the higher level operations. A change in file organization might result in changes in performance speed but not in the retrieval performance of the system. At the conceptual level that interests us here it does not matter how the document files are organized.

10. Latent semantic indexing

Latent semantic indexing (LSI) is one of the most sophisticated modern attempts at high quality automatic indexing. It is based on clustering of terms based on co-occurrence and the identification of documents associated with these clusters. By relying on co-occurrence data, LSI is also able to deal with the problem of the variety of terms that can be used to express similar ideas. LSI is described by its creators in Deerwester, Dumais, Furnas, Landauer, and Harshman (1990). Gordan and Dumais (1998) describe its application in the search for relationships in widely separated literatures.

As an example of LSI’s capability of dealing with divergent terminology, let’s imagine documents on the repair and maintenance of automobiles. Different documents may use a number of different terms like “automobile,” “car,” “motor vehicle,” “sedan,” plus the names of particular brands and models – “Buick,” “Plymouth,” “Cherokee.” The LSI program is likely to associate these terms together because of their high level of co-occurrence with terms like “oil,” “gasoline,” “fuel,” “carburetor,” “tires,” “air-conditioning,” etc. The LSI program creates clusters of highly related (through co-occurrence) terms, so that when a sufficient number of these terms occurs in a document, the document can be linked to that cluster. In this way, a search for the care and maintenance of carburetors in gasoline-powered automobiles can be made without regard to the particular words used for automobile. All words that mean more or less the same thing as automobile will be linked to the same cluster, as long as a sufficient number of other co-occurring terms match with terms in the cluster. LSI is an example of current attempts to use computing algorithms to get underneath the actual words used, in order to identify the underlying ideas expressed.

11. Citation indexes

Reference citations have always been a useful basis for indicating (indexing!) possibly fruitful relationships. Almost every writer of a term paper, to say nothing of more serious researchers, has

pursued reference citations in good documents they have already found as a way to identify other documents of interest. In effect, a string cluster is created, with each link through a reference citation leading to an older article that was cited in the later paper. This kind of citation indexing can only lead backwards in time, because it is impossible to cite a document that has not yet been created!

Creating indexes that could trace reference citations forward in time was extremely laborious before the advent of the computer. Such citation indexing was limited largely to the legal literature until the Institute for Scientific Information (1961–1999, etc.) introduced the *Science citation index* in 1961, followed by the *Social science citation Index* in 1969 and the *Arts and humanities citation index* in 1976. These indexes have now become standard tools, available in both print and electronic forms. They permit the user to begin with a given document and to trace its citation in subsequent documents forward in time. To the extent that a reference citation indicates a link between messages related with respect to topic, purpose, meaning or significance, these links can be quite useful for literature searching.

12. Bibliographic coupling

Bibliographic coupling is a special form of clustering based on reference citations. The underlying idea is that two or more documents are related (“bibliographically coupled”) if they share the same reference citations. The more reference citations they share, the more closely related they may be. The technique of bibliographic coupling was described by Kessler (1963).

Clusters based on bibliographic coupling are static, because the reference citations in a given document never change after the document is finished and published. Small (1973) invented, or discovered, co-citation clustering as an alternative method for creating dynamic clusters of related documents that tend to focus on new (as opposed to older) documents and are thought to point to active research efforts.

13. Co-citation

In co-citation clustering, clusters are not based on reference citations shared in common (as in bibliographic coupling) but on two or more documents being cited together in a subsequent document. If new papers, hot off the press, frequently cite both documents A and B, then, the reasoning goes, documents A and B must be related, and the more often they are co-cited (cited together in later papers), then the closer the relationship is. Because new documents keep coming out, with different sets of reference citations, the co-citation clusters keep changing over time, showing new patterns of emerging relationships among documents, authors, and the topics they address. This constant change, incorporating new citation patterns, is the basis of the claim (or hope) of its proponents that these co-citation clusters can identify hot topics and emerging research fronts.

