Large Scale Search Engines
Credit: 3-0-0

Course Description

In order to find relevant information timely from very large amount of semi-structured and unstructured data, over the years the search technologies have undergone considerable amount of changes – both, architectural as well as algorithmic. The main goal of this course is to teach the state of the art methods and designs underlying the modern search engines that handle very large amount of data.

Course Objectives

The broad objective of the course is to teach large scale search engines architecture, the approaches for storing large volume of data and the efficient algorithms for query processing. Specifically, the course will cover the following major topics.

- File systems for big data
- Distributed in-memory index construction, maintenance and partitioning strategies
- Shard-based search in scale-up and scale-out platforms
- Efficient multi-stage learning to rank architecture
- Specialized search: Personal big data, Search from enterprise data, Federated search
- Evaluation: Evaluation-as-a-service model, cost effective relevance estimation from very large collection

Course Content

- Challenges in large-scale IR: Infrastructural challenges, algorithmic challenges (2)
- Search System Infrastructure: Google file system, Map-reduce, Big Table (7)
- Distributed indexing: Index construction, maintenance and partitioning (6)
- Scalable index compression: PForDelta, Compression-decompression with SIMD (2)
- Shard-based Search: Scale-up architecture, Scale-out architecture, Selective Search (6)
- Learning to rank with multi-stage search architecture (3)
- Real-time search: Indexing and query processing (3)
- Federated, aggregated, vertical and enterprise search (3)
- Lifelogging: Searching from personal big data (2)
- Evaluation: Evaluation-as-a-service model, sampling-based methods for relevance density estimation (3)

Books

1. Data-Intensive Text Processing with MapReduce; Jimmy Lin and Chris Dyer; Morgan & Claypool Publishers, 2010
2. Search Engines: Information Retrieval in Practice; Croft, Metzler and Strohman; Pearson, 2010.

References

1. Selective Search: Efficient and Effective Search of Large Textual Collections; Kulkarni and Callan, TOIS, 2015.
2. Searching the Enterprise; Kruschwitz and Hull; FnTIR, 2017.
3. A General SIMD-Based Approach to Accelerating Compression Algorithms; Zhao et. al; TOIS, 2015.
4. Challenges in building large-scale information retrieval systems; Dean, WSDM, 2009.
5. Partitioning and Segment Organization Strategies for Real-Time Selective Search on Document Streams; Wang and Lin; WSDM, 2017.
6. Efficient Cost-Aware Cascade Ranking in Multi-Stage Retrieval; Chen et. al; SIGIR, 2017.
7. The Google file system; Ghemawat et. al; SOSP, 2003.
8. Bigtable: A Distributed Storage System for Structured Data; Chang et. Al,, TOIS, 2008.
9. MapReduce: Simplified Data Processing on Large Clusters; Dean and Ghemawat, OSDI, 2004.
10. Retrievability in API-Based "Evaluation as a Service"; Paik and Lin; ICTIR, 2016
11. LifeLogging: Personal Big Data; Gurrin et. al; FnTIR, 2014.
12. Federated Search; Shokouhi; FnTIR, 2011.