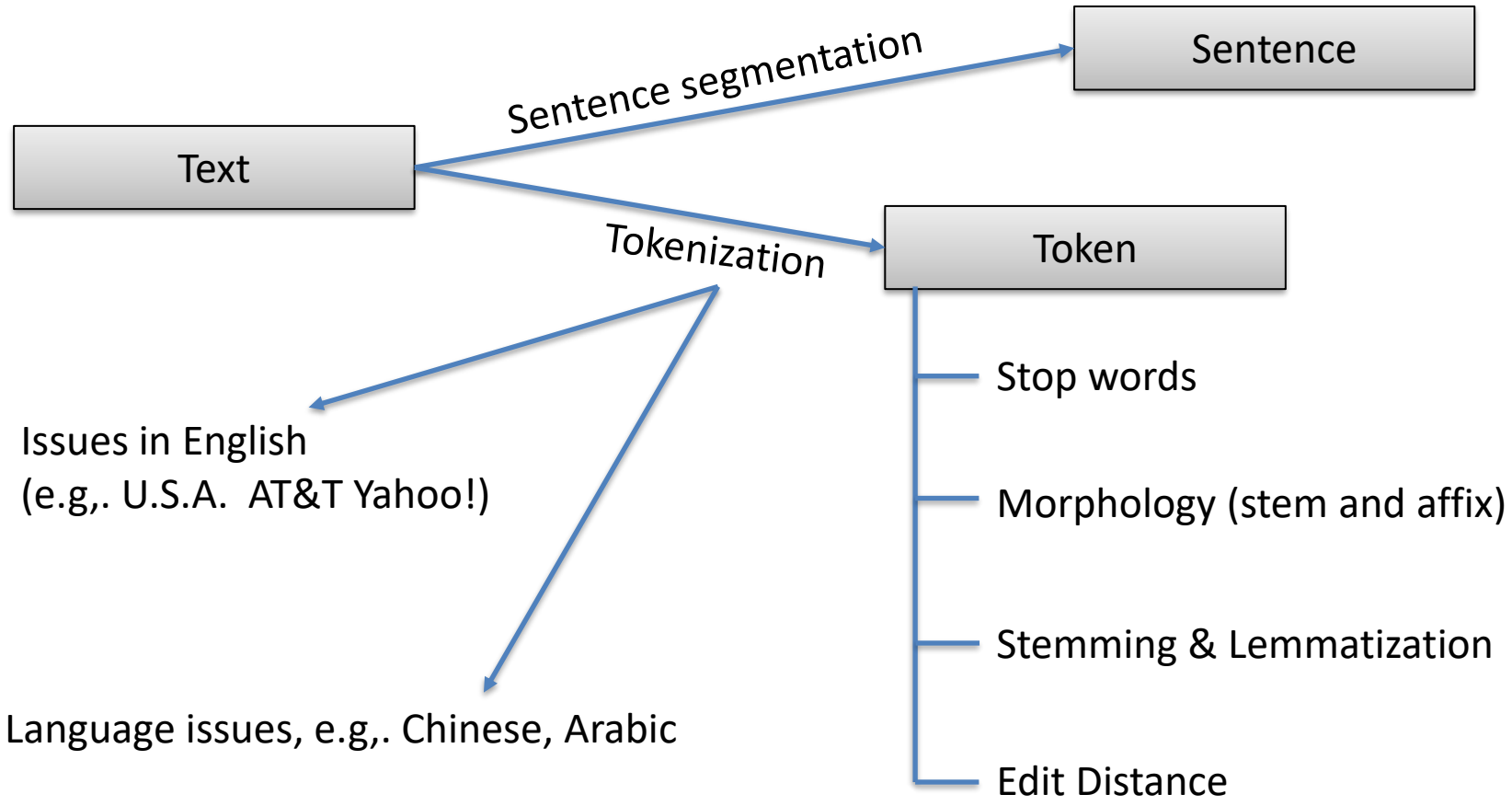
