Information Retrieval

Matthias Hagen Martin Potthast Benno Stein

Contents

Learning Objectives

- Understand a retrieval system's components and explain their interactions
- Understand and compare text preprocessing options
- Understand and compare options for retrieval system indexes
- Acquire a solid background in the theory of retrieval models
- Understand and apply the principles and practice of retrieval evaluation
- Implement and evaluate retrieval systems in practice

Related Fields

1. Statistics [paradigms, models]
2. Mathematics

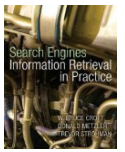
3. Data Mining [methods, algorithms]
4. Machine Learning
5. Natural Language Processing
6. Knowledge Processing

7. Search Engines [applications]
8. Recommender Systems
9. Decision Support Systems

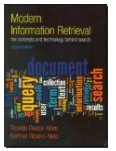
Literature

Information Retrieval:

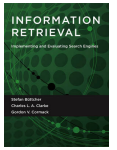
- W.B. Croft, D. Metzler, T. Strohman.
  Search Engines: Information Retrieval in Practice
  Pearson 2009. ciir.cs.umass.edu/downloads/SEIRiP.pdf



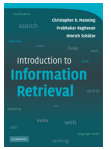
- R. Baeza-Yates, B. Ribeiro-Neto.
  Modern Information Retrieval: The Concepts and Technology behind Search
  Pearson 2011.



- S. Büttcher, C.L.A. Clarke, G.V. Cormack.
  Information Retrieval: Implementing and Evaluating Search Engines
  MIT Press 2010.

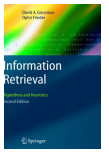


- C.D. Manning, P. Raghavan, H. Schütze.
  Introduction to Information Retrieval
  Cambridge University Press 2008. nlp.stanford.edu/ir-book/



- P. Ingwersen, K. Järvelin.
  The Turn: Integration of Information Seeking and Retrieval in Context
  Springer 2005.

- D.A. Grossman, O. Frieder.
  Information Retrieval: Algorithms and Heuristics
  Springer 2004.



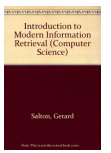
Literature

Information Retrieval: (continued)

- R. Ferber.
  Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web
  dpunkt 2003. information-retrieval.de/irb/ir.html

- I.H. Witten, A. Moffat, T.C. Bell.
  Managing Gigabytes: Compressing and Indexing Documents and Images
  Morgan Kaufmann 1999.

- G. Salton, M.J. McGill.
  Introduction to Modern Information Retrieval
  McGraw-Hill 1983. sigir.org/resources/museum/



- C.J. van Rijsbergen.
  Information Retrieval
  Butterworths 1979. www.dcs.gla.ac.uk/keith/preface.html

Data Mining:

- S. Chakrabarti.
  Mining the Web: Discovering Knowledge from Hypertext Data
  Morgan Kaufmann 2003.

Remarks:

- Peer-reviewed research in information retrieval is published at a number of “core” conferences, journals, and monograph series. The Information Retrieval Anthology provides an overview.
- Probably most important IR conferences: SIGIR, WSDM, WWW, CIKM, ECIR, ICTIR, SIGIR-AP + shared tasks at TREC, CLEF, NTCIR, FIRE
- Important conferences from related fields:
  - Natural language processing: ACL, EMNLP, NAACL, COLING, CoNLL, AACL, EACL
  - Digital libraries: JCDL, TPDL
  - Information science: ASIST, iConference
  - Human-computer interaction: CHI, CUI, IUI, UMAP
  - Knowledge discovery and data mining: KDD, SDM, ICDM, PAKDD
  - Machine learning: NeurIPS, ICML, ECML
  - Artificial Intelligence: AAAI, IJCAI, ECAI
  - Web science: WebScience, Hypertext
  - Social media: ICWSM, ASONAM

Remarks: (continued)

- Informational retrieval research and development is based on open source software projects.
- Retrieval libraries (industry): Elasticsearch, Lucene, Solr, Sphinx, Vespa, Xapian
- Retrieval libraries (research):
  - Terrier, PyTerrier [Macdonald et al., 2012, 2020]
  - Anserini [Yang et al., 2017], Pyserini [Lin et al., 2021]
  - Capreolus [Yates et al., 2020]
  - PyGaggle [Pradeep et al., 2023]
  - Sentence Transformer [Reimers and Gurevych, 2019]
  - Matchmaker, MatchZoo, OpenMatch, OpenNIR, Tevatron, Zettair, ...
  - PISA [Mallia et al., 2019]
  - ir_axioms [Bondarenko et al., 2022]
  - Lemur (Indri, Lucindri, Galago)
- Search engines (research): ChatNoir Search [Bevendorff et al., 2018]

Remarks: (continued)

- Data sources:
  - Common Crawl
  - OpenWebSearch.eu [Granitzer er al., 2024]
  - ir_datasets [MacAveney et al., 2021]
- Evaluation tools:
  - trec_eval
  - TrecTools [Palotti et al., 2019]
  - ir-measures [MacAvaney et al., 2022]
  - TIREx [Fröbe et al., 2023]