TURING 图灵原版计算机科学系列

Springer

Information Retrieval

Algorithms and Heuristics

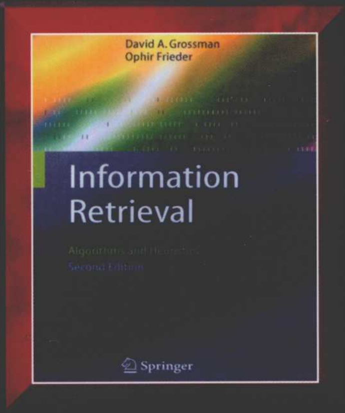
Second Edition

信息检索

算法与启发式方法

(英文版 · 第2版)

[美] David A. Grossman
Ophir Frieder 著



David A. Grossman
Ophir Frieder

Information Retrieval

Algorithms and Heuristics
Second Edition

Springer

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原版计算机科学系列

Information Retrieval

Algorithms and Heuristics

Second Edition

信息检索

算法与启发式方法

（英文版・第2版）

[美] David A. Grossman
Ophir Frieder 著

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

信息检索：算法与启发式方法：第2版＝Information Retrieval: Algorithms and Heuristics Second Edition：英文/（美）格罗斯曼（Grossman, D. A.），（美）弗里德（Frieder, O.）著.—北京：人民邮电出版社，2009.10

（图灵原版计算机科学系列）

ISBN 978-7-115-21225-2

I. 信… II. ①格… ②弗… III. 情报检索－英文 IV. G252.7

中国版本图书馆CIP数据核字（2009）第143915号

内容提要

本书是“信息检索”课程的优秀教材，书中对信息检索的概念、原理和算法进行了详细介绍，内容主要包括检索策略、检索实用工具、跨语言信息检索、查询处理、集成结构化及数据和文本、并行信息检索以及分布式信息检索等，并给出了阐述算法的大量实例。

本书有一定的深度和广度，而且所有的内容都用当前的技术阐述，是高等院校计算机及信息管理等相关专业本科生和研究生的理想教材，对信息检索领域的科研和技术人员也是很好的参考书。

图灵原版计算机科学系列

信息检索：算法与启发式方法（英文版·第2版）

◆ 著　[美] David A.Grossman　Ophir Frieder
责任编辑　杨海玲
◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街14号
邮编　100061　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京铭成印刷有限公司印刷
◆ 开本：700×1000　1/16
印张：21.5
字数：418千字　　2009年10月第1版
印数：1-2 000册　　2009年10月北京第1次印刷
著作权合同登记号　图字：01-2009-4809号
ISBN 978-7-115-21225-2

定价：59.00元

读者服务热线：（010）51095186　印装质量热线：（010）67129223
反盗版热线：（010）67171154

版权声明

For our wives:
Mary Catherine and Nazli
and of course, for
Isaac Sean Grossman,
born 2/2/04.

Foreword

The past five years, as Grossman and Frieder acknowledge in their preface, have been a period of considerable progress for the field of information retrieval (IR). To the general public, this is reflected in the maturing of commercial Web search engines. To the IR practitioner, research has led to an improved understanding of the scope and limitations of the Web search problem, new insights into the retrieval process through the development of the formal underpinnings and models for IR, and a variety of exciting new applications such as cross-language retrieval, peer-to-peer search, and music retrieval, which have expanded the horizons of the research landscape. In addition, there has been an increasing realization on the part of the database and IR communities that solving the information problems of the future will involve the integration of techniques for unstructured and structured data. The revised edition of this book addresses many of these important new developments, and is currently the only textbook that does so.

Two particular examples that stood out for me are the descriptions of language models for IR and cross-language retrieval. Language models have become an important topic at the major IR conferences and many researchers are adapting this framework due to its power and simplicity, as well as the availability of tools for experimentation and application building. Grossman and Frieder provide an excellent overview of the topic in the retrieval strategies chapter, together with examples of different smoothing techniques. Cross-language retrieval, which involves the retrieval of text in different languages than the query source language, has been driven by government interest in Europe and the U.S. A number of approaches have been developed that can exploit available resources such as parallel and comparable corpora, and the effectiveness of these systems now approaches (or even surpasses in some cases) monolingual retrieval. The revised version of this book contains a chapter on cross-language retrieval that clearly describes the major approaches and gives examples of how the algorithms involved work with real data. The combination of up-to-date coverage, straightforward treatment, and the frequent use of examples makes this book an excellent choice for undergraduate or graduate IR courses.

