DYNAMIC INFORMATION AND LIBRARY PROCESSING

GERARD SALTON

Cornell University

PRENTICE-HALL, INC.
Englewood Cliffs, N J 07632

xiv + 523 pages
$19.95

ISBN 0-13-221325-7

REVIEWED BY RONALD E. WYLLYS

University of Texas Austin 78712

First, an overall characterization of the book It is an outstanding work. Time may well establish it as a masterpiece Salton has succeeded in combining. (1) the presentation of an interesting and, more importantly, a challenging concept--the "dynamic library" --toward which he believes libraries and information agencies ought to direct their research, developmental, and organizational efforts, (2) extensive guides to the relevant literature in several fields, through late 1974, (3) a textbook for at least two semester-length courses, for which my suggested titles would be "Language Processing for Information Storage and Retrieval" and "Library Systems Analysis", plus a good part of a third semester on "Library Automation", and (4) an invaluable reference work for computational linguists, information scientists, and librarians.

Now to the details. Since this review is being prepared for the American Journal of Computational Linguistics, it will be presented in a somewhat unusual format. Instead of beginning at the beginning of the book, I shall start by discussing what seem to me to be the book's highlights for the computational linguist Only after that discussion shall I deal with the general plan of the book and with other specific parts of it

The book's ten chapters are intended to be capable of being read independently of one another, although most readers will want to peruse Chapter 1 ahead of any other in order to understand Salton's underlying theme for the book Of the ten chapters, those most immediately relevant to computational linguistics are undoubtedly the last two, plus Chapter 3. The last two are part of a section called "Dynamic Information Processing," in which Salton connects basic concepts in file organization and language processing with their potential applications in the dynamic library (about which more is said below) At the heart of computational linguistics, Chapter 9, "Language Processing," condenses into 49 pages a frank evaluative review of the state-of-the-art in this field Salton links the research in the field with its potential for applications to information systems by saying:

> A content analysis system going beyond the identification of individual terms . . . requires at least three parts: a *description* of an area of discourse in terms of basic entities, or concepts, of importance in this area, including also the main logical-semantic relationships that must be identifled between these entities; a *linguistic*

> *theory* based on appropriate characterizations of lexical items and on grammatical and semantic rules that would underlie the language analysis system; a *set of procedures* capable of generating for each acceptable input string a deep structure specifying the linguistic-semantic relations between entities obtained from the linguistic analysis, as well as the logical-semantic relations derived from the encyclopedia.

With the stage thus set, Salton presents a moderately detailed and highly readable overview of recent and current approaches to natural-language analysis, with an ample supply of examples The chapter's bibliography can serve as a list of the highlights in computational linguistics during 1963-1973 Not everyone will agree with Salton's somewhat pessimistic view of the usefulness of computational linguistics for information systems in the near future (i e , 10-20 years), but all will find this chapter a masterful presentation

In Chapters 3 and 10, Salton enlarges the horizons of computational linguistics beyond its most frequent area of concern, for whcih one might better use the narrower name of "algebraic linguistics", by discussing what I like to call "quantitative linguistics" --another part of computational linguistics, broadly considered Chapter 3, "Automatic Indexing and Abstracting", treats methods by which it is possible to assess programmmatically the probable usefulness of words and phrases as indicators of the content of documents. Such methods are primarily, but by no means exclusively, statistical, and the discussion includes syntax-analytic methods Salton takes pains to dispose of the standard criticism of automatic indexing as "imperfect", by arguing that

> The assertions concerning the inadequacy of automatic indexing are often bolstered by demonstrations designed to show that the results of certain specified automatic procedures will fail to pass any rational test carried out by independent human observers. And from such demonstrations one concludes that the quality obtained through automatic indexing methods is inferior to that of indexing by specialists.
>
> The trouble with these arguments is that a correct premise--that most automatic indexing products are imperfect--leads wrongly to the conclusion that the automatic product is necessarily inferior to one obtained intellectually by human experts.

He concludes that although "it is hazardous to extrapolate test results obtained in a laboratory environment to operational situations involving possibly hundreds of thousands of items", nevertheless, a number of different, independent tests--several of which he discusses-- have shown that "relatively simple automatic text analysis systems do not produce in a document retrieval environment search results inferior to" those of conventional manual indexing

As befits a final chapter, Chapter 10, entitled (like its superordinate) "Dynamic Information Processing", shows how the theories and techniques developed earlier in the book can be applied to the book's main theme, the dynamic library. As Salton puts it

> In this chapter the characteristics of on-line retrieval systems are taken up with emphasis on novel procedures not now implemented in operational situations in which suitable interactions between users and system may be particularly beneficial. Covered in particular are indexing methods adapted to particular (possibly changing) document collections, thesaurus construction and manipulations, search procedures based on the use of feedback information supplied by the customer population during the search operations, document space

> modification methods in which the document characterizations are changed in accordance with experiences accumulated in the course of operations, and collection growth and retirement procedures. Various methods are suggested for these tasks, and evaluation results are given whenever they are available.

