Fall 2004

If It Ain't Broke ... : Changing Search Philosophies

John M. Weiner
SUNY University at Buffalo

ELECTRONIC JOURNAL OF ACADEMIC AND SPECIAL LIBRARIANSHIP

v.5 no.2-3 (Fall 2004)

If It Ain't Broke ...: Changing Search Philosophies

John M. Weiner
weiner@buffnet.net

Abstract

This study employs simulation analyses to determine the consequences of two search philosophies. The first is called the *a posteriori* approach and involves terms selected arbitrarily by the user without knowledge of the specific content of the documents. The second is the *a priori* approach and involves terms selected because the user knows that the author employed those exact terms in describing his/her findings. Further, the authors' combinations of these terms would be known. The results of simulation studies show that the *a posteriori* approach was comparable to a random walk. If the need to correctly identify documents was given high priority, would the *a posteriori* approach be discarded and replaced by the *a priori* approach? The feasibility of the latter can be seen by the elimination of thousands of hours of effort by users involved in performing flawed searches and independent replication of extractions of the same documents. However, are we far enough into the 21st Century and the Information Age to consider such a change?

Introduction

Articles dealing with literature searches often begin with a statement lamenting the flaws in retrieval provided by current systems [Allen 2001]. The complaints describe the overwhelming deluge of documents retrieved leaving the user with the odious task of manually sifting through them. [Kiernan 2004] Present procedures in finding and using scientific information to build new descriptions or research strategies represent an intermediate position between a paper-oriented philosophy and a computer-supported one. Bibliographic searching is performed using computerized techniques while the documents identified are processed using manual procedures. Many flaws in this present mixture of techniques have been discussed [Aguillo 2000; Allen 2001; Ding 2001; Liu 2004; Lowe 2000; Markov 2004]. One of the more critical ones deals with the use of computerized search strategies [Brinn 1998; Powell 2003; Rappoport 2003; Voss 2001].

Current search strategies employ specific elements. These include terms selected from natural language or fixed vocabularies. The fixed vocabularies have appeal in attaining standardization across documents. However, there is no restriction on the authors' use of vocabulary. This freedom of expression sets the stage for discrepancy between the authors' descriptions of content and the indexers' selection of appropriate descriptive terms from a fixed list [Horowitz 1982]. These fixed vocabularies may or may not be current relative to the authors' introduction of terms and concepts. In addition, these fixed lists are typically not available to the user when the search is constructed.

Even if the list of indexing keywords were made available to the user, their actual use in the text is not. While examples abound that show the defects in simply knowing the terms, each is a single sample from an infinite number of samples dealing with search issues [Bernier 1961; Blair 1986; Funk 1983; Maron 1977; Weinberg 1988]. As an example, in attempting to identify the book, *Fantasies in Processing* [Weiner 2004], entry of the single term "fantasies" in a popular web-based search engine, resulted in identification of over 2.8 million hits. Using the term "processing" yielded more than 34 million hits. This book dealt with World Wide Web-based procedures in automated database construction and management, computerized text analysis and continuing education. Entering the search string "web-based data management OR web-based text analysis OR web-based education" yielded over 1.7 million hits. Adding the string "fantasies in processing" to the "web-based data management, etc." string yielded almost 3,000 hits of which one was correct. When only the string "fantasies in processing" was entered, the search yielded more than 42,000 hits, one of which was correct. Changing the string to "processing fantasies" yielded approximately the same number of hits with the correct item identified. Finally, forcing the search to consider the exact phrase, "fantasies in processing" yielded 17 URL's all describing the correct document.

Web page searching uses a set of terms chosen by the authors to characterize the reports [Lowe 2000; Powell 2003; Rappoport 2003]. These terms may or may not be the same as those used in the actual document. Indeed, services exist to assist in selecting the right mix of keywords to enhance recognition by the search software. [e.g., Microsoft bCentral 2004] Whether or not the terms accurately describe the report, the user is not made aware of the terms used in describing each page.

