

# Modern Information Retrieval

## Introduction<sup>1</sup>

Hamid Beigy

Sharif University of Technology

September 25, 2022



---

<sup>1</sup>Some slides have been adapted from slides of Manning, Yannakoudakis, and Schütze.



1. Course Information
2. Introduction
3. Course overview

## Course Information

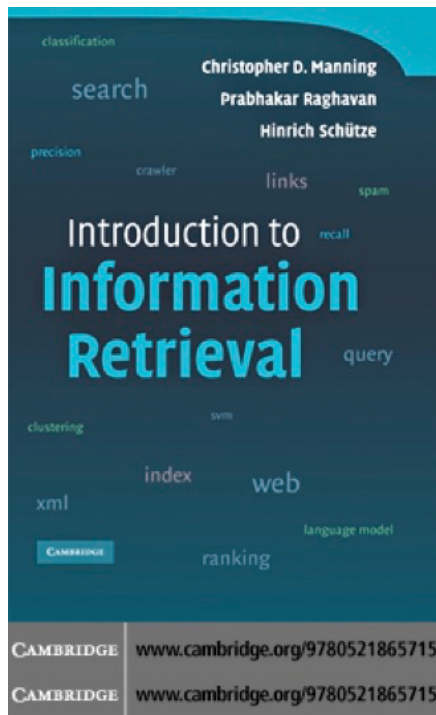
---








1. Course name : **Modern Information Retrieval**
2. Instructor : Hamid Beigy  
Email : [beigy@sharif.edu](mailto:beigy@sharif.edu)
3. Class : **CE 102**
4. Virtual class link: <https://vc.sharif.edu/ch/beigy>
5. Course Website:  
<http://ce.sharif.edu/courses/01-02/1/ce324-1/>  
<http://sharif.edu/~beigy/14011-40324.html>
6. Lectures: **Sat-Mon (9:00-10:30)**
7. TAs :  
Hossein Mirzaiee    Email: [mirhosseinsadegh@yahoo.com](mailto:mirhosseinsadegh@yahoo.com)



► Evaluation:		
Mid-term exam	30%	<a href="#">1401/09/02</a>
Final exam	30%	
Practical Assignments	30%	
Quiz	10%	





-  Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (2011). *Modern Information Retrieval*. 2nd. USA: Addison-Wesley Publishing Company. ISBN: 9780321416919.
-  Kowalski, Gerald (2010). *Information Retrieval Architecture and Algorithms*. 1st. Berlin, Heidelberg: Springer-Verlag. ISBN: 1441977155, 9781441977151.
-  Li, Hang (2011). *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers.
-  Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.
-  Mitra, Bhaskar and Nick Craswell (2018). "An Introduction to Neural Information Retrieval". In: *Foundations and Trends in Information Retrieval*: 13.1, pp. 1–126.

## Introduction

---





1. We define the information retrieval as

## Definition (Information retrieval )

Information retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).

2. **Document Collection**: units we have built an IR system over. Documents can be
  - ▶ memos
  - ▶ book chapters paragraphs
  - ▶ scenes of a movie
  - ▶ turns in a conversation...
3. These days we frequently think first of **web search**, but there are many other cases:
  - ▶ E-mail search
  - ▶ Searching your laptop
  - ▶ Corporate knowledge bases
  - ▶ Legal information retrieval



