# Human Information Retrieval

Julian Warner

The MIT Press
Cambridge, Massachusetts
London, England

© 2010 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Sabon by The MIT Press.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Warner, Julian, 1955–
Human information retrieval / Julian Warner.
p. cm.—(History and foundations of information science)
Includes bibliographical references and index.
ISBN 978-0-262-01344-4 (alk. paper)
1. Information retrieval. I. Title.
ZA3075.W375 2010
025.04—dc22
2009010120

10 9 8 7 6 5 4 3 2 1

# 1

# Introduction

Information retrieval is of high contemporary significance, diffusing into ordinary discourse and everyday practice. Recently information retrieval has changed rapidly, particularly through the influence of Internet search engines. Practical understanding of how to use systems has been in advance of fuller theoretic understanding, particularly when concerned with transformations of meaning. The breadth of the area—the variety of disciplines that have and potentially could contribute—may have inhibited the development of understanding. A deep divide remains between centrally relevant activities and their associated cultures, between library and information science and Internet search engines. The limited theoretical development implies a need for a deeper—more comprehensive and inclusive—understanding of information retrieval, which should reveal structure and underlying patterns in a complex, differentially understood, and apparently chaotic area. Ideally, a deeper understanding or theory should be congruent with practical understanding and everyday practice. It should also be explicitly articulated, yielding knowledge which can be returned to the real world to inform deliberate intervention in system design and use rather than unconsciously reproducing patterns of activity.

The overall intention is to develop a labor theoretic approach to information retrieval. The immediate concern in this chapter involves the initial components for this approach. This chapter also reviews existing evaluative traditions and indicates the possibility for synthesis within a labor theoretic approach. Chapter 2 introduces and develops crucial concepts of selection power and selection labor, both as concepts and activities in themselves and for the relation between them.

Labor, choice, and technology are fundamental to human experience. In the Judeo-Christian tradition, once out of Eden we are condemned to labor and compelled to choose. Technology may have been noticed less explicitly, but it is equally pervasive in post-Edenic experience, both as agrarian and industrial—or *productive*—and information technologies. Physical and mental labor are usually considered separately from each other, but acknowledgment of the mental components of physical labor and physical elements in mental labor have moved recently toward synthesis, and an emerging view of intelligence concerns "quality of our bodies as much as our minds" (Gosden 2003, 31–33, 119). Mental or informational labor has been recognized as both an independent activity and an adjunct to obtaining physical control over the environment (Webster 2002, 15). Types of mental labor have been differentiated, with semantic labor distinguished from syntactic labor (Warner 2005a). Classically from Aristotle, choice, or deliberation, is the product of mental labor. Late-twentieth-century developments in information technology constitute a revolution in the mechanization of mental labor (Minsky 1967, 2), embodied in the computer as a universal information machine that developed from mid- and late-nineteenth-century antecedents in special-purpose information machines. Both productive and information technologies are human constructions—the products of human physical and intellectual labor drawing upon natural resources and preexisting human constructions (Warner 2004, 5–35). An understanding of information retrieval constructed from labor, choice, and technology promises to be deeply rooted in human experience and to offer a radical depth of understanding.

Power in explanation can be demonstrated by the ability to absorb elements from previous models as special cases of the new model, indicative of the history of a true science, while discarding those elements that have obstructed understanding.

## Existing Models

Information science has developed existing evaluative models that should be absorbed into the new model and offer some elements for synthesis and advancement, with diffusion into computer science, librarianship, and indexing. Some discussions of the information society are concerned with informational or mental labor, and this offers a more indirect resource that can also be absorbed.

### Information and Computer Science

The dominant tradition for evaluating information retrieval systems in information science emerged nearly simultaneously and partly independently in the United States and in the United Kingdom during the early- to mid-1950s (Ellis 1996, 1–22) and has since diffused to and been partly absorbed by computer science. The techniques developed for selection and ordering of references and documents have served as exemplars, or demonstrations of possibility, for the increasingly dominant Internet search engines, with some elements of more direct transfer or inheritance in search algorithms derived from information retrieval research (Ellis and Vasconcelos 1999, 8). There were also parallel developments in commercial search services, largely independent in concept but drawing on common technologies, which also served as exemplars and demonstrated commercial feasibility and technical possibility.