W. Bruce Croft
August 2004

Acknowledgments

The first edition of this book was published in 1998. Since then, numerous advancements in Information Retrieval have occurred.

For help with most of the material for the first edition, we are still deeply indebted to Paul Kantor and Don Kraft for their insightful comments during some of the early stages of the development of the book. We also greatly thank Steve Robertson, K.L. Kwok, and Jamie Callan for critically reviewing the sections describing their efforts, and Warren Greiff whose extreme patience taught and retaught us the details of inference networks.

With the addition of language models and cross-language information retrieval we are extremely thankful for the insightful comments and suggestions from ChengXiang Zhai, Doug Oard, and James Mayfield.

General critical feedback came from multiple individuals. Readability for classroom use was enhanced by feedback given from the students in our undergraduate Information Retrieval course and our graduate Advanced Information Retrieval. John Juilfs spent countless hours checking all of the new examples for this edition and checking our math (we take full responsibility for any remaining errors, but he saved us from lots of them). Abdur Chowdhury carefully reviewed and provided input for our new section on duplicate document detection. Wei Gen Yee provided significant input to our modifications in the distributed information retrieval chapter. Steve Beitzel and Eric Jensen provided significant input to updates in our efficiency and parallel processing sections. Rebecca Cathey provided examples for our new language modeling and cross-language sections. Michael Saelee produced and refined our figures. Finally, Wei Guan carefully and repeatedly checked all of our examples and notation.

We also thank Wendy Grossman, Ben Goldfarb, David Roberts, and Jordan Wilberding for detailed comments. We also thank all others who read the book and provided excellent feedback, and our friends and colleagues who provided moral support.

Finally, we would not have ever finished the book without constant moral support from our families and close friends. Special thanks go to Mary Catherine McCabe and Nazli Goharian who made countless sacrifices of their time and gave us constant encouragement and support.

From both of us to all of you, THANKS.

Preface

When we wrote the first edition of this book in 1998, the Web was relatively new, and information retrieval was an old field but it lacked popular appeal. Today the word *Google* has joined the popular lexicon, and *Google* indexes more than four billion Web pages. In 1998, only a few schools taught graduate courses in information retrieval; today, the subject is commonly offered at the undergraduate level. Our experience with teaching information retrieval at the undergraduate level, as well as a detailed analysis of the topics covered and the effectiveness of the class, are given in [Goharian et al., 2004].

The term *Information Retrieval* refers to a search that may cover any form of information: structured data, text, video, image, sound, musical scores, DNA sequences, etc. The reality is that for many years, database systems existed to search structured data, and information retrieval meant the search of documents. The authors come originally from the world of structured search, but for much of the last ten years, we have worked in the area of document retrieval. To us, the world should be data type agnostic. There is no need for a special delineation between structured and unstructured data. In 1998, we included a chapter on data integration, and reviews suggested the only reason it was there was because it covered some of our recent research. Today, such an allegation makes no sense, since information *mediators* have been developed which operate with both structured and unstructured data. Furthermore, the eXtensible Markup Language (XML) has become prolific in both the database and information retrieval domains.

We focus on the ad hoc information retrieval problem. Simply put, ad hoc information retrieval allows users to search for documents that are relevant to user-provided queries. It may appear that systems such as *Google* have solved this problem, but effectiveness measures for *Google* have not been published. Typical systems still have an effectiveness (accuracy) of, at best, forty percent [TREC, 2003]. This leaves ample room for improvement, with the prerequisite of a firm understanding of existing approaches.

Information retrieval textbooks on the market are relatively unfocused, and we were uncomfortable using them in our classes. They tend to leave out details of a variety of key retrieval models. Few books detail inference networks, yet an inference network is a core model used by a variety of systems. Additionally, many books lack much detail on efficiency, namely, the execution speed of a query. Efficiency is potentially of limited interest to those who focus only on effectiveness, but for the practitioner, efficiency concerns can override all others.

