Modern Information Retrieval

Chapter 1

Introduction

Information Retrieval
The IR Problem
The IR System
The Web

Information Retrieval (IR)

- IR deals with the representation, storage, organization of, and access to information items
  - Types of information items: documents, Web pages, online catalogs, structured records, multimedia objects
- Early goals of the IR area: indexing text and searching for useful documents in a collection
- Nowadays, research in IR includes:
  - Modeling, Web search, text classification, systems architecture, user interfaces, data visualization, filtering and languages

Early Developments

- For more than 5,000 years, man has organized information for later retrieval and searching
  - This has been done by compiling, storing, organizing, and indexing papyrus, hieroglyphics, and books
- For holding the various items, special purpose buildings called *libraries*, or *bibliothekes*, are used
  - The oldest known library was created in Elba, in the Fertile Crescent, between 3,000 and 2,500 BC
  - By 300 BC, Ptolemy Soter, a Macedonian general, created the Great Library at Alexandria
  - Nowadays, libraries are everywhere
    - In 2008, more than 2 billion items were checked out from libraries in the US—an increase of 10% over the previous year

Early Developments

- Since the volume of information in libraries is always growing, it is necessary to build specialized data structures for fast search — *the indexes*
- For centuries indexes have been created manually as sets of *categories*, with labels associated with each category
- The advent of modern computers has allowed the construction of large indexes automatically

Early Developments in IR

- During the 50's, research efforts in IR were initiated by pioneers such as Hans Peter Luhn, Eugene Garfield, Philip Bagley, and Calvin Moores, who allegedly coined the term *Information Retrieval*
- In 1962, Cyril Cleverdon published the Cranfield studies on retrieval evaluation
- In 1963, Joseph Becker and Robert Hayes published the first book on IR
- In the late 60's, key research conducted by Karen Sparck Jones and Gerard Salton, among others, led to the definition of the *TF-IDF term weighting scheme*

Early Developments in IR

- In 1971, Jardine and van Rijsbergen articulated the *cluster hypothesis*
- In 1978, the first ACM SIGIR Internation Conference on Information Retrieval was held in Rochester
- In 1979, van Rijsbergen published a classic book entitled *Information Retrieval*, which focused on the Probabilistic Model
- In 1983, Salton and McGill published a classic book entitled *Introduction to Modern Information Retrieval*, which focused on the Vector Model

Libraries and Digital Libraries

- Libraries were among the first institutions to adopt IR systems for retrieving information
- Initially, such systems consisted of an automation of existing processes such as card catalogs searching
- Increased search functionality was then added
  - Ex: subject headings, keywords, query operators
- Nowadays, the focus has been on improved graphical interfaces, electronic forms, hypertext features

IR at the Center of the Stage

- Until recently, IR was an area of interest restricted mainly to librarians and information experts
- A single fact changed these perceptions—the introduction of the Web, which has become the largest repository of knowledge in human history
- Due to its enormous size, finding useful information on the Web usually requires running a search
- And searching on the Web is all about IR and its technologies

Thus, almost overnight, IR has gained a place with other technologies at the center of the stage

The IR Problem

The IR Problem

- Users of modern IR systems, such as search engine users, have information needs of varying complexity
- An example of complex information need is as follows:

  Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)

The IR Problem

- This full description of the user information need is not necessarily a good query to be submitted to the IR system
- Instead, the user might want to first translate this information need into a query
- This translation process yields a set of *keywords*, or *index terms*, which summarize the user information need
- Given the user query, the key goal of the IR system is to retrieve information that is useful or relevant to the user

The IR Problem

- That is, the IR system must rank the information items according to a degree of relevance to the user query
- The IR Problem