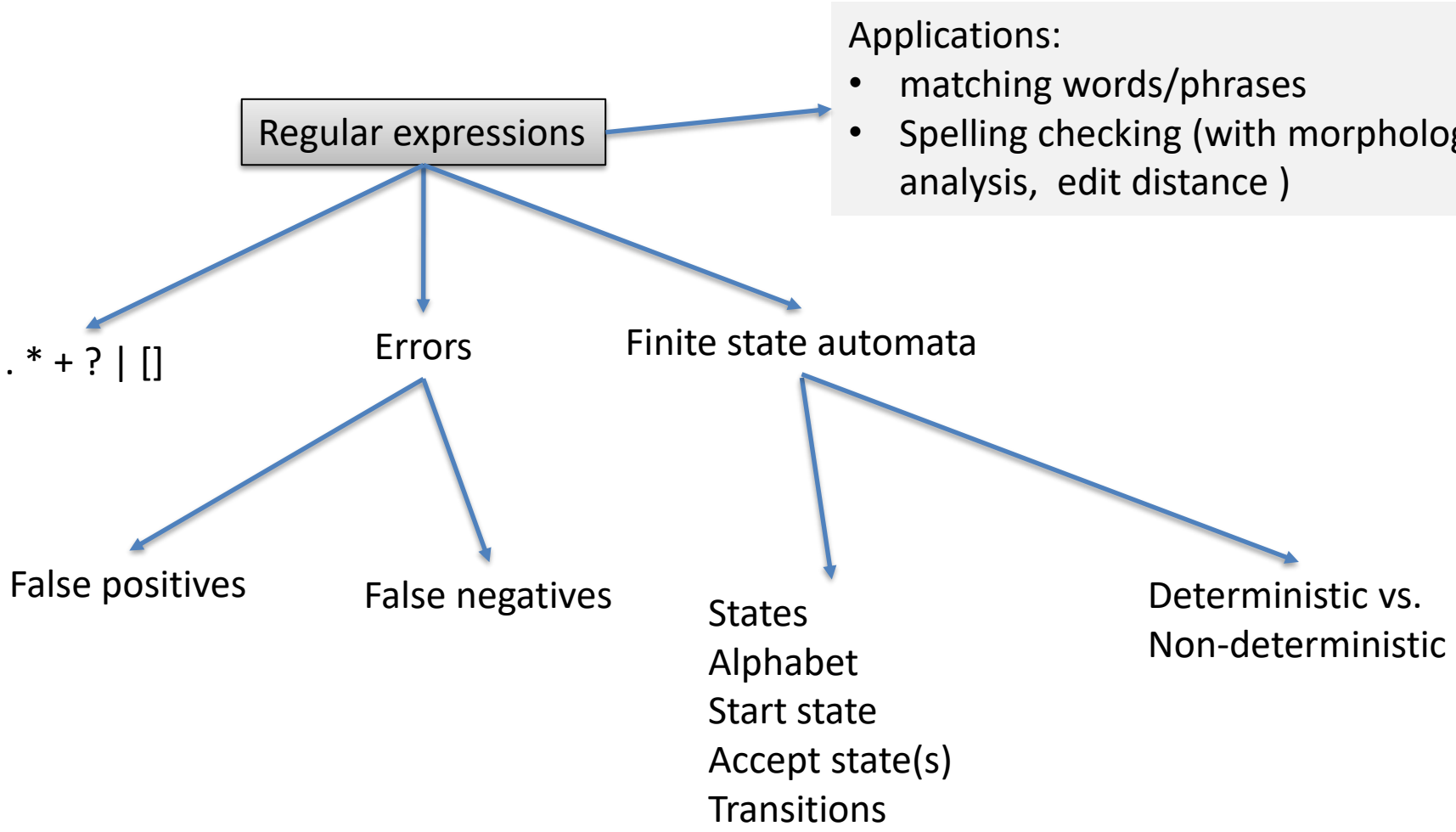
AI6122 Text Data Management & Analysis