Similarly, in searching text, the terms provided by the author could be used as key indexing terms. One study suggested that there was about 50% agreement between authors' selection of keywords and indexers' choices. [Horowitz 1982] Again, the user may not be aware of these author-selected keywords. Further, those provided by the author may not actually occur in the text.

These findings suggest that using the present search process involving the *a posteriori* approach will not be effective in identifying the correct document without considerable manual effort subsequent to the computer software's efforts.

Since the user frequently intends to identify and process documents in order to learn new information, there is no guarantee, *a priori*, that these exact queries will be known. More frequently, the user must enter arbitrary individual terms or topics, none of which are known to be in the documents.

In earlier times, the reason for this was clear. The effort involved in reading and identifying the authors' vocabulary and ideas was considerable. With the advent of high-speed text processing in Windows-based desktop or laptop computers, the need to function without real knowledge of the authors' actual vocabulary, ideas and conceptual issues is less compelling.

Is it worth the effort to have the computer "read" the text and provide true awareness of contents? The example showed that almost three million documents had the term "fantasies" and over ten times that many had the term "processing." Finding the correct document would be difficult. In contrast, in the last search using the exact phrase "fantasies in processing," the user knew precisely what to request because he/she knew what the authors had used. The retrieval changed in quality to one of specific, relevant documents.

This study explored the implications of using simulation techniques [Bruce 2000] representing the present *a posteriori* process versus the alternative *a priori* process. The results of that investigation provided probability distributions representing the consequences of each search approach. This simulation approach is a feasible alternative to the thousands of hours of manual effort required in studying the consequences of computerized searching.

Methods

In the *a posteriori* approach, the computer software performs the following tasks:

1. Reads the text.
2. Matches terms used in the search with those in the text.
3. Identifies the document containing the terms either singly or in combination.
4. Prepares a display of the selected data.

These software packages also may include criteria to aid in selecting the more relevant documents. This approach takes seconds of computer time to obtain the display and

hours of manual processing time to identify the correct ones. Combinations of terms using the Boolean AND or OR operators are called *a posteriori* Boolean combinations.

The *a priori* approach is associated with a different strategy. Here, the authors provide the terms and combinations (called a priori Boolean combinations) to be used in search and retrieval. The computer software performs the following tasks:

1. Reads the text.
2. Identifies the sentences.
3. Identifies the informative terms (authors' vocabulary) within each sentence.
4. Identifies the couplets of these informative terms within each sentence.
5. Links each couplet with the associated sentence.
6. Links each couplet and sentence with the appropriate bibliographic data.
7. Stores the extracted data in a web-based knowledge resource.

The resource displays the vocabulary used by the authors and for each term, the related terms representing the couplets. Selection of a particular couplet displays the sentences containing that couplet and, if desired, a seamless link to PubMed enabling retrieval of the specific document. The complete process (items 1-7) requires approximately 0.6 minutes per document.

A simulation programming language entitled Resampling Stats [Bruce 2000] provides an easy to learn, powerful capability in describing and assessing situations difficult to study in real life. Using this tool, the *a posteriori* approach can be estimated by generating random terms and comparing these against two designated terms. By modulating the matching criterion, probability distributions of proportion of matches in searches can be constructed. In addition, the program can determine the probability distribution for the matching of both terms.

This same program, with a different matching criterion, can be used to approximate the *a priori* approach. The probability for each term and for the pair must approach certainty (Probability = 1.0).

Results

A Posteriori Boolean Combinations

A program was developed that mimics the selection of two designated terms, each of 6 characters, and the testing of those against the terms found in documents containing 200 terms. The search was performed 1000 times using 100 documents in each set. The number of articles was counted describing individual matches of designated terms, as well as matched pairs. The frequency of term matching for each set was displayed as a

proportion of the 100 documents. The matching criterion was selected so that the proportion of articles in a set approximated an average of 80% to 90%. Pairs of matched terms also were identified.