The information and computer science tradition has not always been explicit about its own values or examined its own assumptions; without full notice, it has sometimes departed pragmatically from its initial assumptions. It can, however, be broadly characterized as query transformation, with the query articulated verbally in advance of searching and then transformed by a system into a set of records (Heine 1980). The central value of query transformation can be summarized as a system that should deliver all and, insofar as possible, only all records relevant to a given query. Retrieved records are assessed for their relevance to the query and before generating measures of performance, including precision and recall. The adopted methodology induces a bias toward fixing and possibly reifying relevance, reducing it from a concept to a relation between query and document (Ellis 1984, 28–29; 1992; 1996, 11–20). Bibliographic systems, rather than full text, have been the dominant—although not exclusive—subject of study. Humanly assigned indexing has tended to be received as a given, leaving the rationale for such indexes—and their associated labor and costs—unexplored. There has also been an implicit teleology that aims for a perfect system. Evaluation has become an end in itself, sometimes obstructing understanding (Ellis 1984; 1992).

The relation between information technology and the research tradition in information and computer science could be characterized as repression, wherein the repressed reemerges but not at a fully conscious level. Repression is discernible in the insistence that the retrieval processes created and studied are independent of their particular technological instantiation while simultaneously allowing procedures to be strongly determined by contemporary technological possibilities. For instance, the stress on query transformation corresponded to the batch processing embodied in the technology of the 1950s. The theoretical legacy of query transformation has proved difficult to adapt to modern systems, which do not necessarily demand a verbally articulated query in advance of searching and which can be—and are—used interactively. Critiques argued that the assumption of the necessity for a verbally articulated query was intratheoretic (Heine 1980) rather than intrinsic to information seeking; this argument has been substantiated by changes in practice enabled partly by subsequent technological developments. Reemergence of the repressed can be found in the late articulation and still-limited acknowledgment of the identity between primitive operations of information retrieval and logic or computation. Analysis has revealed that the potential transformations for information retrieval on written records or descriptions are variations on primitive operations of sorting or partitioning and the transformation of one symbol into another (Buckland and Plaunt 1994). This can be a regarded as a special case of the known potential for reducing mathematical and logical operations on an object language to the writing, erasure, and substitution of symbols (Ramsey 1925/1990, 165–174) and also corresponding to the primitive computational operations (Warner 1994, 102–103). The paradigm of query transformation can be regarded as largely but not entirely exhausted, becoming increasingly distant from the empirical reality of interactive and distributed systems (Ellis and Vasconcelos 1999, 8), exposing its rigidity if the original distinctions are retained, or surviving by ad hoc modifications to its theoretical base and thereby losing relevance in the first direction and internal intellectual coherence in the other.

Two paradigms—the cognitive and the physical—have been distinguished in information retrieval research, but they share the assumption of the value of delivering relevant records (Ellis 1984, 19; Belkin and Vickery 1985, 114). For the purposes of discussion here, they can be considered as a single heterogeneous paradigm, linked but not united by this common assumption. The value placed on query transformation is dissonant with common practice, where users may prefer to explore an area and may value fully informed exploration. Some dissenting research discussions have been more congruent with practice, advocating exploratory capability—the ability to explore and make discriminations between representations of objects—as the fundamental design principle for information retrieval systems.

We can acknowledge the utility of techniques developed for selection and ordering of references and documents—in both the experimental tradition and commercial practice—and simultaneously recognize that these techniques are derived from known fundamental computational operations. The techniques have been realized in the special-purpose tools and machines used for information retrieval at the beginning of the research tradition in the 1950s and by the programmed universal information machine of the modern computer. We can fully acknowledge technology rather than repress it and still make a distinction between techniques and values in order to preserve and carry forward what may be valuable from information retrieval research in information and computer science.