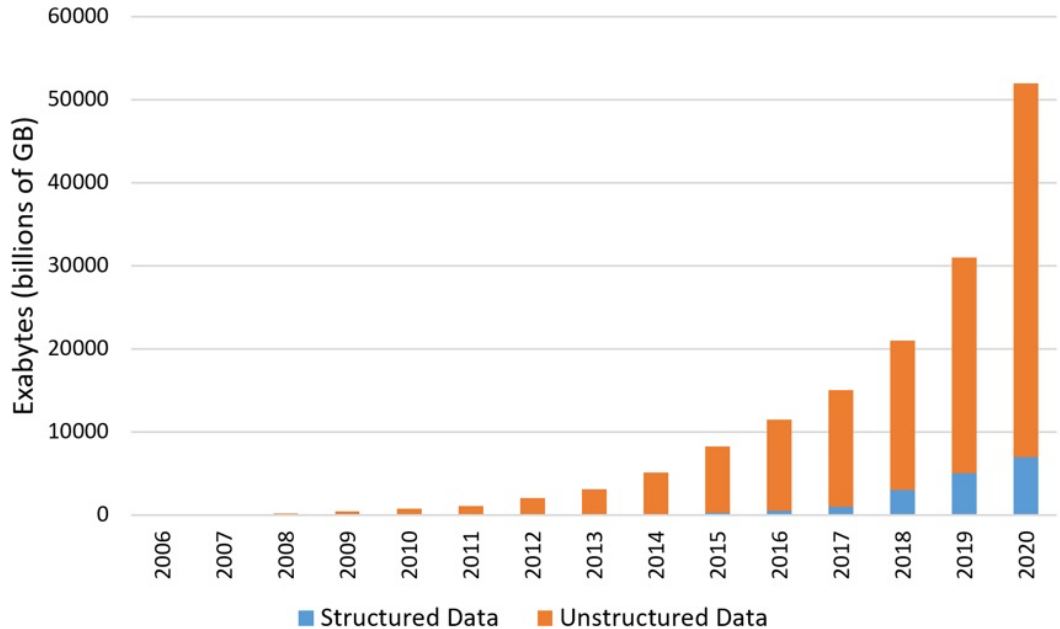
1. Unstructured data means that a formal, semantically overt, easy-for-computer structure is missing.
2. In contrast to the rigidly structured data used in DB style searching (e.g. product inventories, personnel records)  

```
SELECT * FROM BUSINESS-CATALOGUE WHERE CATEGORY = "FLORIST" AND CITY-ZIP = "CB1"
```
3. This does not mean that there is no structure in the data
  - ▶ Document structure (headings, paragraphs, lists. . . )
  - ▶ Explicit markup formatting (e.g. in HTML, XML. . . )
  - ▶ Linguistic structure (latent, hidden)



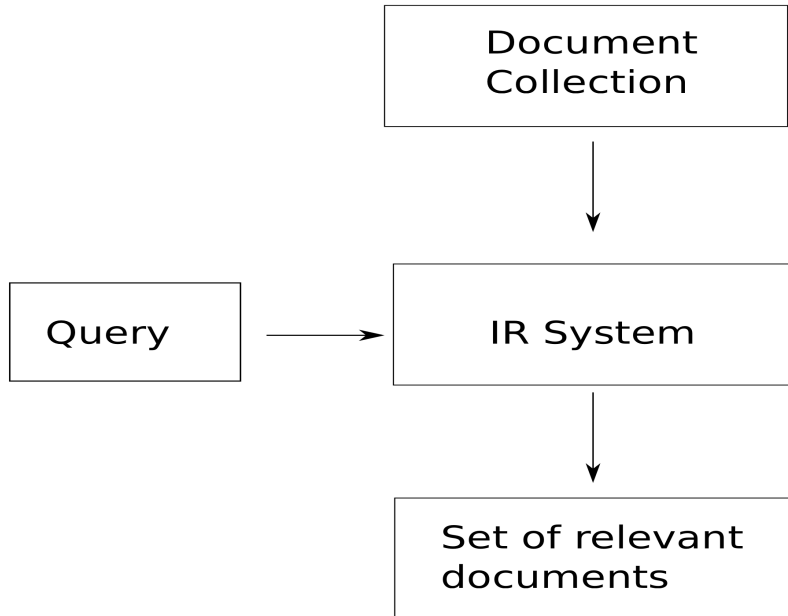
1. Information retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).
2. An **information need** is the topic about which the user desires to know more about.
3. A **query** is what the user conveys to the computer in an attempt to communicate the information need.
4. Types of information needs
  - ▶ Known-item search
  - ▶ Precise information seeking search
  - ▶ Open-ended search (“topical search”)