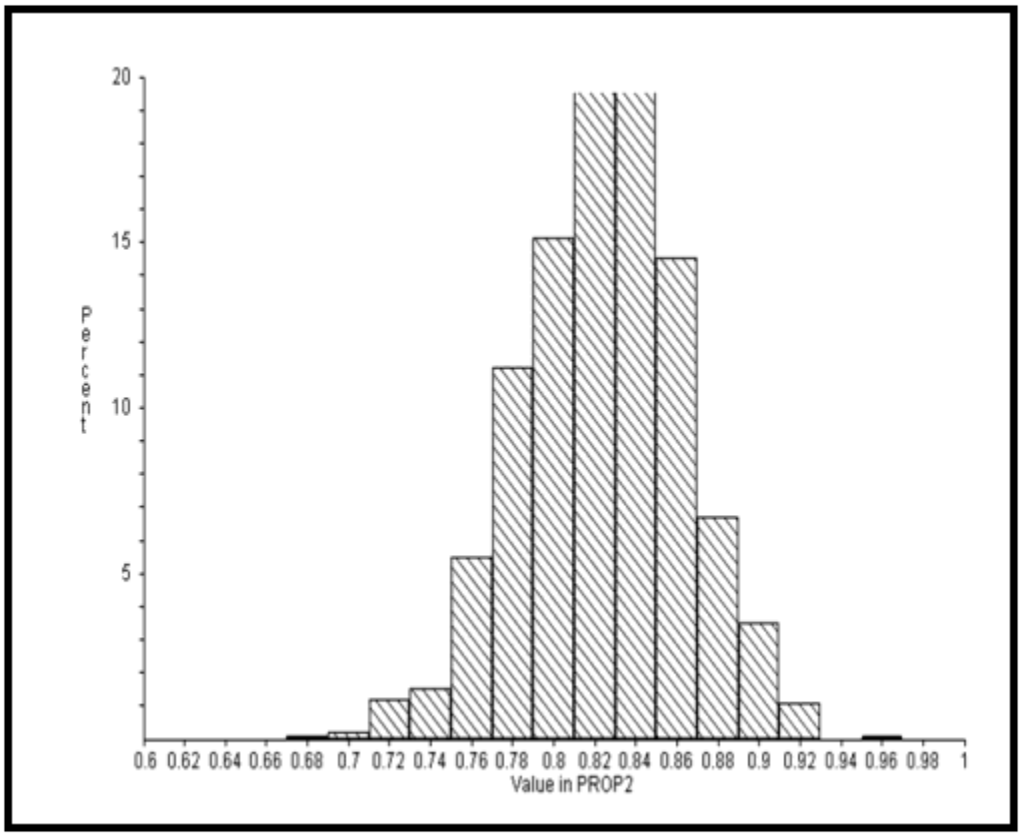

Figure 1. *A Posteriori* Matching using Two Designated Terms. The First of the Pair is Matched with Terms in the Documents.

Figure 1 shows the results of the matching process involving the first term in the search query. The proportion of matches ranged in value from 66% to 98%. The central value approximated 83%. The proportion of matches associated with the second search query term also ranged from 78% to 98% with a central value of 88% (data not shown).