Review - IR Topics

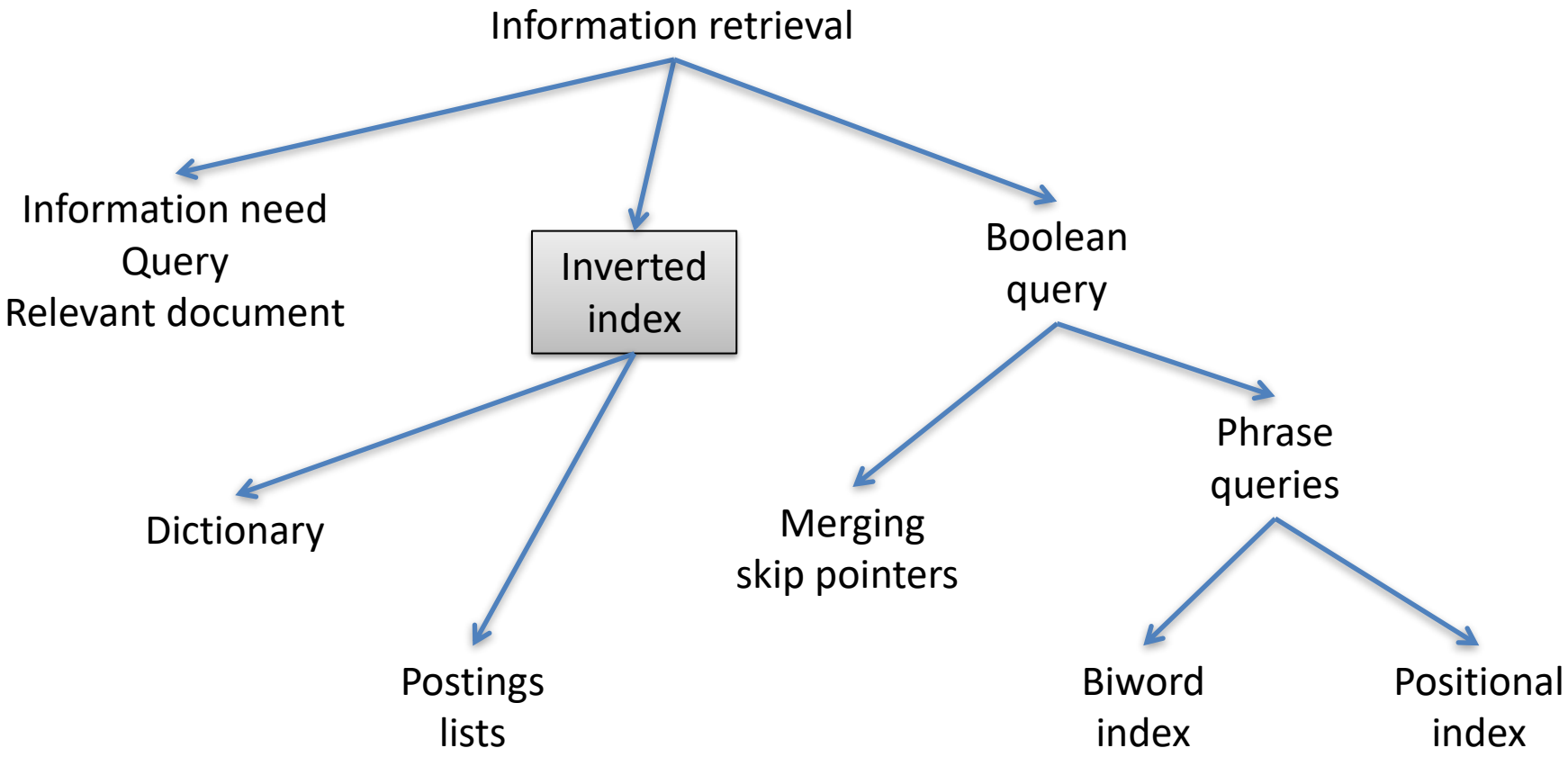
Tokenization, Morphology



Regular expressions