14. Relevance feedback

All automatic indexing techniques are the creations of human designers, but in most cases, once these techniques are set in motion, they go their own way with no opportunity for human

intervention or feedback. When automatic indexing techniques are combined with machine matching, human searchers have the opportunity to influence the matching procedures or criteria by specifying a range of parameters. In addition to choosing or changing search terms, they may, depending on what a particular search system permits, be able to indicate relative importance or weights for terms, truncate terms, require certain proximity of terms in text, specify whether case (upper or lower case letters) is to be considered, and similar modifications. But even with these searcher-specified modifications, once the automatic matching process is invoked, it proceeds according to algorithm without modification.

The idea behind relevance feedback is to permit evaluation of preliminary results in order to influence (hopefully to improve!) subsequent performance. This process is analogous to the common practice of searchers browsing displayed indexes. As they examine potential entries, they are constantly making judgments of relevance, and making adjustments in their searching in accordance with these judgments.

Relevance feedback represents an effort to build this natural feedback process into searching and automatic matching procedures. Of course, even without the techniques described here, a searcher can review the results of a search, make modifications in the original search statement based on these results, eliminating some terms, adding some terms, and perhaps changing weights or other parameters. The more formal relevance feedback techniques are designed to make the process easier and a little more automatic. Current applications have been described by Salton and Buckley (1990) and Spink (1995).

The usual approach to relevance feedback is for a preliminary machine matching search to proceed using terms (and modifications such as term weights, truncation, proximity limits, etc.) provided by the user. The results of this initial search are presented to the user, along with an evaluative questionnaire in which the user can indicate preliminary relevance judgments concerning the value of the retrieved documents. These judgments are then used by the system to modify the initial search, and a second search is performed. This interaction can continue as long as the user wishes.

Two types of modifications are generally made. Search terms are added or deleted, and weights for terms are increased or decreased. Search terms that are closely associated with documents that received low or negative relevance ratings may be given lower weights or may be removed from the search altogether. Search terms that are associated with documents that received high relevance ratings can be given higher weights. Also, terms not in the original search but which are closely associated with highly rated documents can be added to the search.

This same approach to ongoing semi-automatic modification of search statements can be used to update search filters for continuously identifying new documents of possible interest as they are added to databases. The resulting profile of user interests serves as the basis for “selective dissemination of information” (SDI) to clients over time. The newer applications of this ongoing indexing and searching of new material are often called “information filtering” (Belkin & Croft, 1992).

Current research is attempting to find ways to improve automatic indexing by tracking patterns of human searching behavior and decision making, and using the results to influence subsequent matching procedures. Whereas current efforts at relevance feedback tend to focus on individual searchers and individual searches, the hope is that longer term group patterns can be used to tailor systems to more closely match the needs of groups of users.

15. Subject analysis and indexing in indexing and abstracting services

Modern indexing and abstracting services (also known as “A&I” services), such as BHA: Bibliography of the history of art/Bibliographie d’histoire de l’Art (J. Paul Getty Trust), BIOSIS/Biological abstracts, Chemical abstracts, EMBASE (Elsevier Science), Engineering index/Compendex, MLA international bibliography (Modern Language Association of America), MEDLINE (National Library of Medicine), Psychological abstracts/PsychINFO (American Psychological Association), PAIS (Public Affairs Information Service), and the several indexing services of the H.W. Wilson Company, all use some sort of human subject analysis and indexing to create their IR databases. Milstead (1992) surveyed and discussed their practices in “Methodologies for subject analysis in bibliographic databases”. A major focus of Lancaster’s (1991, 1998) book on *Indexing and abstracting in theory and practice* (1991, 2nd. ed., 1998) is on the practices of major indexing and abstracting services.