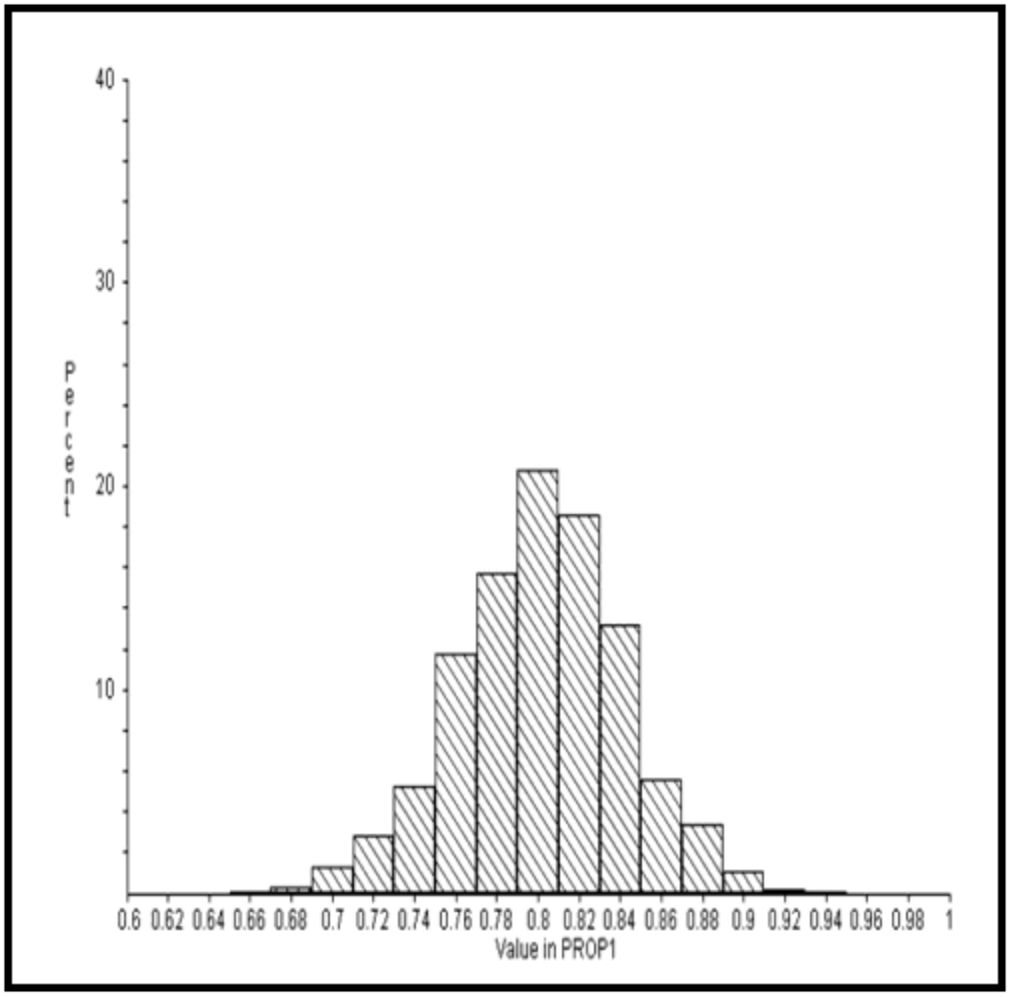

Figure 2. *A Posteriori* Matching using Two Designated Terms. The Pair is Matched with Terms in the Documents.

Figure 2 shows the distribution of proportions of matches for the pair of designated terms. A success is defined when the both designated terms are matched by terms in the document. Approximately 80% of the documents in the 100 document sets contained the pair of designated terms. Matching ranged from 66% to 94%. The sampling distribution is symmetrical and approximates the Gaussian form. These results are consistent with the product of the individual probabilities (i.e., 0.83 x 0.88 = 0.79).

A Priori Boolean Combinations

The criterion for recognizing a match was changed to reflect the knowledge possessed by the user. In this search process, the precise terms used by the authors of the scientific reports, in combination, are known and used in the search query.