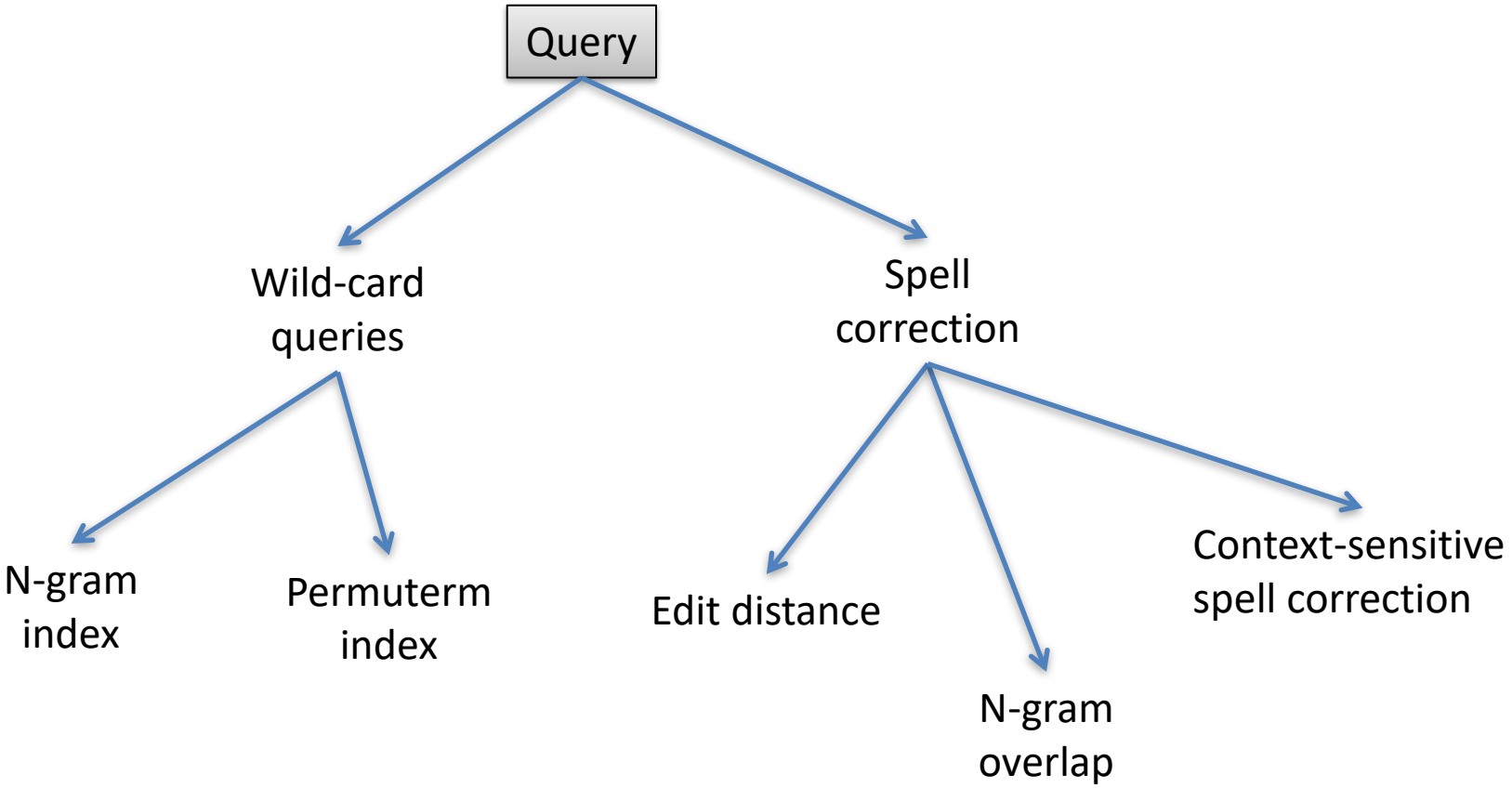


Boolean retrieval

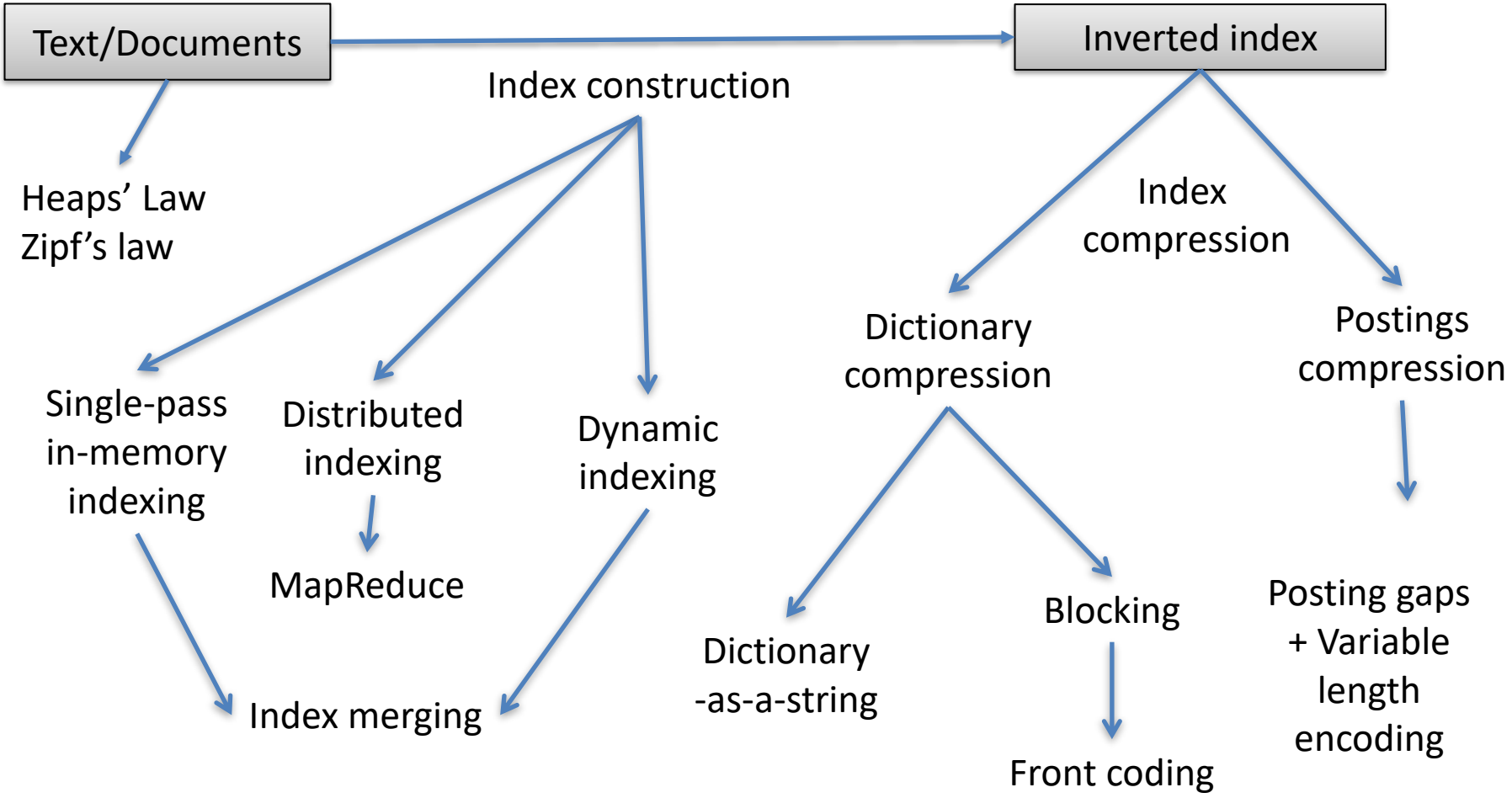


