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 itens: 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, etc.
- For holding the various items, special purpose buildings called **libraries** are used
- The oldest known library was created in Elba, in the Fertile Crescent, between 3,000 and 2,500 BC
- Nowadays, they are everywhere and constitute the collective memory of the human race

Early Developments

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

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, etc.

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 World Wide Web
- Web is today the largest human repository of knowledge in history
- Finding useful information on the Web is not always a simple task and usually requires running a search
  - And searching on Web is all about IR and its techs
- 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 does not necessarily provide the best formulation for querying an IR system
- Instead, the user might want to first translate this information need into a query
- This translation 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

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

  Retrieving all the itens that are relevant to a user query while avoid retrieving nonrelevant itens

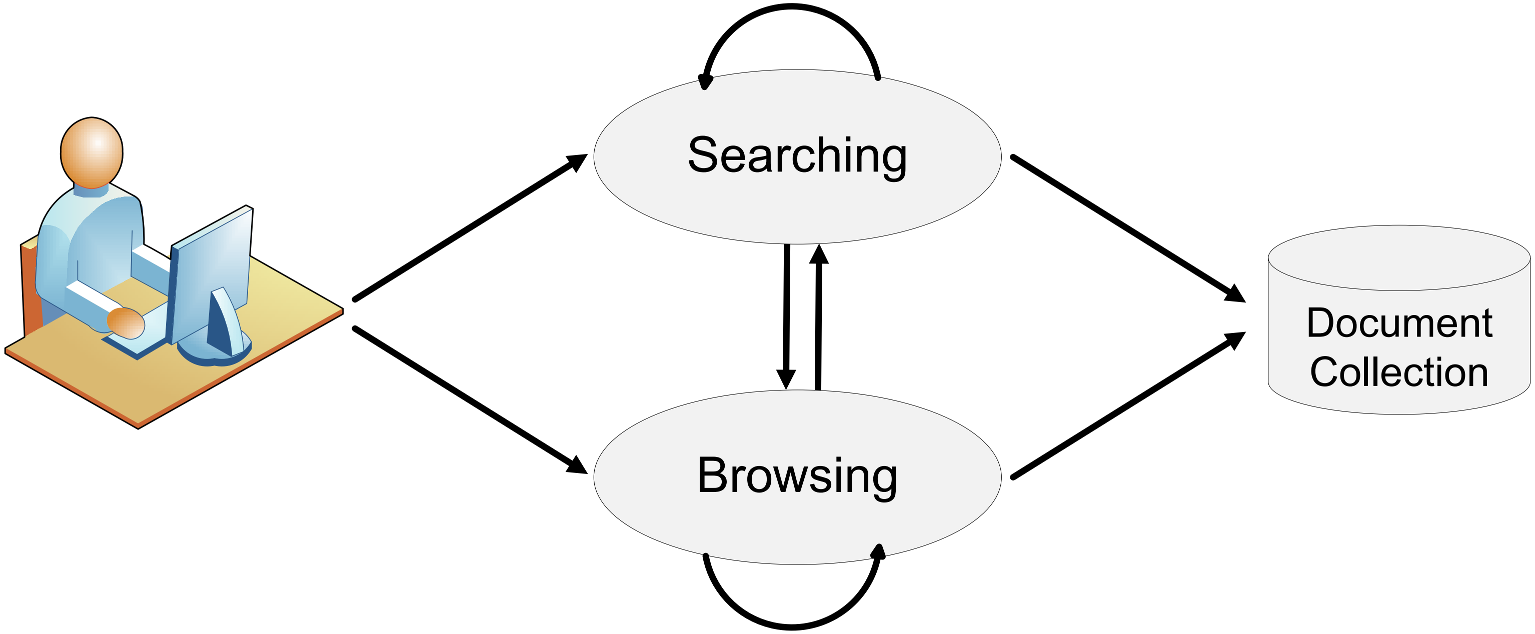
- The notion of relevance is of central importance in IR

The User’s Task

- The user of an IR system has to translate their information need into a query
- This usually implies specifying a set of words that convey the semantics of the information need
    - 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 decides to glance related documents about Formula 1 racing and Formula Indy
    - We say that the user is **browsing** or **navigating** the documents in the collection, not searching