# Structured vs Unstructured data growth



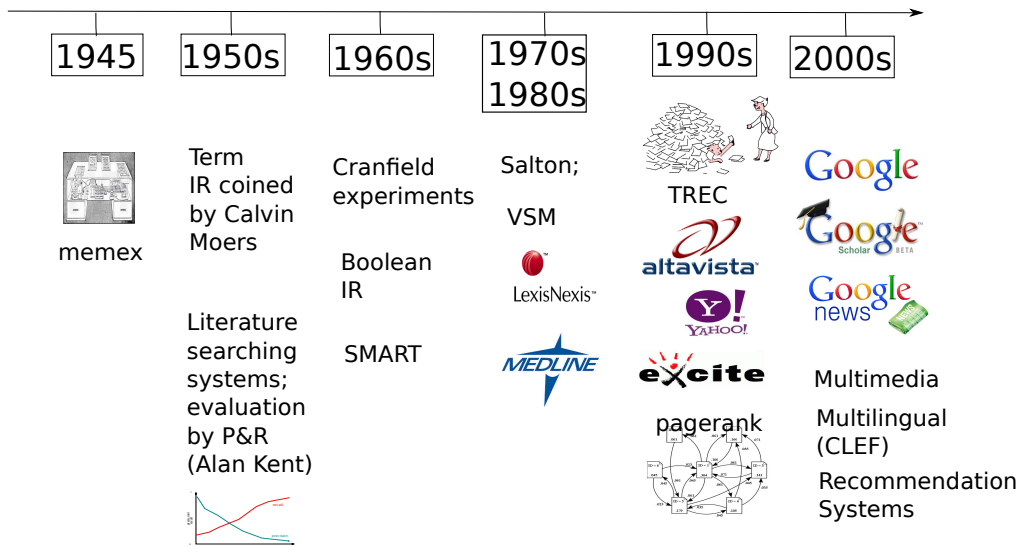


1. A document is **relevant** if the user perceives that it contains information of value with respect to their personal information need.
2. Are the retrieved documents
  - ▶ about the target subject ?
  - ▶ up-to-date?
  - ▶ from a trusted source?
  - ▶ satisfying the user's needs?
3. How should we rank documents in terms of these factors?





- ▶ The **effectiveness** of an IR system (i.e., the quality of its search results) is determined by two key statistics about the system's returned results for a query:
  - ▶ **Precision**: What fraction of the returned results are relevant to the information need?
  - ▶ **Recall**: What fraction of the relevant documents in the collection were returned by the system?
  - ▶ What is the best balance between the two?
    - ▶ Easy to get **perfect recall**: **just retrieve everything**
    - ▶ Easy to get **good precision**: **retrieve only the most relevant**







## 1960-1970 <sup>2</sup>

- ▶ Initial exploration of text retrieval systems for "small" corpora of scientific abstracts, and law and business documents.
- ▶ Development of the basic Boolean and vector-space models of retrieval.
- ▶ Prof. Salton and his students at Cornell University are the leading researchers in the area

## 1970-1980

- ▶ Large document database systems, many run by companies ([Lexis-Nexis](#) and [Dialog](#) and [MEDLINE](#))

## 1980-1990

- ▶ Searching FTPable documents on the Internet ([Archie](#) and [WAIS](#))
- ▶ Searching the World Wide Web ([Lycos](#) and [Yahoo](#) and [Altavista](#))

## 1990-2000

- ▶ Searching FTPable documents on the Internet ([Archie](#) and [WAIS](#))
- ▶ Searching the World Wide Web ([Lycos](#) and [Yahoo](#) and [Altavista](#))
- ▶ Organized Competitions ([NIST](#) and [TREC](#))
- ▶ Searching the World Wide Web ([Ringo](#) and [Amazon](#) and [NetPerceptions](#))