4

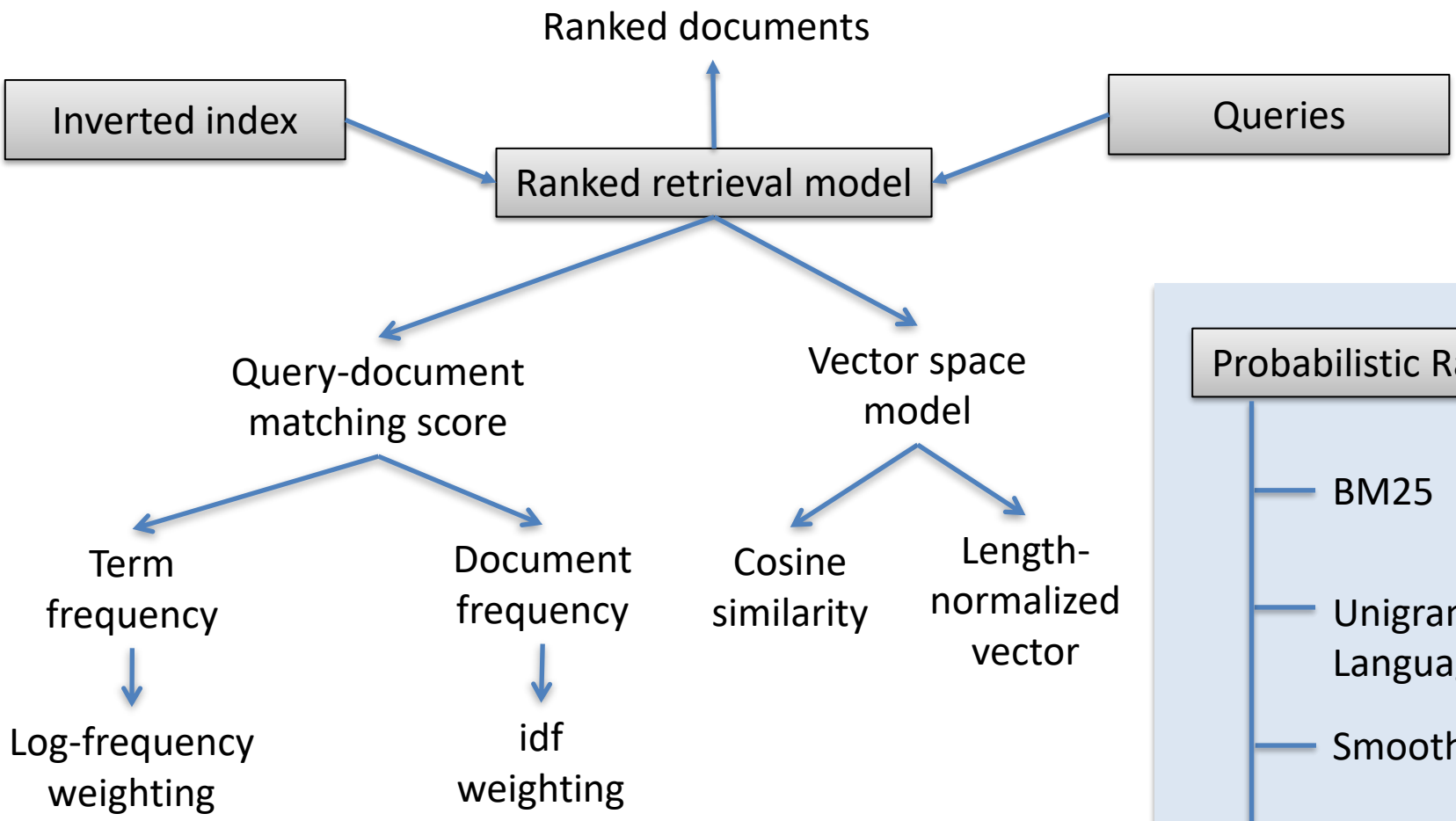
Tolerant retrieval