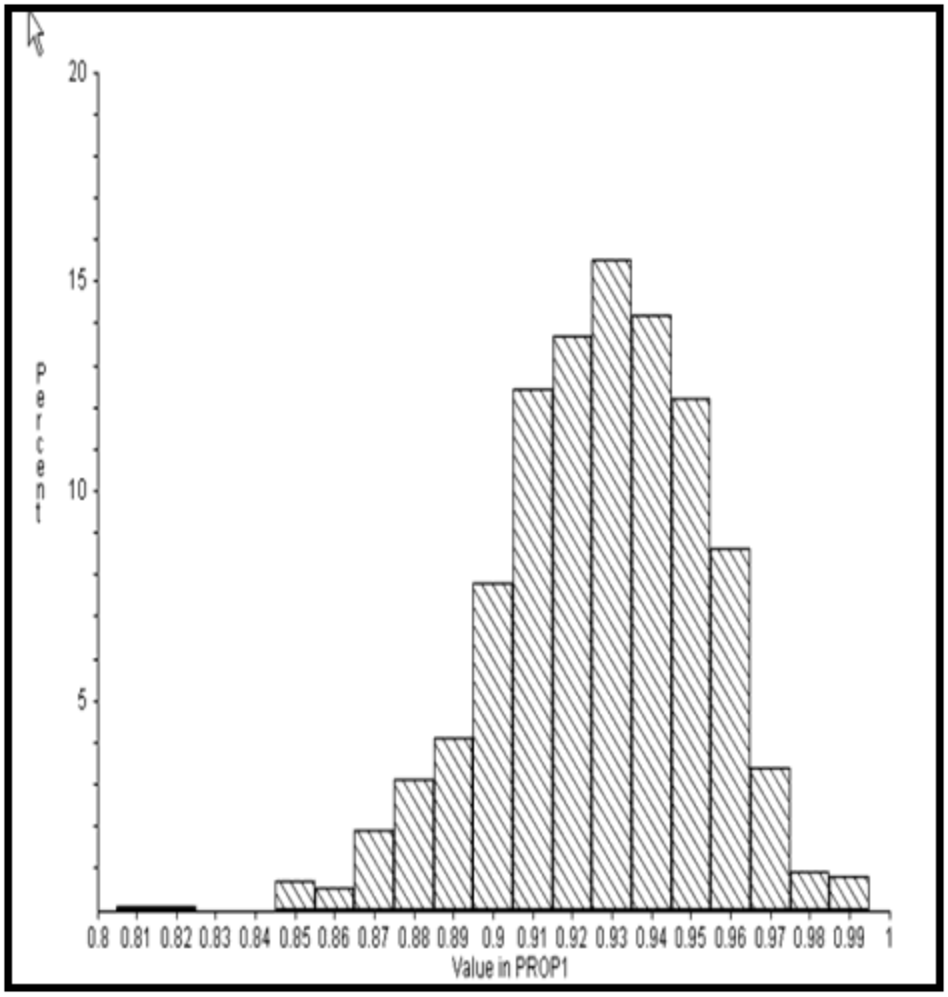

Figure 3. *A Priori* Matching using Two Designated Terms. The Pair is Matched with Terms in the Documents.

Figure 3 shows the distribution of proportions of paired matches ranging from 81% to 98% with central value approximating 93%. The distribution for the matching of the first designated term ranged from 85% to 98% and the distribution for the second designated term ranged from 94% to 100%.

This example approached the exactness of the *a priori* process although errors (e.g., typing or mechanical) were assumed to be present. It showed that as term matching improved, the recognition of pairs also would improve. When error is eliminated, the probability is one in describing the recognition of the pair of terms provided by the author.