- ▶ Automated Text Categorization & Clustering

## 2000-2010

- ▶ Link analysis for Web Search ([Google](#))
- ▶ Parallel Processing ([Map-Reduce](#))
- ▶ Question Answering ([TREC Q/A track](#))
- ▶ Multimedia IR ([Image](#) and [Video](#) and [Audio and music](#))
- ▶ Cross-Language IR
- ▶ Document Summarization

## 2010-2020

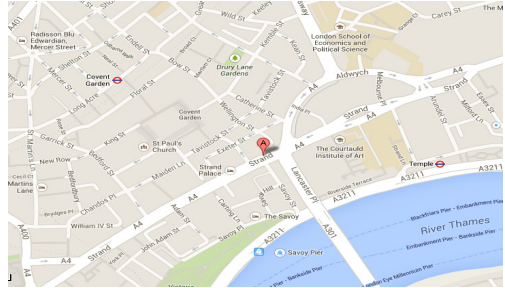
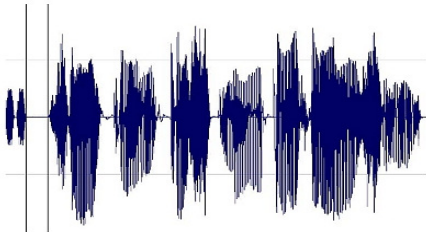
- ▶ Intelligent Personal Assistants ([Siri](#), [Cortana](#), [Google](#), and [Alexa](#))
- ▶ Complex Question Answering ([IBM Watson](#))
- ▶ Distributional Semantics
- ▶ Deep Learning

## 2020-\*\*\*\*

- ▶ By 2025, the researchers believes that we have [rich multi-sensorial experiences that will be capable of producing hallucinations which blend or alter perceived reality.](#)