## Librarianship and Indexing

Compared with the research tradition developed in information science and subsequently diffused to computer science, the historical antecedents for understanding information retrieval in librarianship and indexing are far longer but less widely influential today. They have tended to be less explicit about their evaluative criteria and aims for information retrieval systems, and far less concerned with producing measures of effectiveness. In contrast to information and computer science, they have been associated with the technologies of writing and printing and have had a pronounced preference for direct human description of information objects. Although less immediately pronounced, we can discern a similar pattern of repression regarding technology. The need for descriptions less extensive than the documents described, imposed by storage constraints of inscribed media, and for direct human intervention in the creation of these descriptions, connected with the technical characteristics of writing and printing, have tended to be universalized and treated as if they were independent of their dominant technological realizations (Wilson 2001). The information and computer science tradition probably inherited the assumed need for brief index descriptions directly from existing information products and not from theories that informed the construction of those products (Cleverdon 1962; Cleverdon, Mills, and Keen 1966). A further limitation of library studies involves its focus on training in the use of information retrieval systems, which often concentrates on the level of system commands rather than understanding their value in communication (Roberts 1989). Disturbing evidence suggests that formal information retrieval systems are marginal in communication (particularly scholarly communication), especially in the sense of information, topic, or subject retrieval rather than document identification (known author or title) and supply (Bath University Library 1971; Smithson 1994).

Two valuable elements are carried forward from librarianship and indexing. The first is a partly implicit stress on selection power, conceived as bibliographic control in librarianship (Wilson 1968) and implied by valuing the discriminatory power of index terms in indexing but made fully explicit here. The second is an acknowledgment of the role of direct human intellectual labor in creating selection power, transforming into a fuller understanding distinguished from specific technological constraints and their partly covert influence on theory and practice.

### Information Society Discussions

Information society discussions have given some rather limited attention to information retrieval. For instance, Lyotard comments:

> It is reasonable to suppose that the proliferation of information-processing machines is having, and will continue to have, as much an effect on the circulation of learning as did advancements in human circulation (transportation systems) and later, in the circulation of sounds and visual images (the media). (1984, 4)

Other comments remain similarly unfocused, recognizing the significance of information retrieval but not providing full research or intellectual context for its consideration. In particular, some information society discussions have treated technology unsatisfactorily (Webster 2002), possibly due to wariness about being stigmatized as technologically determinist (Wilson 1996a), and there has been a limited understanding of fundamental computer operations. An analytically valuable category of informational labor has begun to be distinguished by some writers (Webster 2002, 15); this book will adopt and further differentiate that distinction, acknowledging the possibility of transferring some forms of mental labor to information technology.

**Summary**

Different elements from the information retrieval tradition developed in information science from librarianship, indexing, and information society discussions will be selected and carried forward. The utility of the techniques developed by information retrieval research—but not the associated value of query transformation—is acknowledged, with the recollection that techniques are variations on primitive computational transformations. Selection power is adopted from librarianship and indexing as the primary value and the role of direct human labor is both substantiated and critiqued. Informational labor transformed into mental labor to incorporate its historical antecedents is derived from information society discussions. Technology is restored, not repressed, and understanding of the types of mental labor transferable to information technology is informed by the distinction between semantic and syntactic mental labor. A synthesis of existing approaches is envisaged, producing a set of concepts and categories that are simultaneously simpler and more powerful than the query transformation of classic information-retrieval research, more explicit and discriminating than librarianship and indexing (particularly regarding the significance and costs of human mental labor), and fuller and more technologically informed than information society discussions.

This book adopts an inclusive understanding of information retrieval systems, developed from common understandings and conveyed by ostensive exemplification rather than restrictive definition. In particular, the common antithesis between experimental and operational systems is dissolved. The real source of contrast between the types of system likely has been different forms of the description process, particularly the experimental preference for machine generation rather than human selection of index terms and for non-Boolean searches for those descriptions. When made explicit, the basis for the distinctions between experimental and operational information retrieval systems appears theoretically weak. The distinction is being increasingly eroded in practice, with operational systems possibly selecting records or documents by directly Boolean operations but ordering retrieved documents on the basis of other indicators.