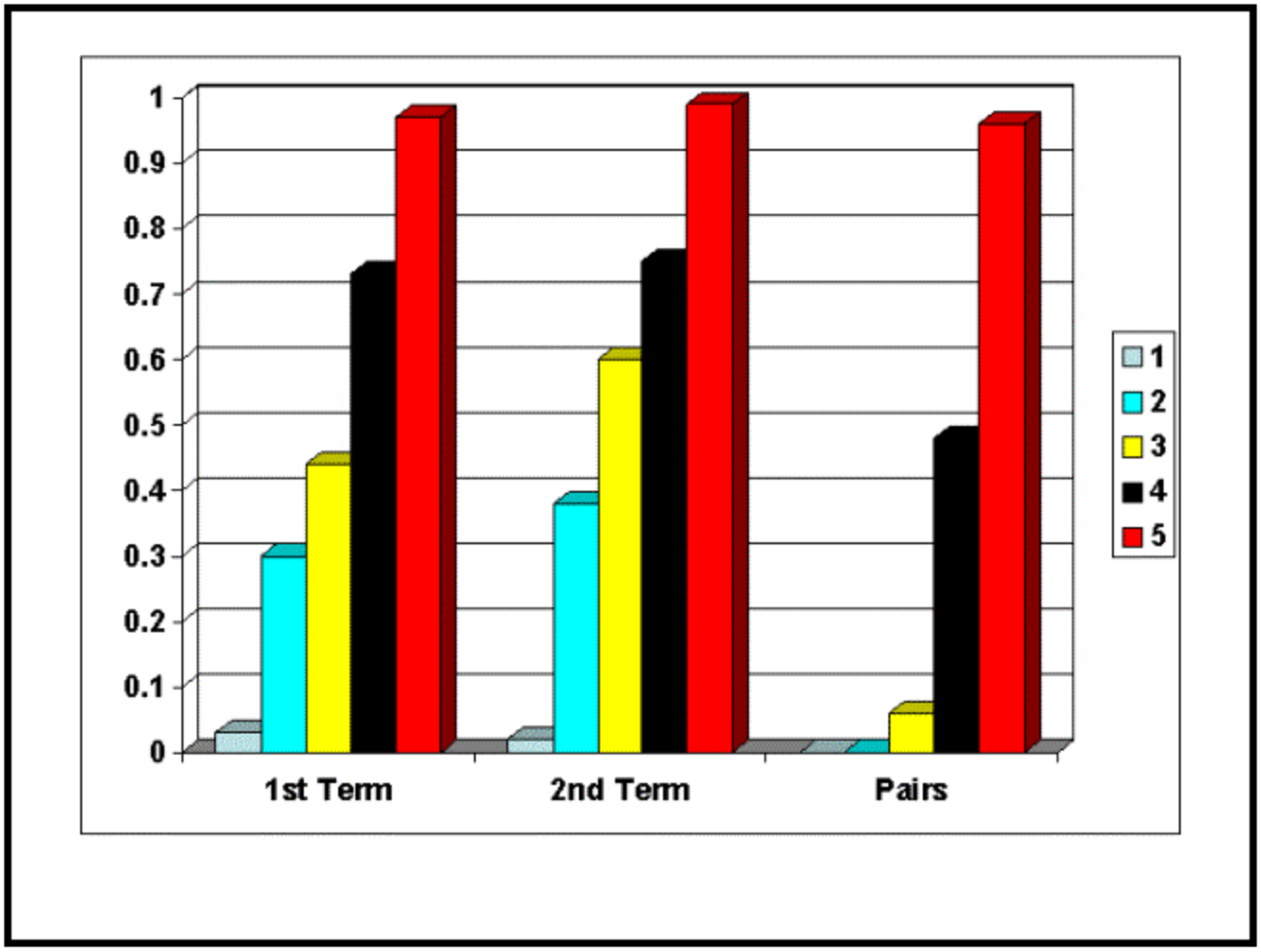

Figure 4. Median Proportions of Matching for 1st, 2nd and Pair of Designated Terms.

Figure 4 summarizes the median proportions of matches associated with five trials demonstrating differing degrees of accuracy in identifying individual terms and pairs. The figure shows the proportion of matches for the first term, the second term and the pair of terms as the matching criteria change. These criteria were set to yield comparable values in the recognition of first and second designated terms. This simulation involved 10 searches of 100 documents in each trial.

Discussion

This study explored two different search philosophies using computer simulation. The traditional approach forces the user to guess at the terms that might represent the content of the documents of interest. Even if the terms are present, the use of these by the authors remains unknown. This approach has been called the *a posteriori* search process because the actual terms and their use can be determined only **after** the documents are retrieved.

