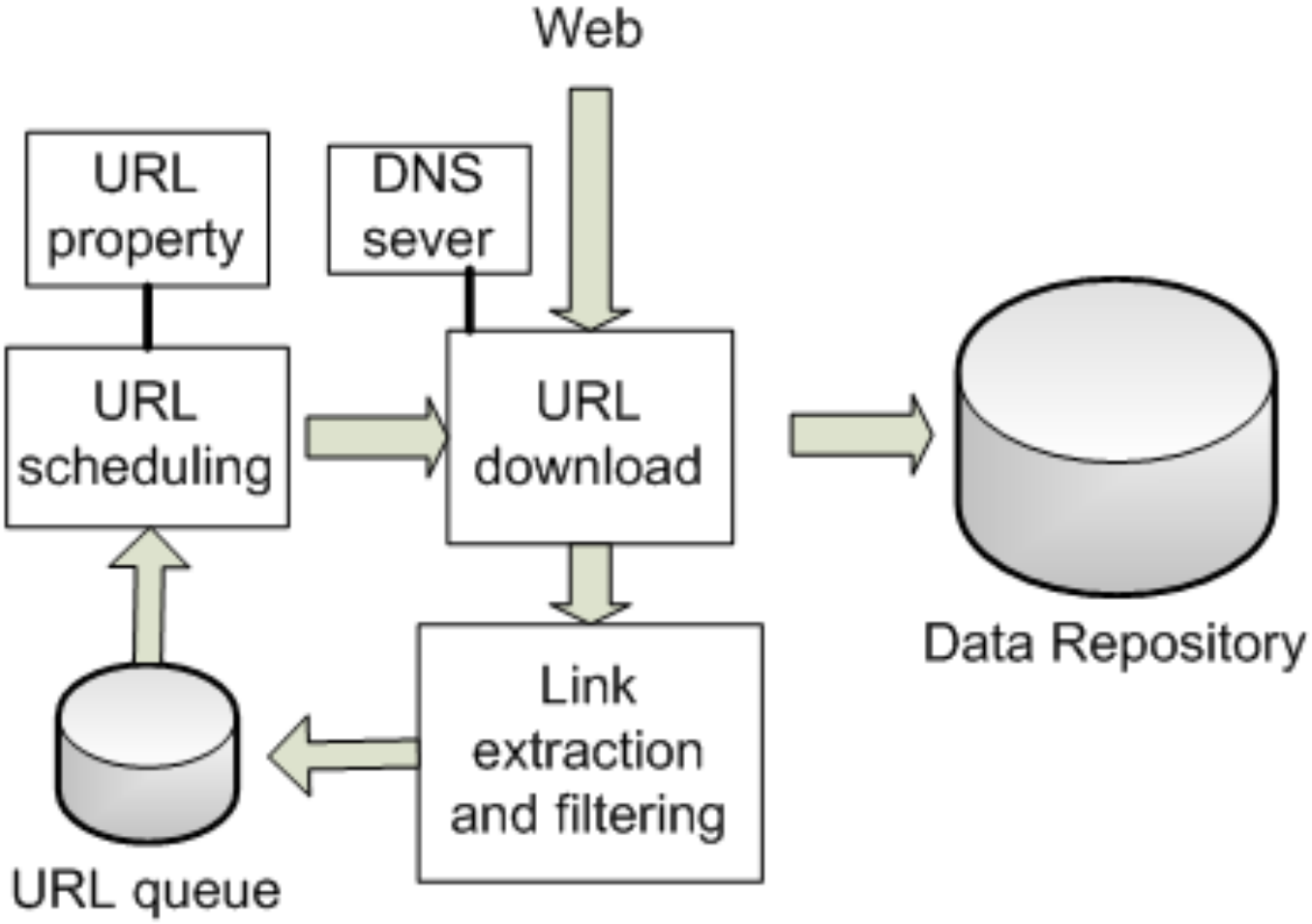
CS293S Summary

2017 Tao Yang

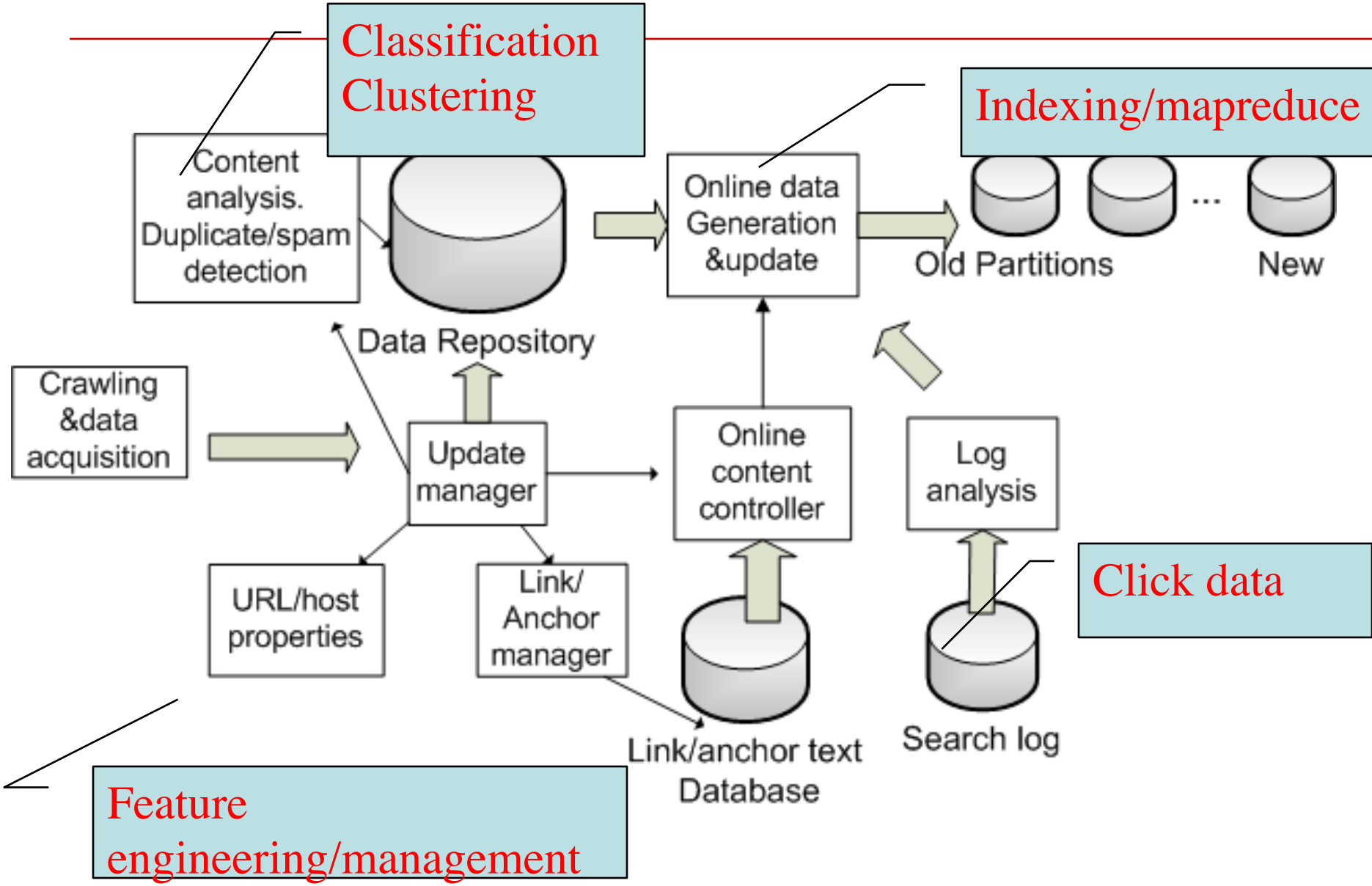
Search Result Reply Pages

A Crawler Architecture