The User’s Task

- The task of the users might be then of two distinct types: **searching** and **browsing**



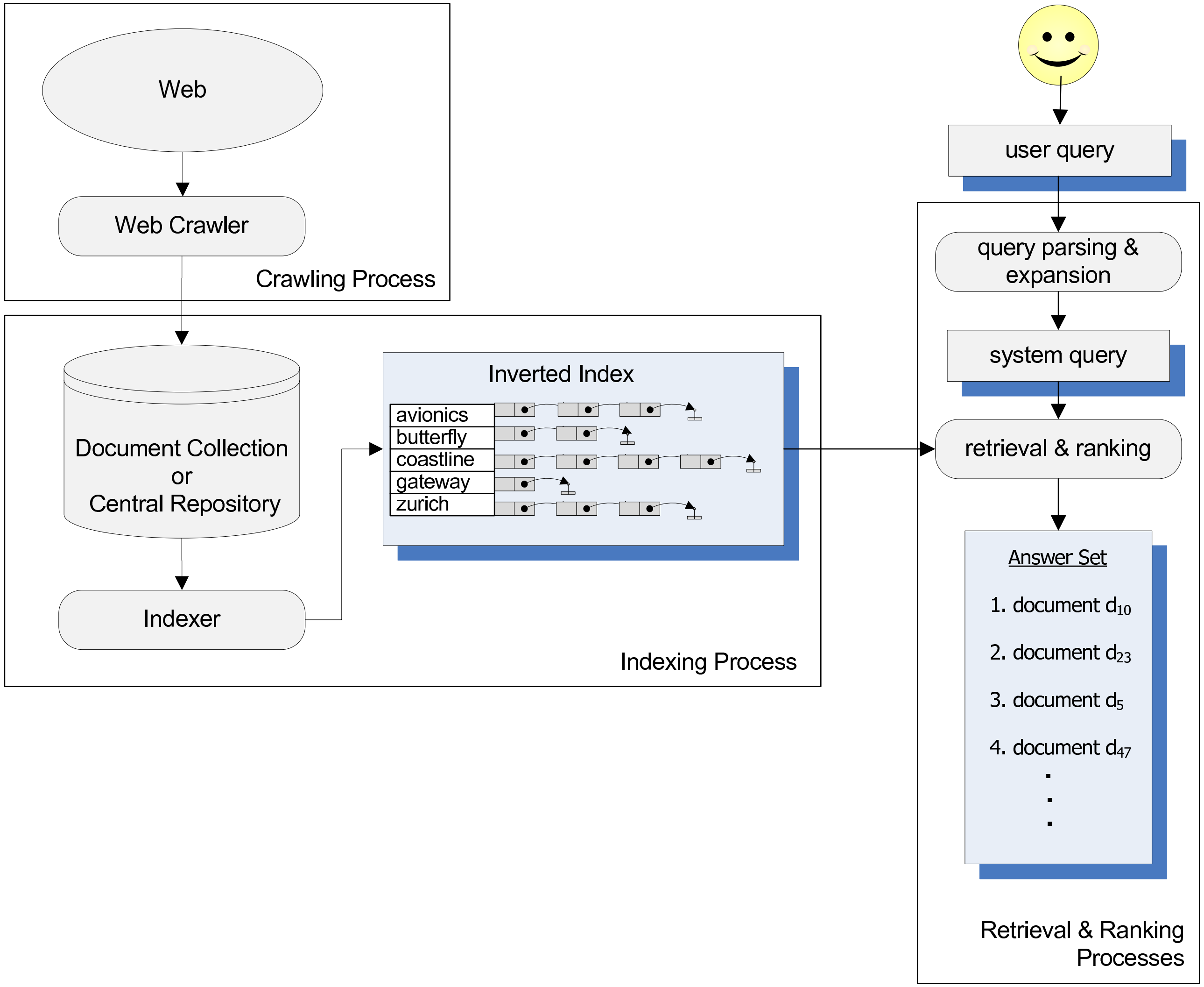
Information × Data Retrieval

- **Data retrieval**: to determe which documents of a collection contain the keywords of the user query
- Data retrieval system
  - Ex: relational databases
  - Deals with data that has a well defined structure and semantic
  - 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