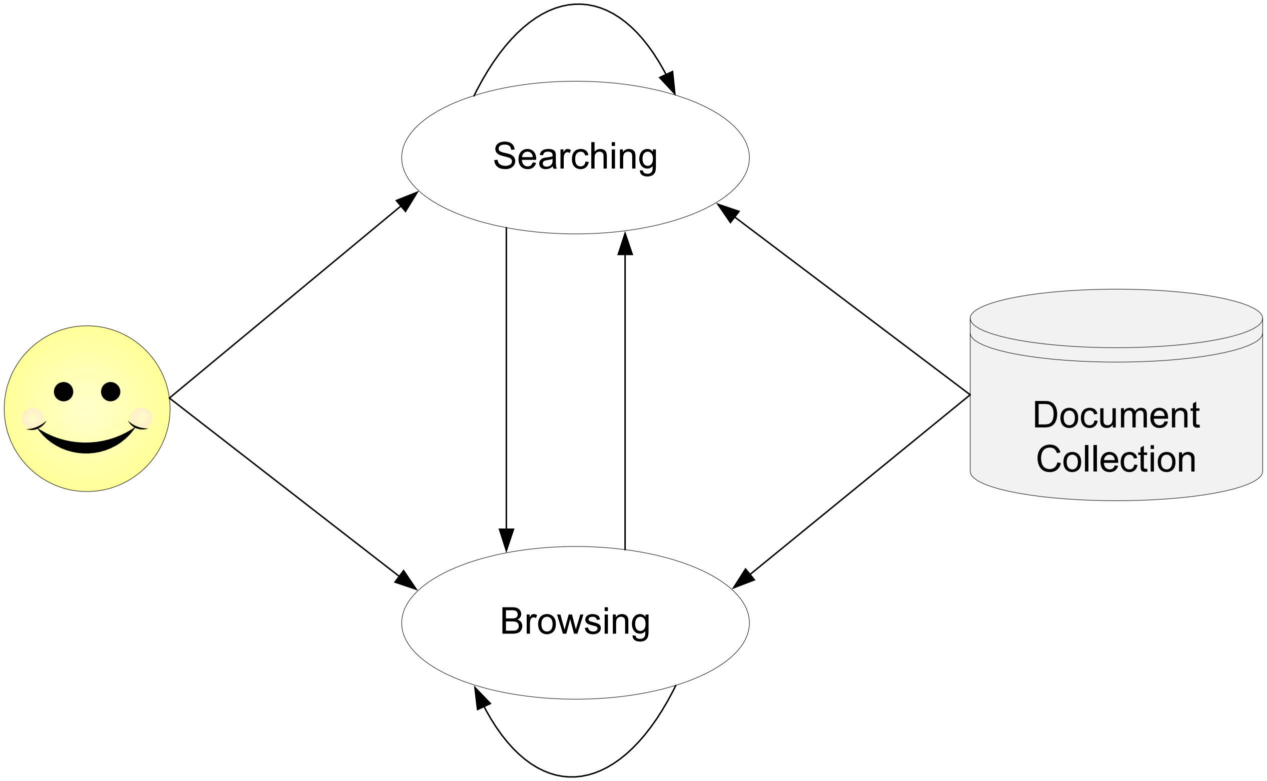
  > *The key goal of an IR system is to retrieve all the items that are relevant to a user query, while retrieving as few nonrelevant items as possible*

- The notion of relevance is of central importance in IR

The User’s Task

- Consider a user who seeks information on a topic of their interest
  - This user first translates their information need into a query, which requires specifying the words that compose the query
  - In this case, we say that the user is *searching* or *querying* for information of their interest
- Consider now a user who has an interest that is either poorly defined or inherently broad
  - For instance, the user has an interest in car racing and wants to browse documents on Formula 1 and Formula Indy
  - In this case, we say that the user is *browsing* or *navigating* the documents of the collection