Actual analysis methods, the topic of this essay, are only one part of indexing practice, and in discussions and descriptions of production methods, procedures, and rules, the basic methods of analysis often play second fiddle to aspects that are outside the scope of this essay, such as syntactic rules or patterns used for creating headings, exhaustivity policies for appropriate detail in analysis and indexing, specificity policies for terms or descriptors, vocabulary management and use of thesauri, and the nature of displayed indexes and interfaces. However, some A&I services have attempted to describe and regularize some of the analysis procedures that indexers use or are expected to employ, and some work has begun in moving analysis rules, patterns, and procedures into expert systems that can be used to assist in both the analysis and the indexing stages.

A prominent example of attempts to regularize and encode analysis procedure is the Medical Indexing Expert (MedIndEx) project of the National Library of Medicine. MedIndEx is a prototype for knowledge-based indexing which makes use of “encoded domain knowledge, including domain-specific relationships between concepts, as well as computer-executable rules that, using this knowledge, participate both actively and specifically ... in the steps of a complex intellectual task” (Humphrey, 1994, p. 166). Humphrey gives an example of an indexer having indicated a medical therapy: “estrogen replacement therapy”. MedIndEx prompts the indexer to indicate the “problem” by naming the particular disease to which the therapy is being applied, if known. A list of possibilities is automatically provided from Medical Subject Headings (MeSH), the thesaurus of medical terminology maintained by the National Library of Medicine. When the indexer chooses “osteoporosis, postmenopausal,” MedIndEx asks for body location, by suggesting “one of the following organ-disease terms, if appropriate: spinal disease; orbital diseases; maxillary diseases; mandibular diseases; jaw diseases” (Humphrey, 1994, p. 167).

A less ambitious example, also from the National Library of Medicine, but also in use elsewhere, is the use of check tags. Check tags are lists of important facets or aspects of subject analysis. Their purpose is to remind indexers to consider all important aspects of a topic. Thus, in indexing for MEDLINE at the National Library of Medicine, the “Automated Indexing and Management System (AIMS)” has used check tags to remind indexers to indicate gender and species of patients or subjects of experiment when relevant. In some cases, the system would automatically insert the term for a concept. For example, if an indexer indicated the concept

“pregnancy”, AIMS would insert the term “female”, but a check tag would ask for information as to “human” or “animal” (Humphrey, 1994, p. 165).

Increasingly, indexing and abstracting services have experimented with and implemented various methods to take fuller advantage of capabilities offered by advancing research in automatic indexing. Hodge and Milstead (1998) focused on this evolving aspect of operational large-scale indexing operations in a survey of major A&I services. Their results are reported in *Computer support to indexing*, a special report for the National Federation of Abstracting and Information Services. They conclude that:

“The final resolution to the debate over natural language processing [i.e., advanced automatic indexing] versus human indexing remains to be determined. However, it is certain that the degree to which natural language processing and its incorporation into Internet-based search technologies can be perfected will have a profound effect on the future of indexing. Both natural language processing and increased demand for human indexing and knowledge organization (whether by the building of organized virtual-digital libraries or by the development of knowledge representation systems based on thesauri) will require increased computer support” (p. 80).

16. Growing role of automatic analysis and indexing

It is clear from research and the experience of users that automatic machine-based indexing and human intellectual analysis-based indexing both make important, but very different, contributions to successful information retrieval. At the same time, expert human indexing keeps getting more expensive, while automatic indexing becomes, comparatively, less and less expensive and more effective. Therefore, it seems likely that future IR databases will seek to maximize benefits by allocating human analysis and indexing to situations where the benefits of human expertise are most apparent and immediate.

In order to improve the effectiveness and efficiency of the information retrieval enterprise, librarians, database producers, and other information professionals need to stop treating every document as if all documents, all texts, and all messages were equally important. We know this is not the case. We need to be more judgmental and discriminating, in the best sense of these terms. We all learn about the so-called “80–20 rule” that suggests that in any large collection of documents, 20% will get 80% of the use, or, to put it differently, 20% of the documents will answer 80% of the questions, or respond to 80% of the needs or desires of users. To allocate human analysis expertise in a rational, cost-effective manner, we need to develop methods for predicting the more important documents and devoting human analysis to them. All documents can receive inexpensive, relatively effective automatic, machine-based analysis and indexing. For important documents, automatic indexing can be augmented by human indexing, to make these documents even more accessible to a broader clientele.