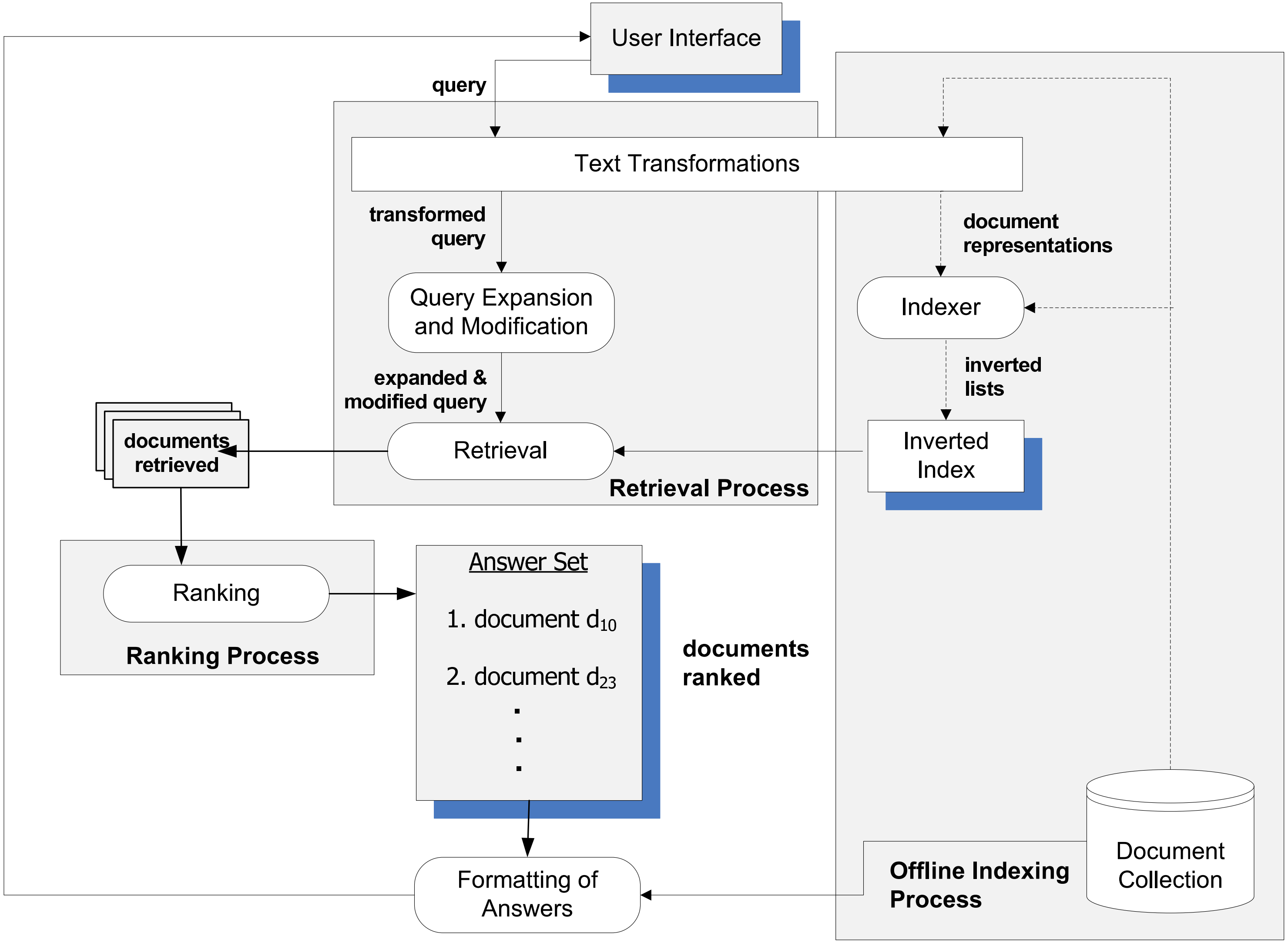
Architecture of the IR System

- 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 the technologies learnt during the war to peace times
- Bush first produced a report entitled *Science, The Endless Frontier*
  - This report directly influenced the creation of the National Science Foundation
- Then, he wrote *As We May Think*, a remarkable paper that 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 introduced the **hypertext** concept
  - The term was coined by Ted Nelson in his Project Xanadu
- 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
  - 20 billion of Web pages
  - 1.7 billion of users
- The advent of the Web changed the world in a way that few people could have anticipated
- The fundamental shift in human relationships, introduced by the Web, was **freedom to publish**
- That is, the freedom to publish that marks the birth of a new era, we refer to as **The e-Publishing Era**

