Modern Information Retrieval

Introduction

Hamid Beigy

Sharif University of Technology

1403-11-20

Table of contents

1. Course Information

2. Introduction

3. Course overview

4. References

Course Information

Course Information

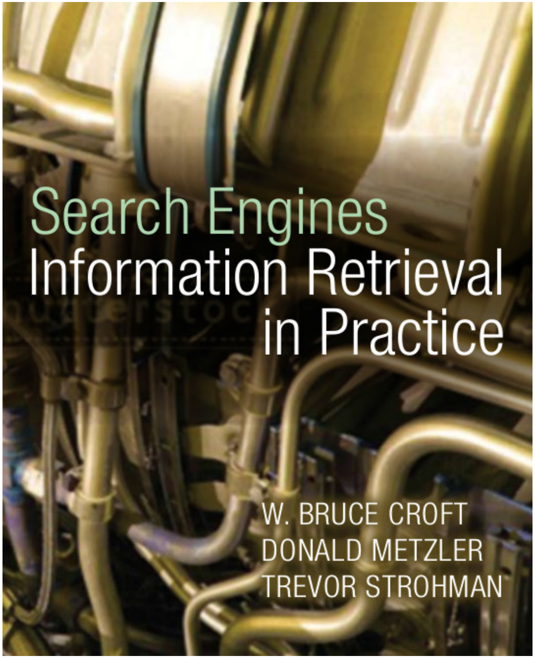
1. Course name : **Modern Information Retrieval**
2. Instructor : Hamid Beigy Email : `beigy@sharif.edu`
3. Class : CE 201
4. Virtual class link: `https://vc.sharif.edu/beigy`
5. Course Website: `http://sharif.edu/~beigy/14032-40324.html`
6. Lectures: Sat-Mon (9:00-10:30)
7. Teaching Assistant : Reza Tavakoli Email: `seyedreza.shiyade@gmail.com`

Course evaluation

- Evaluation:

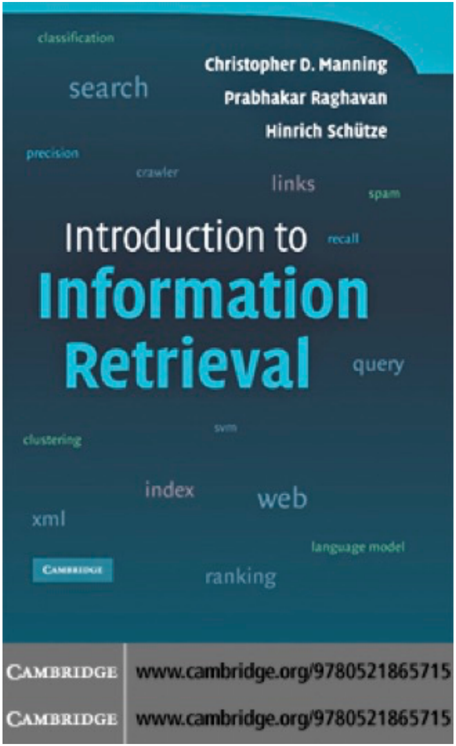
Mid-term exam	25%	1404-01-30
Final exam	30%	
Practical Assignments	35%	
Quiz	15%	

Main References



Search Engines
Information Retrieval
in Practice

W. BRUCE CROFT
DONALD METZLER
TREVOR STROHMAN



Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze

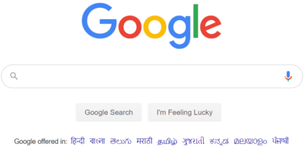
Introduction to
Information
Retrieval

References

- Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (2011). *Modern Information Retrieval*. 2nd. USA: Addison-Wesley Publishing Company. ISBN: 9780321416919.
- Croft, W. Bruce, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.
- 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.