In highly selective libraries and databases, it will probably continue to make sense to apply human analysis to all selected documents, but in IR databases that seek to be comprehensive within their domains, then it makes sense to apply some discriminating criteria within the database collection in an attempt to identify the more important documents. As IR database grow

ever larger, such discriminating criteria can be beneficial for users who often want not everything on a topic, but only the best and most appropriate items.

Human indexing invested in more important documents can make those documents more accessible to users by identifying themes, relationships, methodological approaches, points of view, prejudices, biases, slants, purposes, values, and qualitative aspects that cannot be easily identified through automatic techniques. The role of the human indexer can take on that of the “readers adviser” in traditional reference services, moving beyond simple topical description. Such human indexing should be guided by the subject scope and domain, mission and objectives of the IR database, which if well formulated, will reflect the needs and desires of potential users.

Human indexing also has a potentially important contribution to make in combating the inevitable scatter of relevant messages in large collections or IR databases. Bates (1998, pp. 1193–1198) summarizes the naturally occurring statistical distributions that characterize information phenomena and IR databases. Zipfian distributions (so named because George Kingsley Zipf first described them in mathematical terms) characterize many information phenomena, such as the distribution of word frequencies in language texts. One of the most important, called Bradford’s Law in honor of Samuel Bradford, describes the greater and greater dispersion of relevant documents beyond a core group of the most relevant documents. Bates asserts that “even in the best-designed database, users will usually retrieve a number of irrelevant records, and the more thorough they are in trying to scoop up every last relevant record by extending their search formulation with additional term variants, the greater the proportion of irrelevant records they will encounter. However, as the ‘core’ terms will probably retrieve a relatively small percentage of the relevant records (certainly under half, in most cases), they must nonetheless tolerate sifting through lots of irrelevant records in order to find the relevant 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 there,** [emphasis added] but it should be understood that even the best human indexing is not likely to defeat the underlying Zipfian patterns” (p. 1198). This points to the need to design tools (and systems) that help users to do the sifting more efficiently.

Certainly, it is unlikely that the deployment of human indexing expertise to combat the Zipfian or Bradfordian scatter of relevant documents can be undertaken willy-nilly across the board. Instead, human expertise can better be devoted to documents that possess some predicable attributes associated with potential importance.

Here are some examples of ways in which “more important” documents might be identified. Whatever criteria are used should be made explicit, in line with Bates’ plea in her landmark paper “Rigorous systematic bibliography” (Bates, 1976).

- *use*. In libraries, especially special libraries, circulation or other use data (such as interlibrary loan and photocopy requests) can be used to identify high-use documents.
- *citation*. Citations to scholarly literature can be monitored through citation indexes. Most active scholars and researchers learn about important documents in their areas of interest not through indexing and abstracting services or other IR databases, but via the informal networks that characterize the “invisible colleges” of scholars – interaction at conferences, telephone calls, email discussion groups, world-wide websites, personal email, and mention by colleagues. Thus initial use and citation are generated in many cases by means independent

of indexing and abstracting services and other IR databases. The augmented indexing given to documents identified by use and citation data helps to make these same important documents available to users who are not tied into the scholarly networks – students, scholars in other fields, and other interested persons. Indeed, persons who want to find only the most important documents could limit their searches to such documents.