The User’s Task



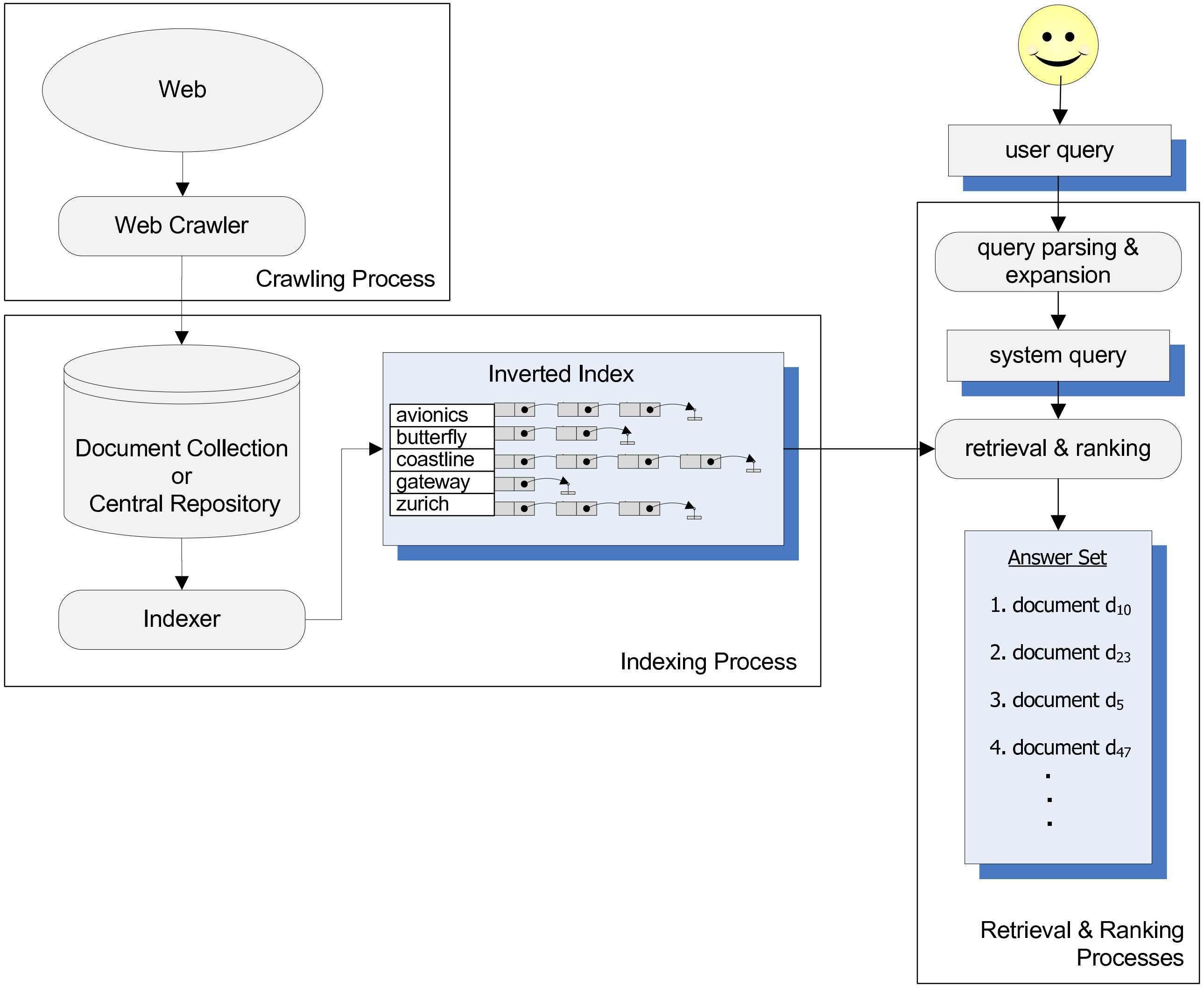
Information × Data Retrieval

- *Data retrieval*: the task of determining which documents of a collection contain the keywords in the user query
- Data retrieval system
  - Ex: relational databases
  - Deals with data that has a well defined structure and semantics
  - A single erroneous object among a thousand retrieved objects means total failure
- Data retrieval does not solve the problem of retrieving information about a subject or topic