If the proposed model is to be regarded as a scientific advance, it must have a dual aspect that comprehends empirical reality and selectively absorbs existing models. Empirical reality should be explained as fully, powerfully, and as parsimoniously as possible. The pervasive presence of labor, choice, and technology in information retrieval practice promises a strong degree of correspondence to empirical reality. Human labor is immediately present as the description labor of cataloging, classification, and database description. Choice has been persistently embodied in practice and, more recently, increasingly recognized theoretically and valued as both selection from retrieval results and the filtering of information. Diffused from the 1950s, modern information technologies, to which aspects of human mental labor can be and increasingly are transferred, are now pervasive in information retrieval.

The reader can now anticipate this book's approach and structure.

The encompassing theories adapted for development and the mode of presentation derives from the human sciences. The author understands human history as a cumulative progression: human labor acting upon the naturally given and humanly modified environment as the primary source for progressive change developed by, but not exclusive to, Marx. Distinctions from Ferdinand de Saussure's *Course in General Linguistics* are adapted for understanding transformations of meaning in full-text retrieval. Material from information theory and computer science is also understood from the perspective of the human sciences. Information theory is utilized to comprehend patterns of occurrence and recurrence of words and phrases. The discussion is informed throughout by an understanding of the computational process derived directly from automata theory, or the theory of computation.

Characteristic of the human sciences, the mode of presentation is primarily discursive and aims to obtain broad intelligibility. Logically expressed schemata are introduced, accompanied by diagrams more often encountered in the formal sciences but intended here to reveal the underlying rigor and economy of the discursively expressed argument, parallel to the discourse rather than independent points of departure. Methodologically, the book deliberately excludes some complexity, particularly regarding full-text retrieval, to enable analytic concentration on central or inescapable issues; it also excludes simplified assumptions that can be false or distorting.

Each chapter develops cumulatively from the preceding chapters.

**Chapter 2: Selection Power and Selection Labor**

Selection power—the ability to make informed choices between objects or representations of objects—is argued from a number of perspectives as the primary aim of information retrieval systems. Similar principles for retrieval are adduced from partly independent discourses, such as the value placed upon an index term's discriminatory power in discussions of indexing and the concept of bibliographic control as mastery over written and published records (UNESCO/Library of Congress 1950, 1). Commonly used systems embody facilities to enhance selection power, and the survival of such systems in the market for information services testifies to the perceived utility of selection power (Swanson 1980, 128). Ordinary discourse comments from consumers of such systems also value control and selection. The concept of exploratory capability developed by critiques of the dominant research tradition is further transformed into selection power, providing more precise and generically applicable analysis. Therefore, understandings embodied in some significant scholarly discourses, in practice and in ordinary discourse, precede theoretical articulation, which continues to value query transformation above selection power in the dominant research tradition. The etymology of intelligence (*inter-legere*: to choose between) implies a link between selection power and both deep and ordinary discourse, or widely diffused aspects of human experience (Stevens 1998, 66). Values for information retrieval are brought into accord with processes by replacing query transformation with selection power.

Selection power is conceived as a fundamental concept that is open to elucidation but not further decomposition into more primitive entities. It is understood as a quality of human consciousness that can be assisted or frustrated by the system's capacity for exploration but is not inherent in the system itself. Under certain historical conditions and levels of technological development, selection power is produced by activities such as cataloguing, classification, description of objects for databases, and searching catalogs and databases, all of which can all be comprehended and understood as selection labor. Thus, selection power is not conceived abstractly, apart from real world circumstances, but operates in relation to considerations of human labor (particularly mental labor), the costs of that labor, and the possibility of transferring particular forms of mental labor to information technology, now primarily computational technologies. A fundamental proposition is developed: selection power is produced by selection labor.

Selection labor is characterized as a form of mental labor and theoretical minima are established for a given collection of objects. The separation of selection labor into description and search labor with the *premodern* technologies of writing and printing on paper is noted. Similarly, the author acknowledges that description and search activities reconverge with computer-based, or *modern*, technologies and also acknowledges the possibility of sustaining analytical distinctions between them.