- *publisher prediction.* Documents published as trade or scholarly monographs represent a tiny percentage of the universe of documents, both print and electronic, in our information environment. When a publisher decides to invest the large amount of money necessary to publish and market a book in print form, in almost all cases, such a document deserves an index based on human intellectual analysis. Fiction, drama, poetry, and other belles-letters are possible exceptions, but a growing number of scholars and other users advocate indexing for these works as well.
- *reviews and awards.* Documents that are reviewed or receive awards should receive augmented indexing. Such documents represent a tiny percentage of all documents, and their selection is based on judgments by reviewers, editors, or practitioners that a document is significant for some reason.
- *searcher nomination.* As modern IR databases and search interfaces become more interactive, they can include features to encourage users to provide feedback to indexers and publishers of databases on a variety of issues, such as indexing terminology and the selection of terms for the description of particular messages. Another area of user input could be the nomination of “important” documents to receive augmented indexing. Searchers who find particular documents useful could say so in a response form, and these responses could be incorporated into the record for the document so that searchers interested in the comments of other users could see them. At the same time, user nomination could also be used to select documents to receive human indexing.

Searcher nomination or recommendation is spreading on the world-wide web, where users often comment on other websites and add links to them in their own websites, thus making the recommended sites easier to find by others. There are also a growing number of “recommender systems” on the web, which use the experience of previous users to recommend documents or sites to new users. The invitation of Amazon.com (and similar sites) to users to add their recommendations or reviews are examples of searcher nominations.

- *advisory board.* IR databases that have a board or network of advisors to assist with source domain coverage can also make use of these expert users to nominate documents or classes of documents that deserve treatment as important documents.
- *indexer nomination.* Expert indexers, especially those with experience in a subject domain, can also identify potentially important documents.
- *exemplary documents.* Blair and Kimbrough (2000) have advocated the identification of exemplary documents to facilitate focused information retrieval, especially for new-comers to a field. Such documents, which would “describe or exhibit the intellectual structure of a particular field of interest”, are a prime example of important documents that would merit the attention of human indexers. Indeed, a primary role of human indexers might be to identify such documents. These documents could then be used, in accordance with Blair and Kimbrough’s proposal, as gateways to literature connected, either through vocabulary or other links, to a broader range of documents.

17. Censorship or guidance?

Some students and colleagues argue that limiting expensive, expert human analysis and indexing to important documents constitutes a kind of censorship. One response is that such limits would be censorship only if criteria related to political, social, religious, ideological, personal, or other prejudices were used to limit access.

But human judgments are always mixed with cultural world views, prejudices, biases and values, both conscious and unconscious, so it could be argued that choosing documents based on actual use rather than on the judgments of a relatively small number of humans could be considered a safeguard against censorship. When measures of importance are based on actual use and recommendations by users, gathered on a relatively large and representative scale, then the user community can be said to be choosing what is important to it.

In addition, selection of useful documents based on nomination or recommendation by advisory groups and indexing staff might further insure an openness to consider documents that might not otherwise be given additional analysis. What should be avoided is a closed circle of popularity, such as that created when an IR database limits its documentary scope to journals selected by votes of librarians, who may tend to purchase journals that are indexed, and then to vote for the journals that they own.

Some critics argue that popularity criteria are similar in kind to those used by all sorts of marketers of entertainment, news, and other consumer products, resulting in excessive blandness. This is only partly true. In marketing, consumer surveys are conducted to ascertain consumer tastes and views, but then advertisers go out of their way to seek to influence and change these tastes and opinions. In a way, librarians and information experts who seek to influence the use of documents are comparable to advertisers. In librarianship, there is a long, and sometimes controversial tradition, of providing to our users what they *should* be using, rather than what they ask for. A growing consensus seems to be swinging in the other direction. That the user is the best judge of what is useful, beneficial, and to be desired. This controversy is often associated with collection development in libraries, but the principles are the same when applied to indexing and the creation of IR databases.