The IR System

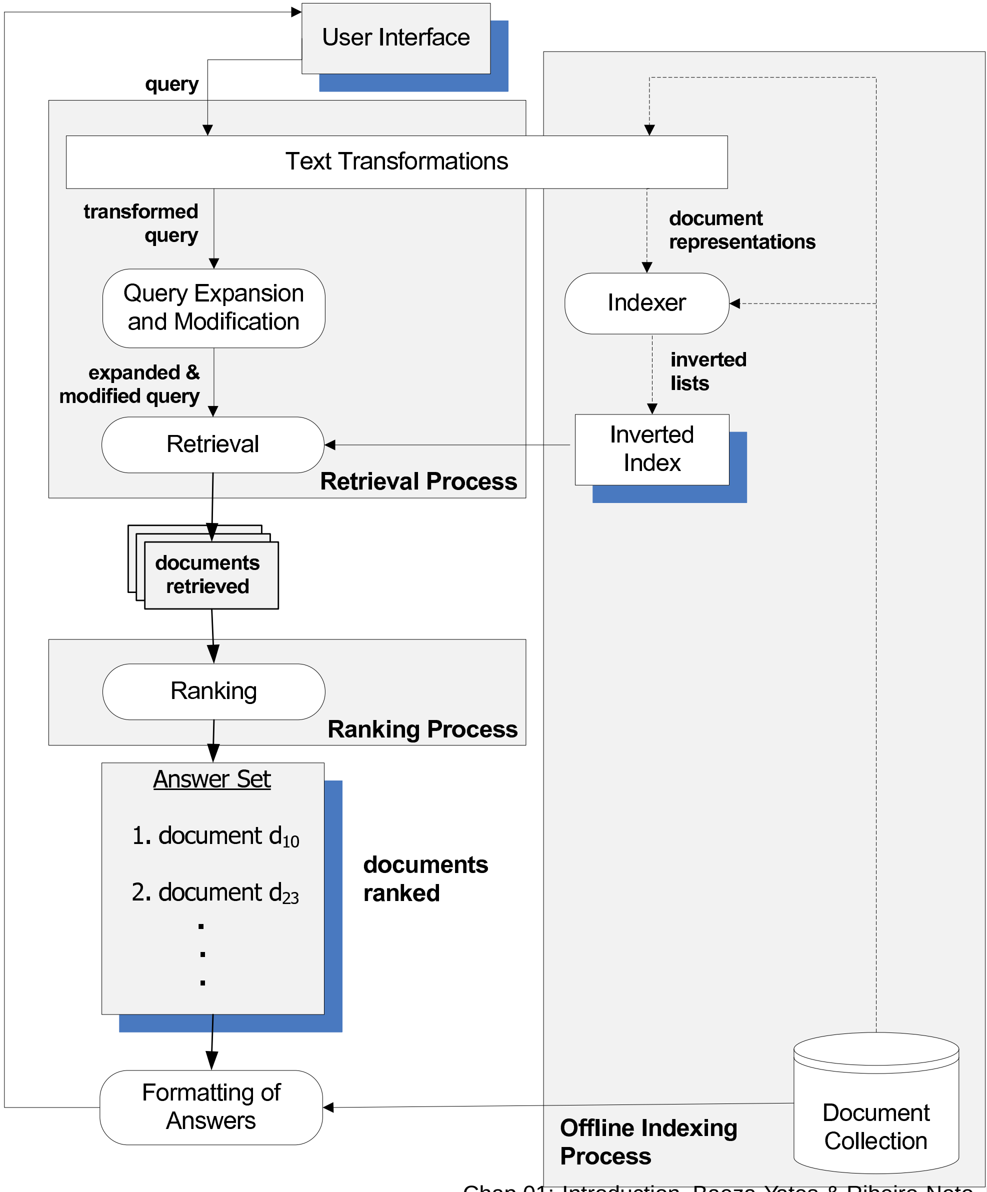
Architecture of the IR System

- High level software architecture of an IR system



Retrieval and Ranking Processes

- The processes of *indexing*, *retrieval*, and *ranking*



The Web

A Brief History

- At the end of World War II, Vannevar Bush looked for applications of new technologies to peace times
- Bush first produced a report entitled *Science, The Endless Frontier*
    - This report directly influenced the creation of the National Science Foundation
- Following, he wrote *As We May Think*, a remarkable paper which discussed new hardware and software gadgets
- In Bush's words:

Whole new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified

A Brief History

- *As We May Think* influenced people like Douglas Engelbart, who invented the computer mouse and introduced the concept of hyperlinked texts
- Ted Nelson, working in his Project Xanadu, pushed the concept further and coined the term *hypertext*
- A hypertext allows the reader to jump from one electronic document to another, which was one important property regarding the problem that Tim Berners-Lee faced in 1989

A Brief History

- At the time, Berners-Lee worked in Geneva at the CERN—*Conseil Européen pour la Recherche Nucléaire*
- There, researchers who wanted to share documentation with others had to reformat their documents to make them compatible with an internal publishing system
- Berners-Lee reasoned that it would be nice if the solution of sharing documents were decentralized
- He saw that a *networked hypertext* would be a good solution and started working on its implementation

A Brief History

- In 1990, Berners-Lee
  - Wrote the *HTTP protocol*
  - Defined the *HTML language*
  - Wrote the first *browser*, which he called *World Wide Web*
  - Wrote the first *Web server*
- In 1991, he made his browser and server software available in the Internet
- The Web was born!