Additionally, for each strategy, we provide a detailed running example. When presenting strategies, it is easy to gloss over the details, but examples keep us honest. We find that students benefit from a single example that runs through the whole book. Furthermore, every section of this book that describes a core retrieval strategy was reviewed by either the inventor of the strategy (and we thank them profusely; more thanks are in the acknowledgments!) or someone intimately familiar with it. Hence, to our knowledge, this book contains some of the gory details of some strategies that cannot be found anywhere else in print.

Our goal is to provide a book that is sharply focused on ad hoc information retrieval. To do this, we developed a taxonomy of the field based on a model that a *strategy* compares a document to a query and a utility can be plugged into any strategy to improve the performance of the given strategy. We cover all of the basic strategies, not just a couple of them, and a variety of utilities. We provide sufficient detail so that a student or practitioner who reads our book can implement any particular strategy or utility. The book, *Managing Gigabytes* [Witten et al., 1999], does an excellent job of describing a variety of detailed inverted index compression strategies. We include the most recently developed and the most efficient of these, but we certainly recommend Managing Gigabytes as an excellent side reference.

So what is new in this second edition? Much of the core retrieval strategies remain unchanged. Since 1998, numerous papers were written about the use of language models for information retrieval. We have added a new section on language models. Furthermore, cross-lingual information retrieval, that is, the posting of a query in one language and finding documents in another language, was just in its infancy at the time of the first version. We have added an entire chapter on the topic that incorporates information from over 100 recent references.

Naturally, we have included some discussion on current topics such as XML, peer-to-peer information retrieval, duplicate document detection, parallel document clustering, fusion of disparate retrieval strategies, and information mediators.

Finally, we fixed a number of bugs found by our alert undergraduate and graduate students. We thank them all for their efforts.

This book is intended primarily as a textbook for an undergraduate or graduate level course in Information Retrieval. It has been used in a graduate course, and we incorporated student feedback when we developed a set of overhead transparencies that can be used when teaching with our text. The presentation is available at *www.ir.iit.edu*.

Additionally, practitioners who build information retrieval systems or applications that use information retrieval systems will find this book useful when selecting retrieval strategies and utilities to deploy for production use. We have heard from several practitioners that the first edition was helpful, and we incorporated their comments and suggested additions into this edition.

We emphasize that the focus of the book is on algorithms, not on commercial products, but, to our knowledge, the basic strategies used by the majority of commercial products are described in the book. We believe practitioners may find that a commercial product is using a given strategy and can then use this book as a reference to learn what is known about the techniques used by the product.

Finally, we note that the information retrieval field changes daily. For the most up to date coverage of the field, the best sources include journals like the *ACM Transactions on Information Systems*, the *Journal of the American Society for Information Science and Technology*, *Information Processing and Management*, and *Information Retrieval.* Other relevant papers are found in the various information retrieval conferences such as ACM SIGIR *www.sigir.org*, NIST TREC *trec.nist.gov*, and the ACM CIKM *www.cikm.org*.

List of Figures

Contents

Chapter 1

INTRODUCTION

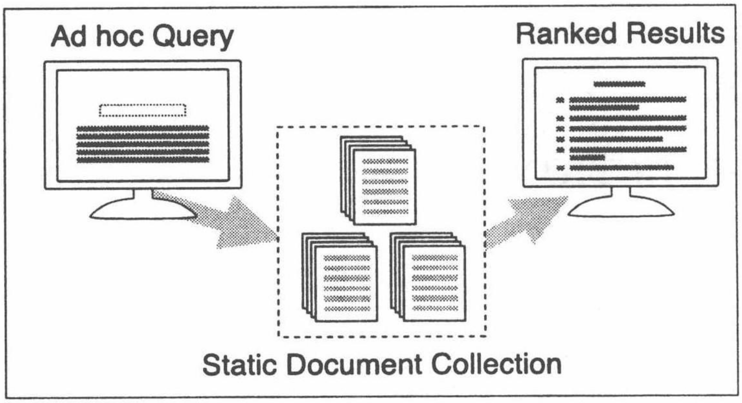
Since the near beginnings of civilization, human beings have focused on written communication. From cave drawings to scroll writings, from printing presses to electronic libraries, communicating was of primary concern to man's existence. Today, with the emergence of digital libraries and electronic information exchange there is clear need for improved techniques to organize large quantities of information. Applied and theoretical research and development in the areas of information authorship, processing, storage, and retrieval is of interest to all sectors of the community. In this book, we survey recent research efforts that focus on the electronic searching and retrieving of documents.

