CS293S Summary

2017 Tao Yang
Search Result Reply Pages

Main results

Advertisements

Suggestions recommendation
A Crawler Architecture

Similarity Analysis

Document → Shingling → Minhashing → Locality-sensitive Hashing

Candidate pairs: those pairs of signatures that we need to test for similarity.

The set of strings of length $k$ that appear in the document

Signatures: short integer vectors that represent the sets, and reflect their similarity
Online Engine: Architecture, Matching, Ranking

Web Search for a Planet: The Google Cluster Architecture
Document Ranking with Text, Quality, and Click Features

• Text features
  - TFIDF, BM25
  - Where do they appear? Title/body
  - Proximity (word distance)

• Document quality and classification
  - Web link scores (e.g. PageRank).
  - Page length, URL type etc.

• User behavior data
  - Presentation: what a user sees before a click
  - Clickthrough: frequency and timing of clicks
  - Browsing: what users do after a click
Learning to rank

- Convert ranking problem to a classification problem.
  - **Point-wise learning**
    - Given a query-document pair, predict a score (e.g. relevancy score)
  - **Pair-wise learning**
    - the input is a pair of documents for a query
  - **List-wise learning**
- Bayes, SVM, decision trees, human rules.
- Bagging/boosting to combine multiple schemes
Recommender vs Search Ranking

- Collaborative filtering:
  Similarity-guided recommendation

Text Content
Link popularity

User click data

Web page ranking

User rating

Content

Item recommendation

Sparse

User a

Item i

\[ p_{a,i} = \bar{r}_a + \sum_{u=1}^{n} w_{a,u} \left( r_{u,i} - \bar{r}_u \right) \]

\[ \sum_{u=1}^{n} w_{a,u} \]
Title:
- ACL-08: HLT Tutorial
  Computational Advertising Tutorial
  Columbus, OH - June 15, 2008
  research.yahoo.com

Creative:

Display URL:

Bid phrase: computational advertising
Bid: $0.5

Landing URL: http://research.yahoo.com/tutorials/acl08_compadv/