**Chapter 3: Description and Search Labor**

Selection labor separates historically into description and search labor and can be analytically decomposed. The activities of description and searching are then more fully characterized empirically as components of selection labor. As forms of mental labor, description and search labor participate in the conditions for labor and mental labor. Concepts and distinctions that apply to physical and mental labor are indicated, introducing the necessity of labor for survival, the idea of technology as a human construction, and the possibility of transferring human labor—including mental labor—to technology. Distinctions specific to mental labor, particularly between semantic and syntactic labor, are introduced. The high cost of human mental labor is also indicated.

Exemplified by cataloging, classification, and database description, description labor is more formally understood as the labor involved in transforming objects into searchable descriptions; it includes interpretation. Search labor is understood as the human labor expended in searching systems. Direct human labor has diminished progressively for both description and search labor, and its syntactic aspects have transferred to technology effectively compelled by the high relative costs of direct human labor compared to machine processes.

**Chapter 4: A Labor Theoretic Approach**

The labor theoretic approach to information retrieval that informed chapters 2 and 3 is made fully explicit in chapter 4. The labor theoretic approach has qualities usually desired for a theory that couples comprehensiveness with economy— parsimony, power, and final simplicity. Although the focus is on the computational mode, it acknowledges inheritances from orality and literacy, and the theory can also comprehend oral and written modes. The labor theoretic approach absorbs library and information and Internet activities into a common schema within the computational mode; different aspects of the schema become prominent for each set of activities. The schema developed within the theory is economical, explicitly reduced to a short sequence of clauses, and also represented in diagrammatic form that includes elements of iconicity. A very powerful analysis results from making fully explicit the dynamics that are strongly implicit in current information activities. Once grasped, the theory becomes simple.[1]

**Chapter 5: Retrieval from Full Text**

The labor theoretic approach can account for the existence of full text retrieval, precisely locating significant changes in description and search labor and processes. However, its analytic power is more fully demonstrated in accounting for the existence of changes and precisely specifying their location than by a fuller understanding of those same qualitative changes. As evidenced by the provision and use of phrase searching, practical understanding of transformations of word meaning and the frequency of word sequences has tended to run ahead of theoretical understanding and articulation. Sources acknowledged in retrieval as relevant to understanding word meanings have exposed a misleading concept of language as a nomenclature but have not fully articulated a positive account of the production of meaning in written language. Therefore, fuller and deliberately articulated understanding requires further development. Ideally, further development should remain consistent with the labor theoretic approach, incorporate existing sources recognized as revealing, and coincide with practical understandings; but it also should develop dialectically from these bases and may draw upon further material.

To obtain deeper insight into inescapable issues of semantics and syntax—the production of meaning from written language and replication and differences in words and words sequence—requires further theoretical contexts. Therefore, the principal sources selected are Saussurean linguistics (for understanding semantics) and information theory (for insight into syntactics understood as patterns of replication and difference in written language), both continually informed by the theory of computation. Selection power is understood to be inescapably produced by human selection labor that modulates over time.

### Chapter 6: A Semantics for Retrieval from Full Text

This chapter is concerned with developing a semantics for written language that incorporates crucial distinctions for understanding retrieval from full text. Established and materially rooted categories in Saussurean linguistics, the syntagma and paradigm—the linear sequence of utterance and the network of associations a word acquires outside a particular syntagma—are adapted to a largely unprecedented purpose: an account of the production of meaning from written language. The inheritance of patterns for the production of meaning and the occurrence and recurrence of words and phrases from oral and written language are also acknowledged and placed in dialogic encounter with the possibilities of computation.

### Chapter 7: A Syntactics for Retrieval from Full Text

Understood as the occurrence and recurrence of patterns, the syntactics of written language are equally relevant to understanding retrieval from full text. Material elements from information theory—the message and messages for selection from a source that correspond directly to the categories chosen from Saussurean linguistics—are adapted to gain understanding of patterns in written discourse, which then can be directly and humanly exploited in searching. The theory of computation continues to inform the understanding of patterns of occurrence and recurrence, particularly for operations that effectively include cutting of words from a line of writing, consistent with the fundamental computational operations of writing, erasing, and substituting symbols.