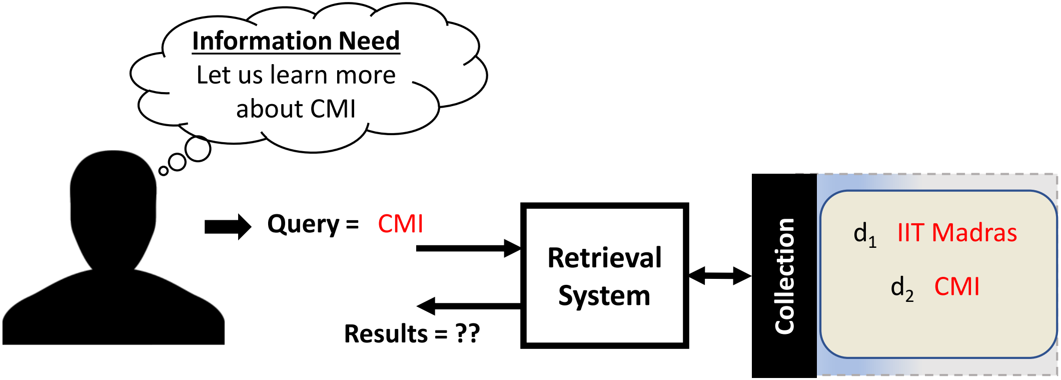
Introduction

Introduction


Google
Google Search
I'm Feeling Lucky
Google offered in: हिन्दी বাংলা తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ

Life without search engines is difficult to imagine!

What is Information Retrieval?



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).

Role of Information

1. If you know
   - Which stock to invest in?
   - Which faculty to work with?
   - How to get into a top college?
   - Which course to register for?
   - What to study?
   - How to prepare for job interviews?
2. **If only you had the information, you could rule this world.**
3. **What happens when you lose access to all your information?**

Document collection

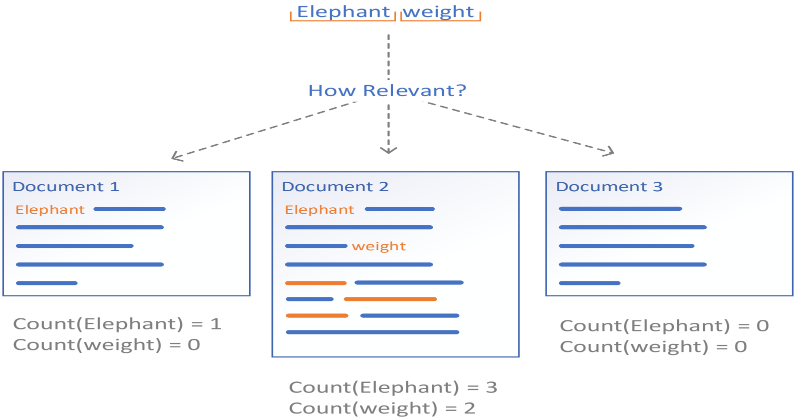
1. Document Collection: units we have built an IR system over such as
    - memos
    - book chapters paragraphs
    - scenes of a movie
    - turns in a conversation...
2. Some applications
    - E-mail search
    - Searching your laptop
    - Corporate knowledge bases
    - Legal information retrieval
3. An information need is the topic about which the user desires to know more about.
4. A query is what the user conveys to the computer in an attempt to communicate the information need.

Relevance

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?

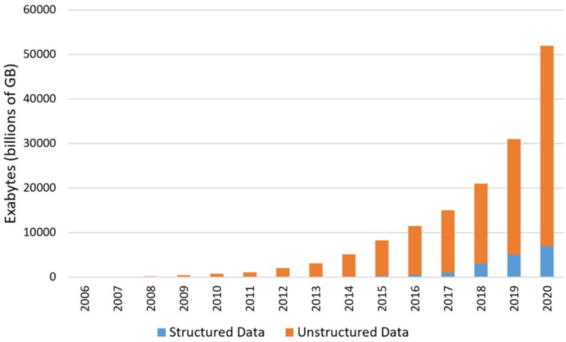
How well has the system performed?

1. The effectiveness of an IR system is determined by two key statistics:
   - Precision: What fraction of returned results are relevant to information need?
   - Recall: What fraction of relevant documents in collection were returned by 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