As in many controversies, both sides have some of the merit. The opinions and desires and information-seeking behavior of our users must be respected. But in addition, in some cases, expert librarians and other information specialists are indeed equipped to make judgments of importance, and they should be encouraged to do just that on behalf of users. Librarians have always done so and must continue to do so in selecting books and other documents for library collections and in defining the coverage of databases with respect to subject and documentary scope and domain. However, in the realm of human indexing and cataloging, indexers and catalogers have traditionally been loathe to express qualitative judgments or opinions. This avoidance is unfortunate and should stop. It is one area where human indexing can be clearly superior to machine indexing. We need to add to the value of our human contributions to information retrieval by expressing qualitative judgments and by highlighting what we believe is best and most appropriate for various categories of users. There is no reason why our indexing should not reflect this. In short, expert opinion can be yet one more criterion for identifying important documents that should receive the attention of human indexers.

Indeed, a few indexing theorists, such as Wilson (1978) and Frohmann (1990) have advocated much more attention to the identification of “truth” or accuracy in messages. Such assessments require careful analysis, and would mean allocating more time and attention on fewer documents, but such assessments might be very valuable in some contexts.

If we ask human indexers to play a larger role in assessing messages, then we may want to let our users know who they are, much like book or movie reviewers who gain a following among those who respect their judgments (or whose names warn others to ignore their views). Expert judgments, in addition to prejudices, biases, and values, can be made more explicit by identifying, and giving credit to, the evaluators. Indeed IR databases featuring the views of such expert indexers will take on the trappings of “expert” websites.

By relying on mass judgments of importance, coupled with experts imposing their assessment of accuracy and value on messages, will we run the risk of finding only the bland and conventional at the cost of the truly original, unconventional, controversial? For this task of discovery, perhaps we must place reliance and hope on intrepid explorers using the kind of full-text browsing techniques that were described previously in the section on phrase indexing.

What we cannot afford to continue to do is to treat all documents that enter our collections and our IR databases as if they were all equally important and equally deserving of our expert analysis and indexing. They simply are not, and to continue to do so is to waste precious resources.

The exploding internet and world-wide web is a prime example where selectivity in deploying human expertise is absolutely essential!

To bring this discussion of humans versus machines in indexing to a close, here is Wellisch’s (1998) concluding quote, reflecting predictions from Wiener’s (1950) book *The human use of human beings* in an essay on “Indexing after the millennium...: the indexer as helmsman”: “[C]ontrol exercised by machines, far from enslaving human beings, will liberate men and women for tasks only they can perform. Let us hope that the coming century will see an increasing use of human indexers for the forging of keys to the hidden treasures of information in all its forms”.

References

- Bates, M. J. (1976). Rigorous systematic bibliography. *RQ*, 16(1), 7–26.
- 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.
- Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12), 29–38.
- Blair, D. C., & Kimbrough, S. O. (2000). Exemplary documents: a foundation for information retrieval design. *Information Processing and Management*, in press.
- Brooks, T. A. (1998). Orthography as a fundamental impediment to online information retrieval. *Journal of the American Society for Information Science*, 49(8), 731–741.
- Callan, J. P., Croft, W. B., & Broglio, J. (1995). TREC and TIPSTER experiments with inquiry. *Information Processing and Management*, 31(3), 327–332, 343. Reprinted in: K. Sparck Jones, P. Willet (Eds.), *Readings in information retrieval* (pp. 436–439). San Francisco: Morgan Kaufman, c1997.
- Cavnar, W. B. (1994). Using an n-gram-based document representation with a vector processing retrieval model. In D. Harman (Ed.), *TREC-3: Proceedings of the third Text REtrieval Conference* (pp. 269 ff); 2–3 November 1994; Gaithersburg, MD. Co-sponsored by the National Institute of Standards and Technology (NIST) and the Advanced Research Projects Agency (ARPA). Gaithersburg, MD: US Department of Commerce, Technology