Our focus is strictly on the retrieval of information in response to user queries. That is, we discuss algorithms and approaches for ad hoc information retrieval, or simply, information retrieval. Figure 1.1 illustrates the basic process of ad hoc information retrieval. A static, or relatively static, document collection is indexed prior to any user query. A query is issued and a set of documents that are deemed relevant to the query are ranked based on their computed similarity to the query and presented to the user. Numerous techniques exist to identify how these documents are ranked, and that is a key focus of this book (effectiveness). Other techniques also exist to rank documents quickly, and these are also discussed (efficiency).

Information Retrieval (IR) is devoted to finding *relevant* documents, not finding simple matches to patterns. Yet, often when information retrieval systems are evaluated, they are found to miss numerous relevant documents [Blair and Maron, 1985]. Moreover, users have become complacent in their expectation of accuracy of information retrieval systems [Gordon, 1997].

A related problem is that of document routing or filtering. Here, the queries are static and the document collection constantly changes. An environment where corporate e-mail is routed based on predefined queries to different parts

Figure 1.1. Document Retrieval

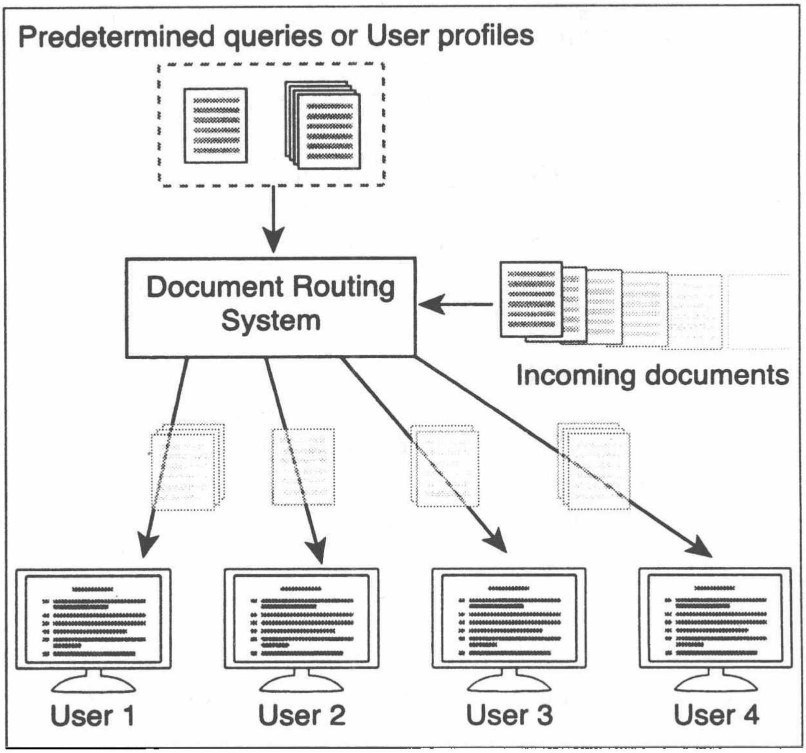


of the organization (i.e., e-mail about sales is routed to the sales department, marketing e-mail goes to marketing, etc.) is an example of an application of document routing. Figure 1.2 illustrates document routing. Document routing algorithms and approaches also widely appear in the literature, but are not addressed in this book.