Structured vs Unstructured Data

1. Unstructured data: a formal, semantically overt, easy-for-computer structure is missing.

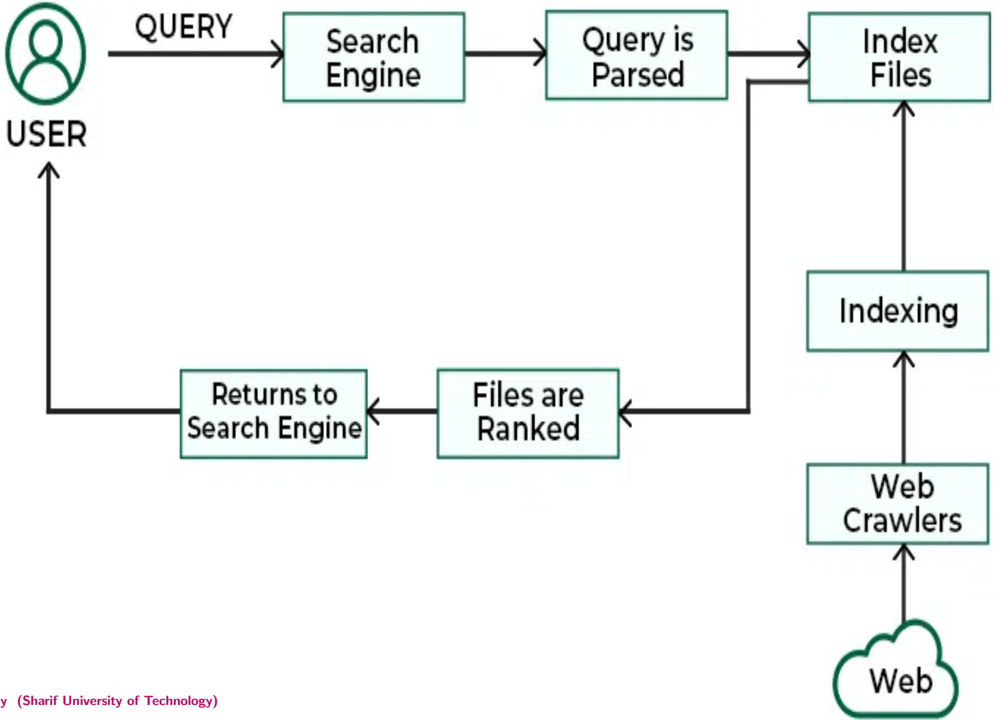


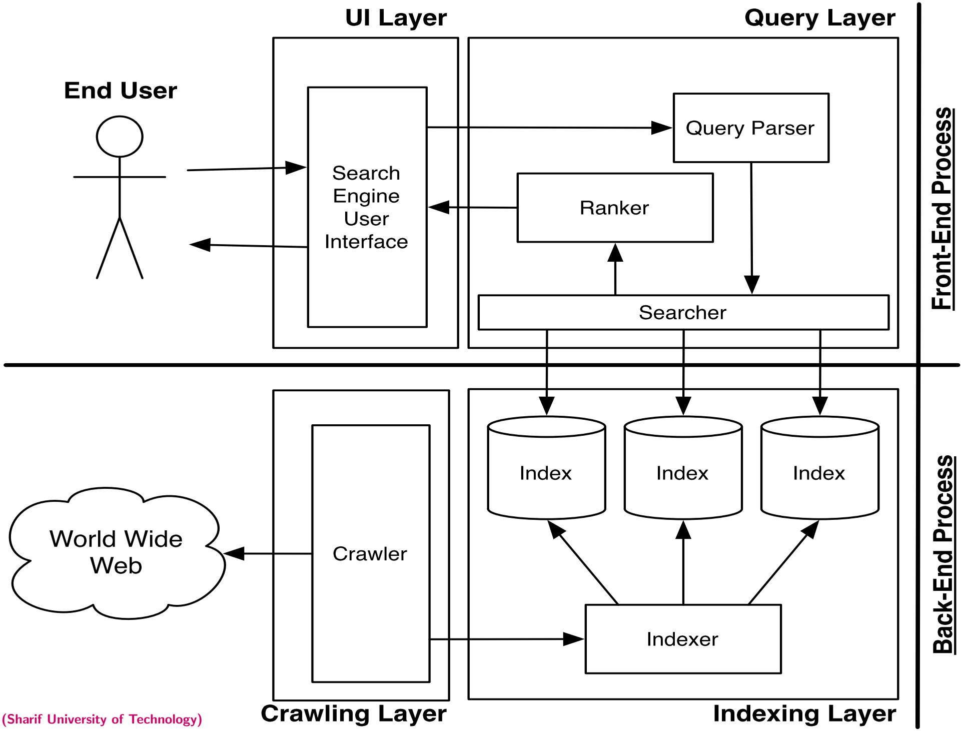
2. Structured data used in DB style searching,

   SELECT * FROM CATALOGUE WHERE CATEGORY = "FLORIST" AND 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)

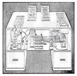
Search engine architecture




UI Layer
Query Layer
End User
Search Engine User Interface
Query Parser
Ranker
Searcher
Front-End Process
Index
Index
Index
World Wide Web
Crawler
Indexer
Back-End Process
Crawling Layer
Indexing Layer