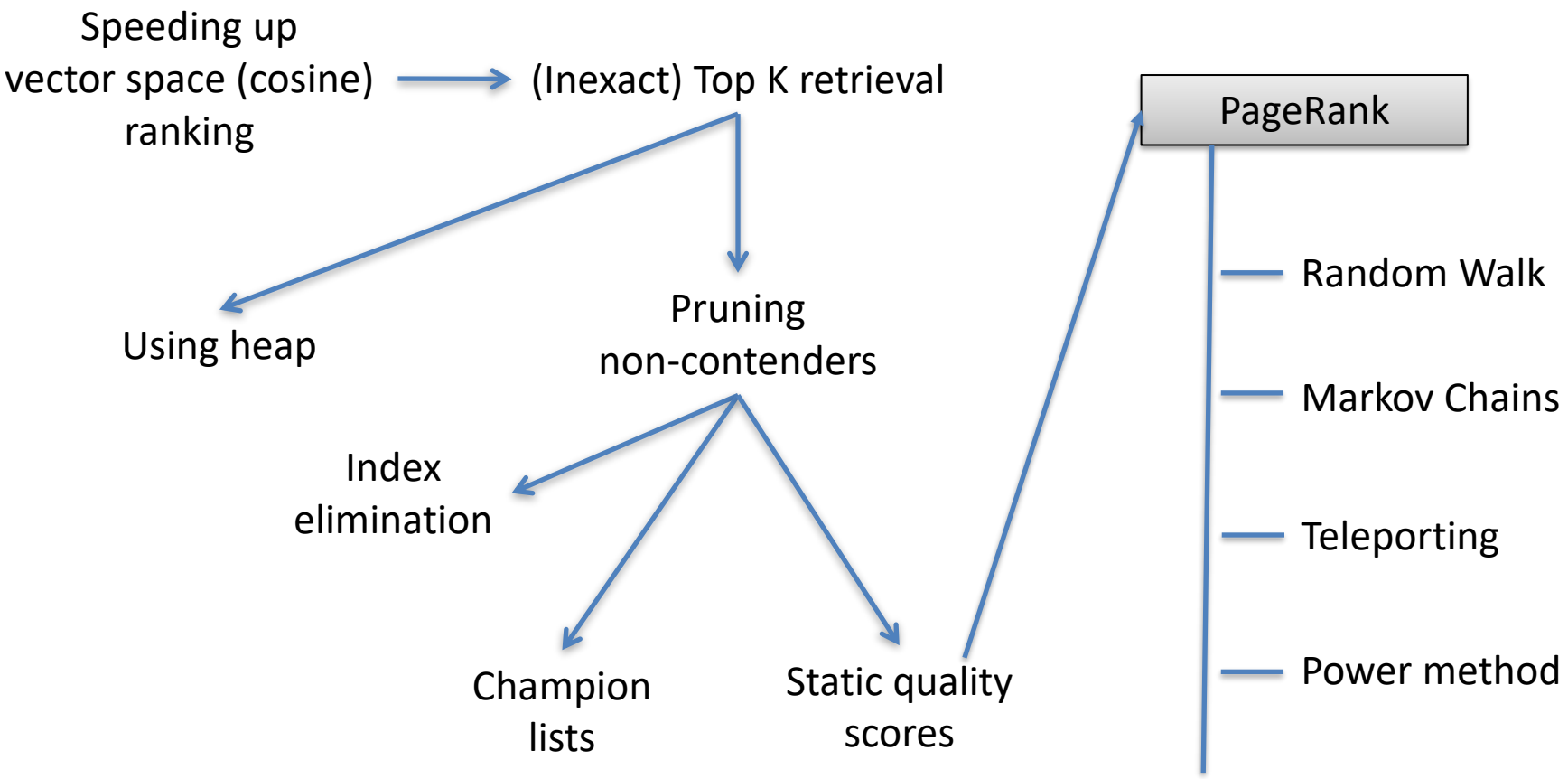


Index construction and Compression



TFIDF and Vector Space Model



- Probabilistic Ranking
  - BM25
  - Unigram Language model
  - Smoothing
  - LM vs VSM

Efficient scoring