Articles have challenged the effectiveness of the indexing process supporting the *a posteriori* system. Flaws have been described associated with the selection of keywords or subject headings. [Bernier 1961; Funk 1983; Kiernan 2004; Spink 2000; Weinberg 1988] The fallibility of manual methods of reading, interpreting and then, selecting terms has been suggested. Computerized indexing also has been challenged because the terms can be used in different context and the computer software may not be able to differentiate the distinctions. The latter is an important consideration as keyword identification by computer software forms an integral component in execution of search engines. [Aguillo 2000; Ding 2001; Gondy 2003; Lowe 2000; Powell 2003; Rappoport 2003; Spink 2003]

The title of this report suggests that there is a need for change. When the technology was not adequate, the emphasis on manual processing was appropriate. However, methods have existed for at least a decade [Weiner 1997] enabling computerized identification and extraction of pertinent textual data from scientific abstracts. The time required to effectively extract, organize and display the data is about 0.6 minutes per 200-300 term document. Since this outlay is a one-time expense, and comparing that time cost with the hundreds of hours expended by users, feasibility is assured. Given viable, economical solutions, is the *a posteriori* search philosophy indeed "broken"? A recent survey [Kiernan 2004] suggested that it is, in the opinion of responding professors.

This situation easily could be changed using available computer technology capable of reading the documents, identifying the vocabulary used by the authors, and determining the ideas presented in the text. With knowledge of the terms used by the authors and **how** these are used in the text, the retrieval process is greatly simplified and enhanced. This process is called the *a priori* search system because the user has full prior awareness of the correct terms. With this approach, the user can spend fulltime effort in using the relevant information. This transition from a clerical, mechanical mode to a cognitive one may have additional benefits in terms of more rapid solution of complex problems. [Weiner 2004]

The simulation analyses raise a question regarding the need for change. With identification of documents occurring 80% or more of the time, is the present approach really broken? Clearly, in a time when the information methods were less robust, the answer would be "no." However, with the capability to identify the exact terms and linkages used by the authors, why not take advantage of this? If "yes" designers of present-day search software must make minor changes. The records of the bibliographic databases or web pages do not require change, although the records inspected by the software might be modified. In bibliographic databases, the software already can search different specified fields and, as such, the title and abstract can be analyzed. In web-

pages, the text that could be used is available. The software must change from the field containing keywords to the one containing the descriptive text.

By simply changing the software to identify the provided sentences in the text and then the informative terms in those sentences, the couplets provided by the authors can be stored for use. The recognition of informative terms can be managed essentially by criteria included in the software. The presentation format would require changing. [see for example, Weiner 1997].

Summary

Individual examples of search failure abound. These examples served to develop scenarios leading to the development of simulation models giving probability distributions. The software Resampling Stats was used in this study. This simulation language was used to explore the consequences of two search philosophies. The findings support the notion that the *a posteriori* search process is a random walk through an unknown field. The ability to recognize individual terms (barring various sources of error) does not imply that the associated document will be correct. The simulation exercise suggested that the selection of documents with an *a posteriori* Boolean AND combination was comparable to the product of independent probabilities. Increasing the accuracy in term recognition to 90% or better, the recognition of pairs improves but not to the degree found in the *a priori* process.

References

Aguillo I. A new generation of tools for search, recovery and quality evaluation of World Wide Web medical resources. *Online Information Review*. 2000; 24(2): 138.

Allen BL. Boolean browsing in an information system: An experimental test. *Information Technologies and Libraries.* 2001; 20(1): 12-20.

Bader JL, Theofanos MF. Searching for cancer information on the internet: analyzing natural language search queries. *J Med Internet Res.* 2003 Dec 11; 5(4): e31.

Bernier CL. The indexing problem. *Journal of Chemical Documentation.* 1961; 1(November): 25-27.