- Administration, National Institute of Standards and Technology; Washington, DC: For sale by the Supt. of Docs., US G.P.O., (NIST Special Publication; 500–226). <http://www.trec.nist.gov/>.
- Croft, B. (1989). Automatic indexing. In B. H. Weinberg (Ed.), *Indexing: the state of our knowledge and the state of our ignorance: Proceedings of the 20th annual meeting of the American Society of Indexers* (pp. 86–100); 13 May 1988, New York. Medford: Learned Information.
- Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Francis, W. N., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.
- Frohmann, B. (1990). Rules of indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation*, 46(2), 81–101.
- Gordan, M. D., & Dumais, S. T. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, 49(8), 674–685.
- Gutwin, C., Paynter, G., Witten, I. H., Nevill-Manning, C. G., & Frank, E. (1998). Improving browsing in digital libraries with keyphrase indexes. Technical report 98-1, Computer Science Department, University of Saskatchewan, <http://www.cs.usask.ca/faculty/gutwin/1998/keyphind-techreport/html/keyphind-9.html>.
- Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7–15.
- Harman, D. (1994). Automatic indexing. In R. Fidel, T. B. Hahn, E. M. Rasmussen, P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 247–264). Medford, NJ: Learned Information for the American Society for Information Science.
- Hearst, M. (1999). The use of categories and clusters for organizing retrieval results. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 333–374). Dordrecht: Kluwer Academic Publishers.
- Hodge, G. M., & Milstead, J. L. (1998). *Computer support to indexing*. Philadelphia, PA: National Federation of Abstracting and Information Services.
- Huffman, S. (1995). Acquaintance: language-independent document categorization by n-grams. In D. Harman (Ed.), *TREC-4: Proceedings of the fourth Text REtrieval Conference* (pp. 269 ff); 1–3 November 1995; Gaithersburg, MD. Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). Gaithersburg, MD: US Department of Commerce, Technology Administration, National Institute of Standards and Technology; Washington, DC: For sale by the Supt. of Docs., US G.P.O. (NIST Special Publication; 500–236) <http://www.trec.nist.gov/>.
- Hull, D. A. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70–84.
- Humphrey, S. M. (1994). Knowledge-based systems for indexing. In R. Fidel, T. B. Hahn, E. M. Rasmussen, P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 161–175). Medford, NJ: Learned Information for the American Society for Information Science.
- Institute for Scientific Information (1961–1999). *Science citation index*. Philadelphia: ISI.
- Institute for Scientific Information (1969–1999). *Social sciences citation index*. Philadelphia: ISI.
- Institute for Scientific Information (1976–1999). *Arts and humanities citation index*. Philadelphia: ISI.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.
- Kowalski, G. (1997). *Information retrieval systems: theory and implementation*. Boston: Kluwer Academic Publishers.
- Korfhage, R. R. (1997). *Information storage and retrieval*. New York: Wiley Computer Publishing.
- Krovetz, R. (1993). Viewing morphology as an inference process. In R. Korfhage, E. Rasmussen, P. Willett (Eds.), *Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval* (pp. 191–202); 27 June–1 July 1993; Pittsburgh, PA. New York, Association for Computing Machinery. (Also available as UMass technical report TR-93-35).
- Krovetz, R. (1999). Personal correspondence to J. Pérez-Carballo.
- Lancaster, F. W. (1991, 1998). *Indexing and abstracting in theory and practice*. Champaign, IL: University of Illinois, Graduate School of Library and Information Science; 1991. A 2nd ed. was published in 1998.
- Mayfield, J., & McNamee, P. (1998). Indexing using both n-grams and words. In E. M. Voorhees, D. Harman (Eds.) *TREC-7: Proceedings of the seventh Text REtrieval Conference*; (pp. 419–424); 9–11 November 1998; Gaithersburg, MD. Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced

- Research Projects Agency (DARPA). Gaithersburg, MD: US Department of Commerce, Technology Administration, National Institute of Standards and Technology; Washington, DC: For sale by the Supt. of Docs., US G.P.O. (NIST Special Publication; 500–242). <http://www.trec.nist.gov/>.
- Milstead, J. L. (1992). Methodologies for subject analysis in bibliographic databases. *Information Processing and Management*, 28(3), 407–431.
- Paice, C. D. (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, 47(8), 632–649.
- Pérez-Carballo, J., & Strzalkowski, T. (2000). Natural language information retrieval: progress report. *Information Processing and Management*, 36(1), 155–178.
- Salton, G. (1975). *A theory of indexing*. Philadelphia: Society for Industrial and Applied Mathematics.
- Salton, G., & McGill, M. J. (1983a). *Introduction to modern information retrieval*. Englewood Cliffs, NJ: Prentice-Hall.
- Salton, G., & McGill, M. J. (1983b) The SMART and SIRE experimental retrieval systems. Reprinted in: K. Sparck Jones, P. Willett (Eds.), *Readings in information retrieval* (pp. 118–155). San Francisco: Morgan Kaufman; c1987.
- Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513–523. Reprinted in: K. Sparck Jones, P. Willett (Eds.), *Readings in information retrieval* (pp. 323–328). San Francisco: Morgan Kaufman.
- Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- Small, H. G. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269.
- Sparck Jones, K., & Willett, P. (Eds.) (1997). *Readings in information retrieval*. San Francisco: Morgan Kaufman.
- Sparck Jones, K. (1999). What is the role of NLP in text retrieval? (pp. 1–24) In T. Strzalkowski (Ed.), *Natural language information retrieval*. Dordrecht: Kluwer Academic Publishers.
- Sparck Jones, K., Walker, S., & Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6), 779–840.
- Spink, A. (1995). Term relevance feedback and mediated database searching: implications for information retrieval practice and systems design. *Information Processing and Management*, 31(2), 161–171.
- Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the Association of Computing Machinery*, 8(2), 271–279.
- Strzalkowski, T., Lin, F., & Pérez-Carballo, J. (1997). Natural language information retrieval: TREC-6 report. In E. M. Voorhees, D. Harman (Eds.), *Information technology: the sixth Text REtrieval Conference (TREC-6)* (pp. 209–228); 19–21 November 1997; Gaithersburg, MD Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). Gaithersburg, MD: US Department of Commerce, Technology Administration, National Institute of Standards and Technology; Washington, DC: For sale by the Supt. of Docs., US G.P.O. (NIST special publication; no. 500-240). <http://www.trec.nist.gov/>.
- Strzalkowski, T., Lin, F., Wang, J., & Pérez-Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 113–145). Dordrecht: Kluwer Academic Publishers.
- Strzalkowski, T., Pérez-Carballo, J., & Marinescu, M. (1996). Natural language information retrieval in digital libraries. In E. A. Fox, G. Marchionini (Eds.), *Proceedings of the first ACM international conference on digital libraries* (pp. 117–125); 20–23 March 1996; Bethesda, MD. New York: Association for Computing Machinery.
- TREC (1992–1999). Text Retrieval Conferences. In E. M. Voorhees, D. Harman (Eds.), *Proceedings of the Text Retrieval Conferences 1–8; 1992–1999*; Gaithersburg, MD. Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). Gaithersburg MD: US Department of Commerce, Technology Administration, National Institute of Standards and Technology; Washington, DC: For sale by the Supt. of Docs., US G.P.O. <http://www.trec.nist.gov/>.
- Wellisch, H. H. (1998). Indexing after the millennium 3: the indexer as helmsman. *The Indexer*, 21(2), 89.
- Wiener, N. (1950). *The human use of human beings: cybernetics and society* (2nd ed. rev). London: Eyre and Spottiswoode; 2nd ed., Garden City, NY: Doubleday; 1954.

- Willett, P. (1988). Recent trends in heirarchic document clustering: a critical review. *Information Processing and Management*, 24(5), 577–597.
- Wilson, P. (1978). Some fundamental concepts of information retrieval. *Drexel Library Quarterly*, 14(2), 10–24.
- Witten, I. H., Nevill-Manning, C. G., McNab, R., & Cunningham, S. J. (1998). A public library based on full text retrieval. *Communications of the ACM*, 41(4), 71–75.