How the Web Changed Search

- Web search is today the most prominent application of IR and its techniques
- 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
- 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** derive 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 the 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 is not a completely safe procedure yet
- 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

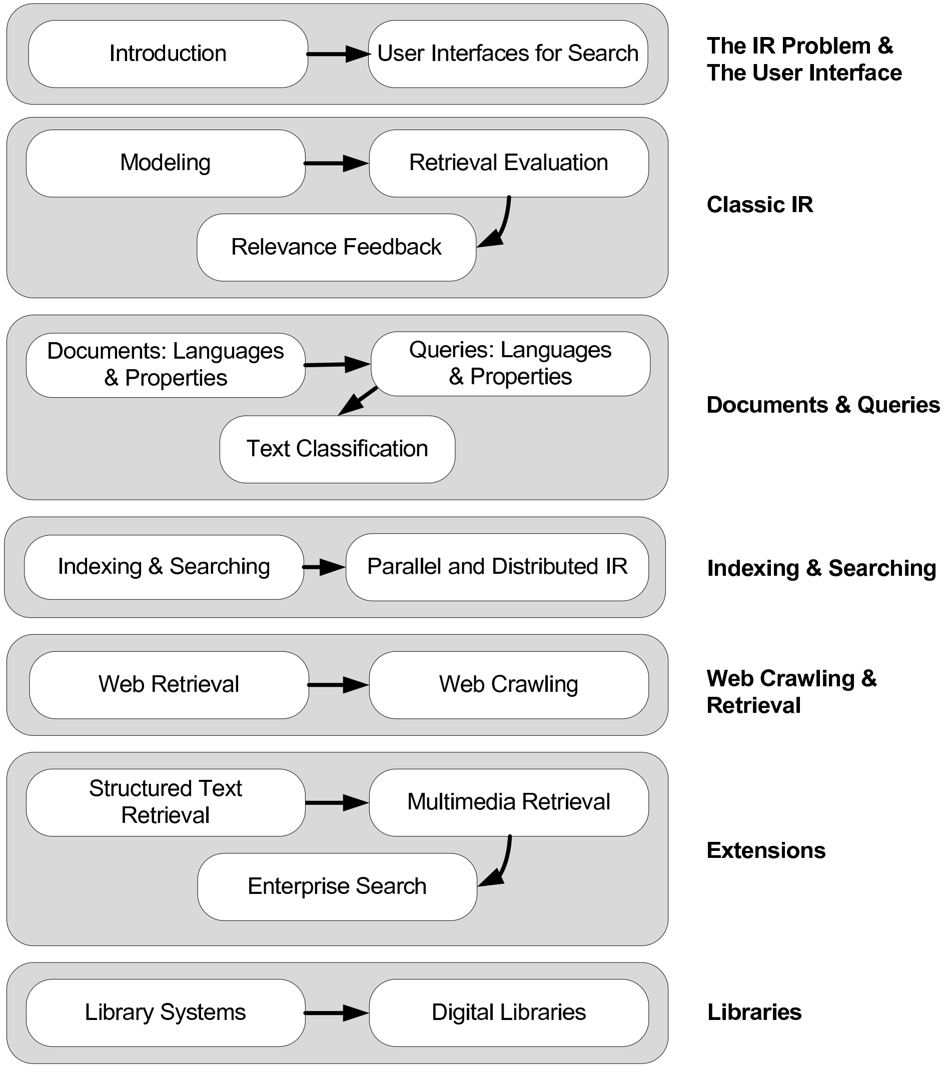
Organization of the Book

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

- Organization of the chapters of the book


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