Blair DC. Indeterminacy in the subject access to documentation. *Information Processing and Management.* 1986; 22(2): 229-241.

Brinn S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems.* 1998; 30(1-7): 107-117.

Bruce P, Simon JL, Oswald T. *Resampling Stats Users Guide.* Virginia: Resampling Stats; 2000.

Ding W, Marchionini, G. A comparative study of Web search service performance. *Proceedings of the 59th Annual Meeting of the American Society for Information Science.* Silver Spring, MA: ASIS; 1996: 136-142.

Funk ME, Reid CA. Indexing consistency in Medline. *Bulletin of the Medical Library Association.* 1983; 71(2): 176-183.

Gondy L, Lally AM, Chen H. The use of dynamic contexts to improve casual internet searching. *ACM Transactions on Information Systems.* 2003; 21(3): 229-253.

Horowitz RS, Fuller SS, Gilman NJ, Stowe SM, Weiner JM. Concurrence in content descriptions: author versus medical subject headings (MeSH). *Proceedings of the American Society for Information Sciences.* 1982; 19: 139-140.

Kiernan V. Professors are unhappy with limitations of Online Resources, Survey finds. *The Chronicle of Higher Education.* 2004(April); A-34.

Liu F, Yu C, Meng W. Personalized Web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering.* 2004; 16(1): 28-39.

Lowe, D. Improving Web Search Relevance: Using Navigational Structures to Provide a Search Context. *Proc. 6th Australian World Wide Web Conf., 2000.* Available at: http://ausweb.scu.edu.au/aw2k/papers/lowe/index.html. Accessed Sept. 8, 2004.

Maron ME. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science.* 1977; 28(1): 38-43.

Markov J. The coming search wars. *The New York Times.* Feb 1, 2004.

Microsoft's bCentral Submit it. Search Engine Marketing for Small Business. Available at: http://www.submit-it.com/. Accessed Sept. 8, 2004.

Piniewski-Bond JF, Buck GM, Horowitz RS, Schuster JHR, Weed DL, Weiner JM. Comparison of Information Processing Technologies. *J Am Med Inform Assoc.* 2001; 8: 174-184.

Powell AL, French JC. Comparing the performance of collection selection algorithms. *ACM Transactions on Information Systems.* 2003; 21(4): 412-445.

Rappoport A. The state of search. *Searcher:Magazine for Database Professionals.* 2003; 11(9): 32-37.

Sonkaj H, Koenig S, Schmiede, R. How should libraries respond to new forms of publication? Digital Resources for the Humanities; 2003; University of Gloucestershire.

Spink A. Web Search: Emerging Patterns. *Library Trends.* 2003; 52(2): 299-306.

Spink A, Jansen BJ, Czumltu HC. Use of query reformulation and relevance feedback by Excite users. *Internet Research.* 2000; 10(4): 317.

Voss D. Better searching through science. *Science.* 2001; 293(5537): 2024-2.

Warner AJ. Quantitative and qualitative assessments of the impact of linguistic theory on information science. *Journal of the American Society for Information Science.* 1991; 42(1): 64-71.

Weinberg BH. Why indexing fails the research. *The Indexer.* 1988; 16(1): 3-6.

Weiner JM, Schuster JHR. *XXIV Century Press.* 1997. Available at: http://www.xxivcentury.com. Accessed Sept. 8, 2004.

Weiner JM, Schuster JHR, Horowitz RS, McAfoos WP, Piniewski-Bond JF. *Fantasies in Processing.* Baltimore, American Literary Press; 2004.

Weiner JM, Shirley S, Gilman, NJ, Stowe SM, Wolf RM. Access to data and the information explosion-oral contraceptives and risk of cancer. *Contraception* 1981; 24(3): 301-313.

Back to Contents

http://southernlibrarianship.icaap.org/content/v05n02/weiner_j01.htm.