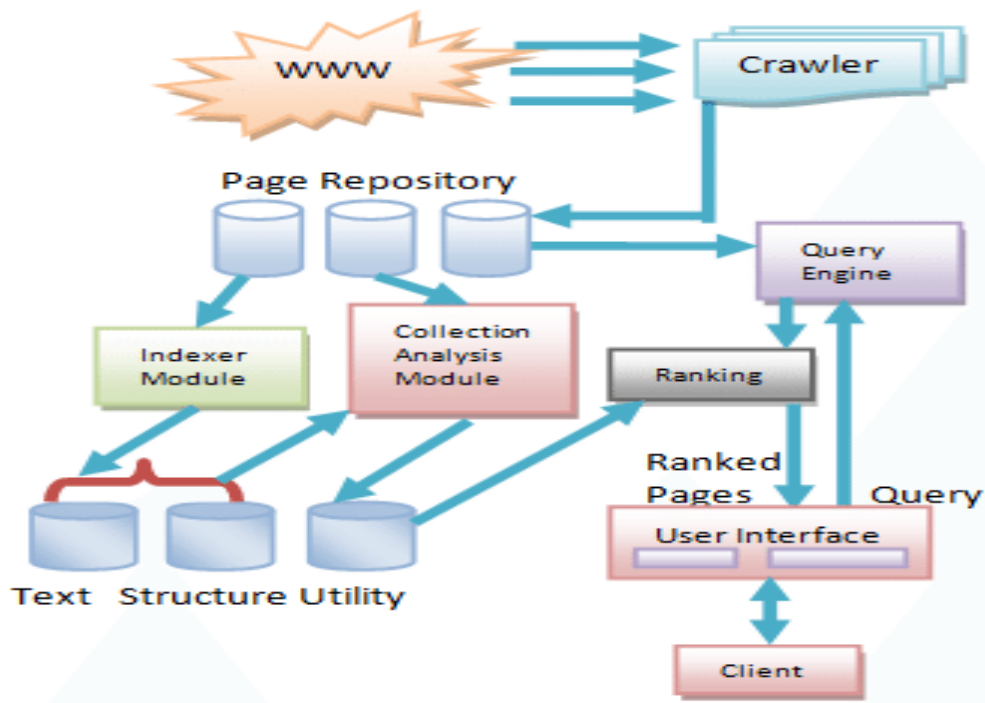
---

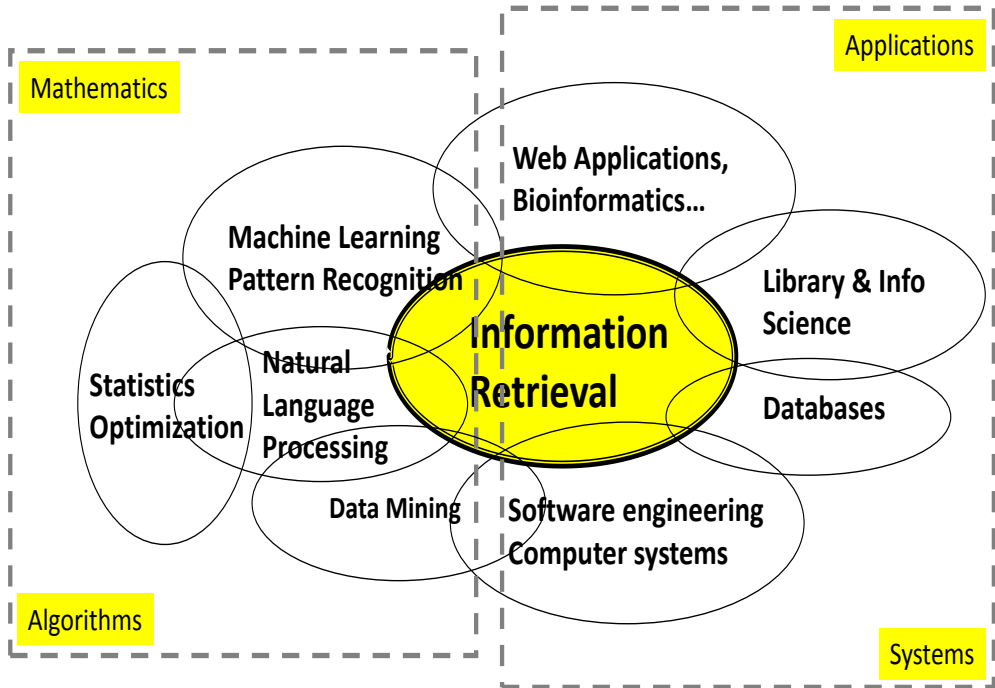
<sup>2</sup>This slide is taken from Prof. Sampath Jayarathna slides.





- ▶ Which plays of Shakespeare contain the words **BRUTUS AND CAESAR**, but **NOT CALPURNIA**?
- ▶ One could grep all of Shakespeare's plays for **BRUTUS** and **CAESAR**, then strip out lines containing **CALPURNIA**.
- ▶ **Why is grep not the solution?**
  - ▶ Slow (for large collections)
  - ▶ grep is line-oriented, IR is document-oriented
  - ▶ "NOT CALPURNIA" is non-trivial
  - ▶ Other operations (e.g., find the word **ROMANS** near **COUNTRYMAN**) not feasible





## Course overview

---



- ▶ Introduction
- ▶ Indexing and text operations
- ▶ IR models ( Boolean, vector space, probabilistic)
- ▶ Evaluation of IR systems
- ▶ Query operations
- ▶ Language models
- ▶ Machine Learning in IR (classification, clustering, and learning to rank)
- ▶ Dimensionality reduction and word embedding
- ▶ Web information retrieval and search engines
- ▶ Some advanced topics
  - ▶ Recommender systems
  - ▶ Personalized IR
  - ▶ Sentiment Analysis
  - ▶ Cross-lingual IR
  - ▶ QA systems
  - ▶ Neural information retrieval