In Figure 1.3, we illustrate the critical document categories that correspond to any issued query. Namely, in the collection there are documents which are retrieved, and there are those documents that are relevant. In a perfect system, these two sets would be equivalent; we would only retrieve relevant documents. In reality, systems retrieve many non-relevant documents. To measure effectiveness, two ratios are used: *precision* and *recall.* Precision is the ratio of the number of relevant documents retrieved to the total number retrieved. Precision provides an indication of the quality of the answer set. However, this does not consider the total number of relevant documents. A system might have good precision by retrieving ten documents and finding that nine are relevant (a 0.9 precision), but the total number of relevant documents also matters. If there were only nine relevant documents, the system would be a huge success — however if millions of documents were relevant and desired, this would not be a good result set.

Recall considers the total number of relevant documents; it is the ratio of the number of relevant documents retrieved to the total number of documents in the collection that are believed to be relevant. Computing the total number of relevant documents is non-trivial. The only sure means of doing this is to read the entire document collection. Since this is clearly not feasible, an approximation of the number is obtained (see Chapter 9). A good survey of

Figure 1.2. Document Routing

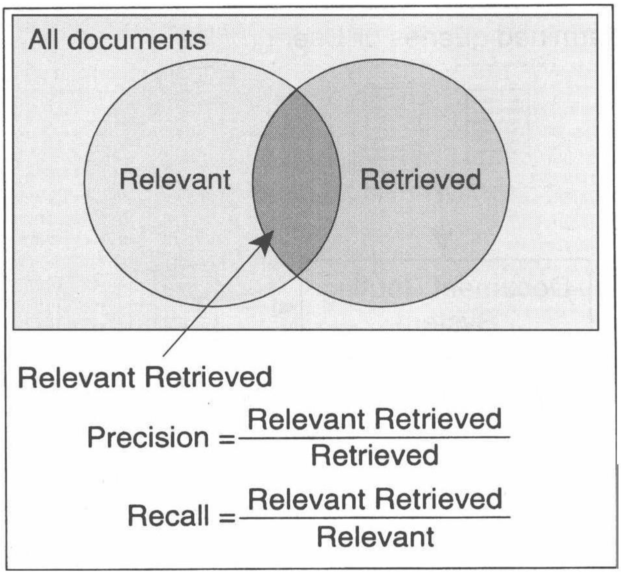


effectiveness measures, as well as a brief overview of information retrieval, is found in [Kantor, 1994].

Precision can be computed at various points of recall. Consider an example query q. For this query, we have estimated that there are two relevant documents. Now assume that when the user submits query q that ten documents are retrieved, including the two relevant documents. In our example, documents two and five are relevant. The sloped line in Figure 1.4 shows that after retrieving two documents, we have found one relevant document, and hence have achieved fifty percent *recall*. At this point, *precision* is fifty percent as we have retrieved two documents and one of them is relevant.

To reach one hundred percent recall, we must continue to retrieve documents until both relevant documents are retrieved. For our example, it is necessary to retrieve five documents to find both relevant documents. At this point, precision is forty percent because two out of five retrieved documents are relevant. Hence, for any desired level of recall, it is possible to compute precision.

Figure 1.3. Result Set: Relevant Retrieved, Relevant, and Retrieved


All documents
Relevant
Retrieved
Relevant Retrieved
Precision = Relevant Retrieved / Retrieved
Recall = Relevant Retrieved / Relevant

Graphing precision at various points of recall is referred to as a *precision/recall curve.*

A typical precision/recall curve is shown in Figure 1.5. Typically, as higher recall is desired, more documents must be retrieved to obtain the desired level of recall. In a perfect system, only relevant documents are retrieved. This means that at any level of recall, precision would be 1.0. The optimal precision/recall line is shown in Figure 1.5.

Average precision refers to an average of precision at various points of recall. Many systems today, when run on a standard document collection, report an average precision of between 0.2 and 0.3. Certainly, there is some element of fuzziness here because relevance is not a clearly defined concept, but it is clear that there is significant room for improvement in the area of effectiveness.

Finding relevant documents is not enough. The goal is to identify relevant documents within an acceptable response time. This book describes the current strategies to find relevant documents *quickly*. The quest to find efficient and effective information retrieval algorithms continues.

We explain each algorithm in detail, and for each topic, include examples for the most crucial algorithms. We then switch gears into survey mode and provide references to related and follow-on work. We explain the key aspects of the algorithms and then provide references for those interested in further