The e-Publishing Era

- Since its inception, the Web became a huge success
    - Well over 20 billion pages are now available and accessible in the Web
    - More than one fourth of humanity now access the Web on a regular basis
- Why is the Web such a success? What is the single most important characteristic of the Web that makes it so revolutionary?
- In search for an answer, let us dwell into the life of a writer who lived at the end of the 18th Century

The e-Publishing Era

- She finished the first draft of her novel in 1796
  - The first attempt of publication was refused without a reading
  - The novel was only published 15 years later!
  - She got a flat fee of $110, which meant that she was not paid anything for the many subsequent editions
  - Further, her authorship was anonymized under the reference “By a Lady”
- We are talking of ...

The e-Publishing Era

- *Pride and Prejudice* is the second or third best loved novel in the UK ever, after *The Lord of the Rings* and *Harry Potter*
- It has been the subject of six TV series and five film versions
  - The last of these, starring Keira Knightley and Matthew Macfadyen, grossed over 100 million dollars
- Jane Austen published anonymously her entire life
- Throughout the 20th century, her novels have never been out of print

The e-Publishing Era

- Jane Austen was discriminated because there was no *freedom to publish* in the beginning of the 19th century
- The Web, unleashed by the inventiveness of Tim Berners-Lee, changed this once and for all
- It did so by universalizing *freedom to publish*

The Web moved mankind into a new era, into a new time, into The e-Publishing Era

How the Web Changed Search

- Web search is today the most prominent application of IR and its techniques—the ranking and indexing components of any search engine are fundamentally IR pieces of technology
- The *first major impact* of the Web on search is related to the characteristics of the document collection itself
  - The Web is composed of pages distributed over millions of sites and connected through hyperlinks
  - This requires collecting all documents and storing copies of them in a central repository, prior to indexing
  - This new phase in the IR process, introduced by the Web, is called *crawling*

How the Web Changed Search

- The *second major impact* of the Web on search is related to:
  - The size of the collection
  - The volume of user queries submitted on a daily basis
  - As a consequence, performance and scalability have become critical characteristics of the IR system
- The *third major impact*: in a very large collection, predicting relevance is much harder than before
  - Fortunately, the Web also includes new sources of evidence
  - Ex: hyperlinks and user clicks in documents in the answer set

How the Web Changed Search

- The *fourth major impact* derives from the fact that the Web is also a medium to do business
  - Search problem has been extended beyond the seeking of text information to also encompass other user needs
  - Ex: the price of a book, the phone number of a hotel, the link for downloading a software
- The *fifth major impact* of the Web on search is Web spam
  - Web spam: abusive availability of commercial information disguised in the form of informational content
  - This difficulty is so large that today we talk of Adversarial Web Retrieval