A short history of IR

1945

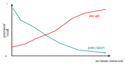


memex

1950s

Term IR coined by Calvin Moers

Literature searching systems; evaluation by P&R (Alan Kent)



1960s

Cranfield experiments

Boolean IR

SMART

1970s 1980s

Salton;

VSM

LexisNexis™

MEDLINE

1990s

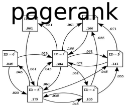
TREC

altavista

YAHOO!

excite

pagerank



2000s

Google

Google Scholar BETA

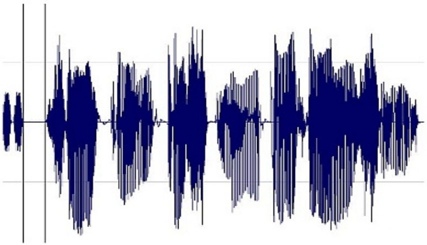
Google news

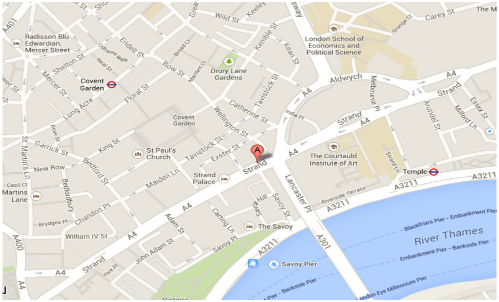
Multimedia

Multilingual (CLEF)

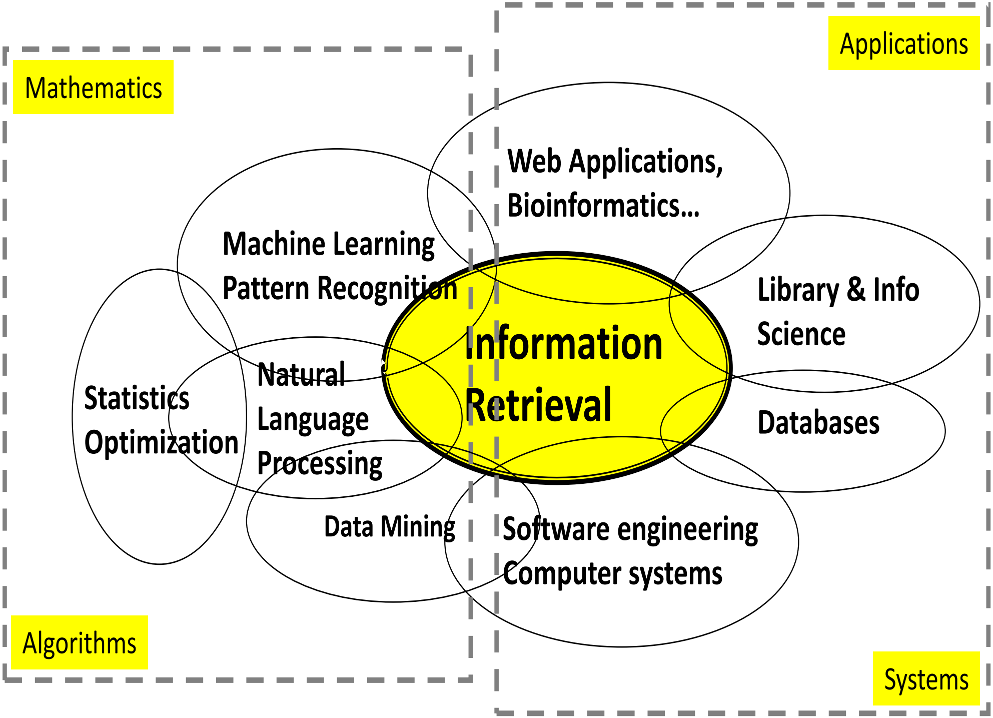
Recommendation Systems

IR for non-textual media





Related areas



Course overview

Course overview

1. Introduction
2. Text processing
3. Ranking with indexes
4. Retrieval models
5. Evaluation of IR systems
6. Machine Learning in IR (classification, clustering, and learning to rank)
7. Web information retrieval and search engines
8. Neural information retrieval
9. Applications
    - Recommender systems
    - Personalized IR
    - Sentiment Analysis
    - Corss-lingual IR
    - QA systems

References

Reading

1. Chapter 1 of Information Retrieval[1]
2. Chapter 1 of Search Engines - Information Retrieval in Practice.[2]

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

[2] W. Bruce Croft, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.

Questions?