Chapter IR:V

V. Retrieval Models

- Overview of Retrieval Models
- Empirical Models
- Boolean Retrieval
- Vector Space Model
- Probabilistic Models
- Binary Independence Model
- Okapi BM25
- Hidden Variable Models
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Generative Models
- Language Models
- Combining Evidence
- Web Search
- Learning to Rank

Learning to Rank

Machine Learning and IR

- Considerable interaction between these fields
  - Rocchio algorithm (1960s) is a simple learning approach
  - 1980s, 1990s: learning ranking algorithms based on user feedback
  - 2000s: text categorization
- Limited by amount of training data
- Web query logs have generated new wave of research
  - e.g., “Learning to Rank”

Learning to Rank

Generative vs. Discriminative

- All probabilistic retrieval models presented so far fall into the category of generative models
  - Assume that documents were generated from some underlying model
  - Use training data to estimate the parameters of the model
  - Probability of belonging to a class (i.e., the relevant documents for a query) is then estimated using Bayes' Rule and the document model
- A discriminative model estimates the probability of belonging to a class directly from the observed features of the document based on the training data
- Generative models perform well with low numbers of training examples
- Discriminative models usually have the advantage given enough training data
  - Can also easily incorporate many features

Learning to Rank

Discriminative Models for IR

- Discriminative models can be trained using explicit relevance judgments or click data in query logs
  - Click data is much cheaper, more noisy
  - E.g., a Ranking Support Vector Machine (SVM) takes as input partial rank information for queries
    - Partial information about which documents should be ranked higher than others
  - Partial rank information comes from relevance judgments (allows multiple levels of relevance) or click data
    - E.g., d_1, d_2 and d_3 are the documents in the first, second and third rank of the search output, only d_3 clicked on
      $\rightarrow \quad (d_3, d_1)$ and (d_3, d_2) will be in desired ranking for this query