**Table 1.1**
Structure of the book

<table>
<tr><td colspan="5">Approaches to information retrieval</td><td>Chapter 1. Introduction</td><td rowspan="4">Chapter Four. A labor theoretic approach</td><td rowspan="7">Chapter Nine. Conclusion</td><td rowspan="7">Postscript</td></tr>
<tr><td>Selection power</td><td colspan="4">Selection labor</td><td rowspan="2">Chapter 2. Selection power and selection labor</td></tr>
<tr><td></td><td colspan="2">Description labor</td><td colspan="2">Search labor</td></tr>
<tr><td></td><td>Description labor semantic</td><td>Description labor syntactic</td><td>Search labor semantic</td><td>Search labor syntactic</td><td>Chapter 3. Description and search labor</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td></td><td>Chapter 5. Retrieval from full text</td><td rowspan="3">Chapter Eight. Semantics and syntactics for retrieval from full text</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td></td><td>Chapter 6. A semantics for retrieval from full text</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td></td><td>Chapter 7. A syntactics for retrieval from full text</td></tr>
</table>

### Chapter 8: Semantics and Syntactics for Retrieval from Full Text

This chapter focuses on bringing semantics and syntactics back to the examples of retrieval given in chapter 5 and then testing and demonstrating their analytic advantages. A fuller example of retrieval is also introduced and considered from the developed perspective.

### Chapter 9: Conclusion

The conclusion explores the implications of semantics and syntactics developed for preexisting theories of information retrieval, for the labor theoretic approach, and for the practical evolution of Internet search engines.

### Postscript

The postscript addresses the changes in what it means to be human that arise from current developments in information retrieval.

A diagram confirms the structure of the book, indicating the topics covered by each chapter in relation to the categories of the labor theoretic approach developed and adopted (see table 1.1). The absence of a category does not mean that the category has been discarded in later chapters, but rather that it either continues to be incorporated into the theory developed without further significant modification, or that it is no longer highly significant in real world practice or has been analytically excluded to enable clarity of attention. For instance, selection power is incorporated into the further theory developed, while semantic description labor and syntactic search labor have diminished in real world practice for retrieval from full text. Accordingly, both are analytically excluded from the later chapters. The diagram offers a guide for placing individual chapters within the overall argument.

**Box 1.1**
Historical valuing of selection power

Ay, in the catalogue ye go for men;
As hounds, and greyhounds, mongrels, spaniels, curs,
Shoughs, water-rugs, and demi-wolves, are clept
All by the name of dogs: the valu'd file
Distinguishes the swift, the slow, the subtle,
The housekeeper, the hunter, every one
According to the gift which bounteous Nature
Hath in him clos'd; whereby he does receive
Particular addition, from the bill
That writes them all alike;

—Shakespeare. *Macbeth*. c.1606. III.i.91

Macbeth's questioning of the murderers (Shakespeare, 1606/1988, 77–78) indicates the value historically attached to subtlety of distinctions in the language or lexicon of information retrieval systems. In this respect, the passage anticipates the principle formulated in modern discussions of indexing and classification—the value of an index term lies in its discriminatory power—and is consistent with the valuing of selection power (see chapter 2).

Commentaries have glossed "valu'd" as an adjective derived from the noun (value) and not as the participle of the verb (Shakespeare 1606/1988, 77), implying that values attached to objects are analogous to attributes in modern databases and index terms in information retrieval systems. It could also be read as being valued or valuable, giving "[p]articular addition" or added value. A "file" (Shakespeare 1606/1988, 142) is a list or roll, a highly linear technological form.

While regarded as an image of order (Shakespeare 1606/1988, 77), the passage also contains elements of disorder: interactions between the breeds are listed in tension with the hierarchy from "dogs" to breeds of dogs—"shoughs" (shag-haired dogs) lead to "water-rugs" (rough-haired water dogs). The transition from the first-named type ("hounds") to the last ("demi-wolves") represents a move from domestic to semiferal, with a hint of lycanthropy implied by the analogy with types of men. The imposition of hierarchy also associates closely with political tyranny.