Practical Issues in the Web

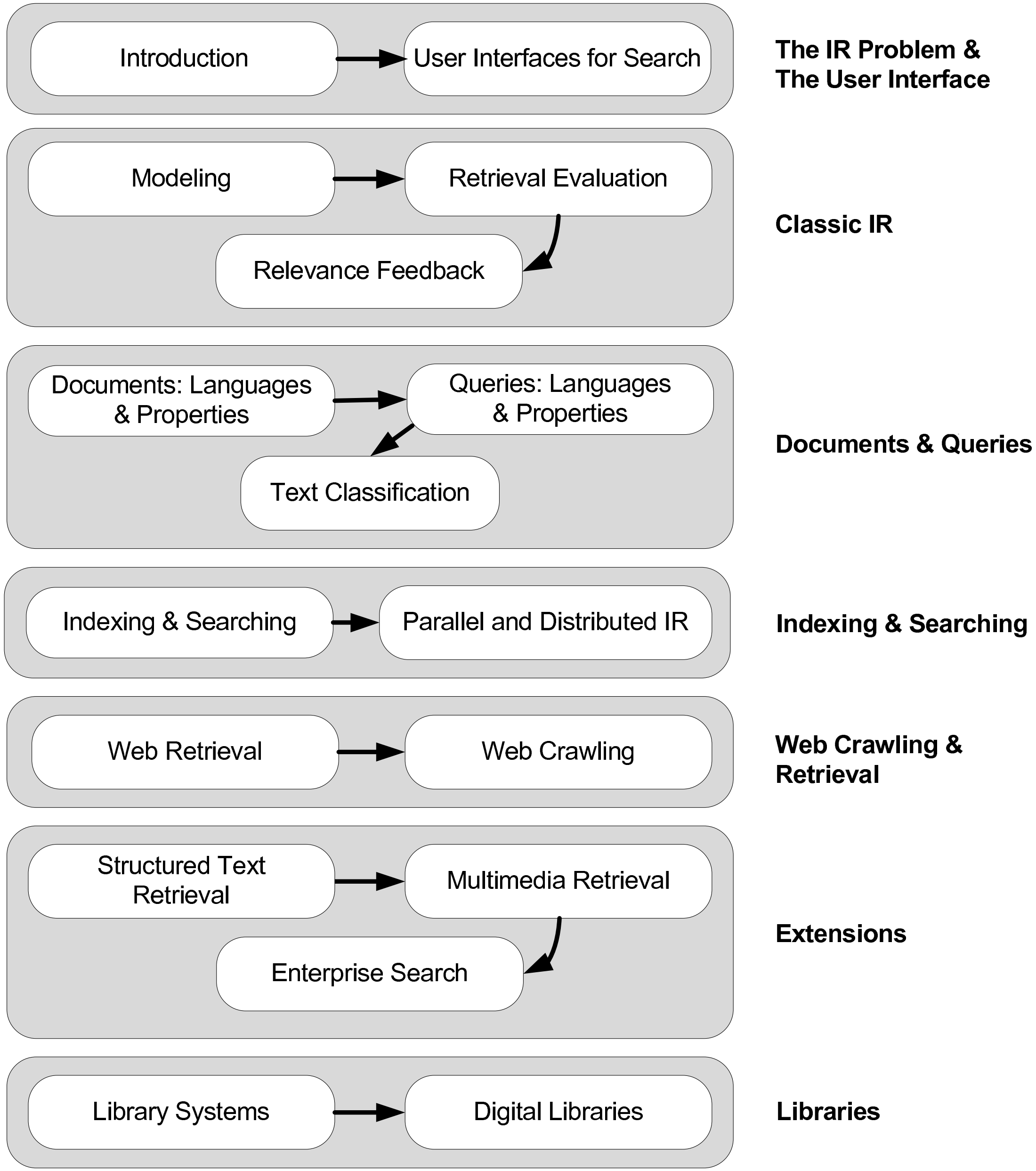
- Security
  - Commercial transations over the Internet are not yet a completely safe procedure
- Privacy
  - Frequently, people are willing to exchange information as long as it does not become public
- Copyright and patent rights
  - It is far from clear how the wide spread of data on the Web affects copyright and patent laws in the various countries
- Scanning, optical character recognition (OCR), and cross-language retrieval

Organization of the Book

Focus of the Book

- The book presents an overall view of research in IR from a computer scientist’s perspective
  - This means that the main focus of the book is on computer algorithms and techniques used in IR systems
- A rather distinct viewpoint is taken by librarians and information science researchers
  - In this viewpoint, the focus is on trying to understand how people interpret and use information
- This human-centered viewpoint is discussed in the user interfaces chapter and in the last two chapters of the book

Book Contents


Introduction
User Interfaces for Search
The IR Problem &
The User Interface
Modeling
Retrieval Evaluation
Relevance Feedback
Classic IR
Documents: Languages
& Properties
Queries: Languages
& Properties
Text Classification
Documents & Queries
Indexing & Searching
Parallel and Distributed IR
Indexing & Searching
Web Retrieval
Web Crawling
Web Crawling &
Retrieval
Structured Text
Retrieval
Multimedia Retrieval
Enterprise Search
Extensions
Library Systems
Digital Libraries
Libraries