Computational linguists may be especially interested in the treatments of how to construct indexing vocabularies and of how to construct, maintain, and manipulate thesauruses. Both of these treatments cover syntax-analytic, as well as quantitative, techniques But all readers will find much of interest in the chapter's combining the foregoing treatments with such ideas as the on-going modification of both queries and document index-term sets, to improve not only the retrieval of documents but also the management of the collection as a whole

Having dealt with the chapters that I suggest will be of primary interest to computational linguists, we can now examine the book as a whole Salton states that his overall purpose in the book is

> to bridge the gap between computer science and information science by introducing a new environment, called the dynamic library, and a set of dynamic information processing tasks to operate in that environment. The idea is to carry out most processing tasks, such as content analysis, classification, information search, and retrieval, interactively under user control, while simultaneously accommodating the file updating and maintenance procedures that are inherent in a changing data processing situation.

The key to achieving the goals of the dynamic library is the use of the "clustered file" concept Since this may not yet be a completely familiar concept, it deserves discussion here In

Salton's words

> . . . a *clustered file* organization is recommended in which documents carrying somewhat similar content descriptions are automatically grouped into clusters. Each cluster is identified by a representative cluster profile, or *centroid*, somewhat akin to the center of gravity of a set of mass points. A cluster centroid is simply a weighted set of terms derived from the document vectors (index-term sets) included in the corresponding cluster.
>
> A clustered file is then similar in concept to a normal classified library file except that the document classes are automatically generated and some overlap may exist between classes, that is, certain documents may be included in more than one class. Furthermore, in the case of the clustered file it is easy to rearrange the cluster composition by moving documents from one class to another if it should prove useful.
>
> A search in a clustered file is carried out in several steps: first, each query is compared with the index file of centroid vectors; then, for those centroids exhibiting a sufficiently high similarity with the query, the individual document vectors in the corresponding clusters are examined, and the document citations are ranked for output purposes in decreasing query-document order. . .
>
> It is clear that the "depth" of the search, as measured by the number of query-document comparisons, can be controlled in a clustered file because it is possible to search only the "best" cluster --the one exhibiting the highest query-centroid similarity--or the top two clusters, or the top ten, as may be required. Moreover, since all document vectors and citations belonging to a given cluster are stored adjacently in the same storage area, for example, on the same track or cylinder of a given disc assembly, only one access operation is needed for each document cluster, as opposed to one access for each document citation in an inverted file.
>
> A detailed comparison of inverted clustered file organizations shows that the clustered file is more economical of storage, leads

> to faster retrieval operations, and permits more flexible search strategies.

This important concept of the clustered file is discussed in detail in Chapter 8, "Automatic Document and Query Classification"

The use of clustered files makes it practical to "maintain the library system in a continuous state of flux"--i e , to make it a dynamic library--by facilitating query processing in which both query vectors and document vectors are continually subjected to small changes As its vector changes accumulate, a document's "classification", i e , its cluster, may change As document changes accumulate, a cluster's centroid may change.

The book as a whole, then is devoted to expounding the theme of the dynamic library and to explicating the necessary details Chapter 1, "Introducing the New Library", does just what its title says, and, as indicated earlier, most readers will want to peruse it to obtain a more detailed idea of what Salton means by the name "dynamic library". In this chapter he argues that libraries present data-processing requirements that are unique as a combination of very large size, high level of file activity, great variety of different operations to be performed, large volume of input and output operations, and need for real-time control He reviews attempts to solve library problems by mechanization and by co-operation, concluding that such efforts can offer no more than partial solutions. As a different approach, he proposes his concept of the dynamic library

In Chapter 2, "Mechanized Housekeeping", Salton provides an excellent overview of the present state of library automation in the areas of cataloging, serials control, and circulation control Chapter 4, "Storage and Retrieval Systems", continues this overview into the areas of reference service, current-awareness systems, and information centers and networks.

Chapter 5, "Library Systems Analysis", deals much too briefly, in my opinion, with systems analysis as such, but it does offer a very readable presentation of the ideas of bibliometrics and of operations-research techniques applied to libraries. In Chapter 6, "System Testing", the difficult problem of evaluating information systems is discussed, the chapter includes a concise treatment of cost-effectiveness and cost-benefit analysis.

Finally, Chapter 7, "Storage Organization", provides an excellent summary of computer-file structures Any instructor who finds that his or her students tend to become overwhelmed by details when they read Knuth should offer them this chapter as a highly readable introduction to file-organization methods Attention is given to the special problems of library files, and the clustered-file concept is introduced. The chapter concludes with a look at some special-purpose file-organization techniques

Salton suggests that the book's chapters could be used for two semester-length courses, as follows, with "()" indicating optional chapters and "_" more advanced topics

For computer-oriented students	1,3,4,(5),7,8,(9),10
For information-science-oriented students	1,2,3,4,5,6,(9),10

My feeling is that each of these sequences is too long for a one-semester course, assuming a reasonable amount of additional reading assignments and exercises for the students My suggestions for courses and chapters are these

A course in "Language Processing for Informaton Storage and Retrieval"	1, 3, 9, 10, 8, (7)
A course in "Library Systems Analysis"	1, 5, 6, 2, 3, 4
As part of a course in "Library Automation'	1, 2, 4

Every reviewer finds a few nits to pick I wish Salton had not used the abbreviation "log" for "natural logarithm" instead of the now standard "ln", or at least that he had explicitly stated his usage A few of the tables contain minor numerical errors, none that I noticed affects the conclusions being drawn, and at least one ("18" instead of "22" in Table 1-2) when corrected strengthens the argument I wish Salton had dealt less curtly with systems analysis, but, after all, the book contains 537 well-filled pages as it stands

In conclusion, I think it likely that this book will come to be viewed as a master contribution to the professional and pedagogical literature in natural-language analysis, information science, and library science