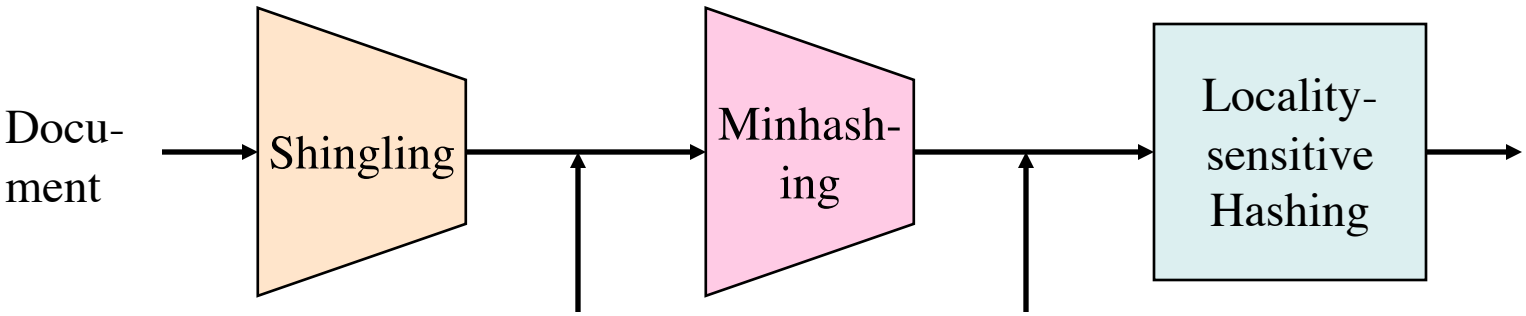


Olston/Najork. Web crawling.
Found. Trends Inf. Retr., 4(3):175--246, March 2010.

Offline Architecture



Similarity Analysis

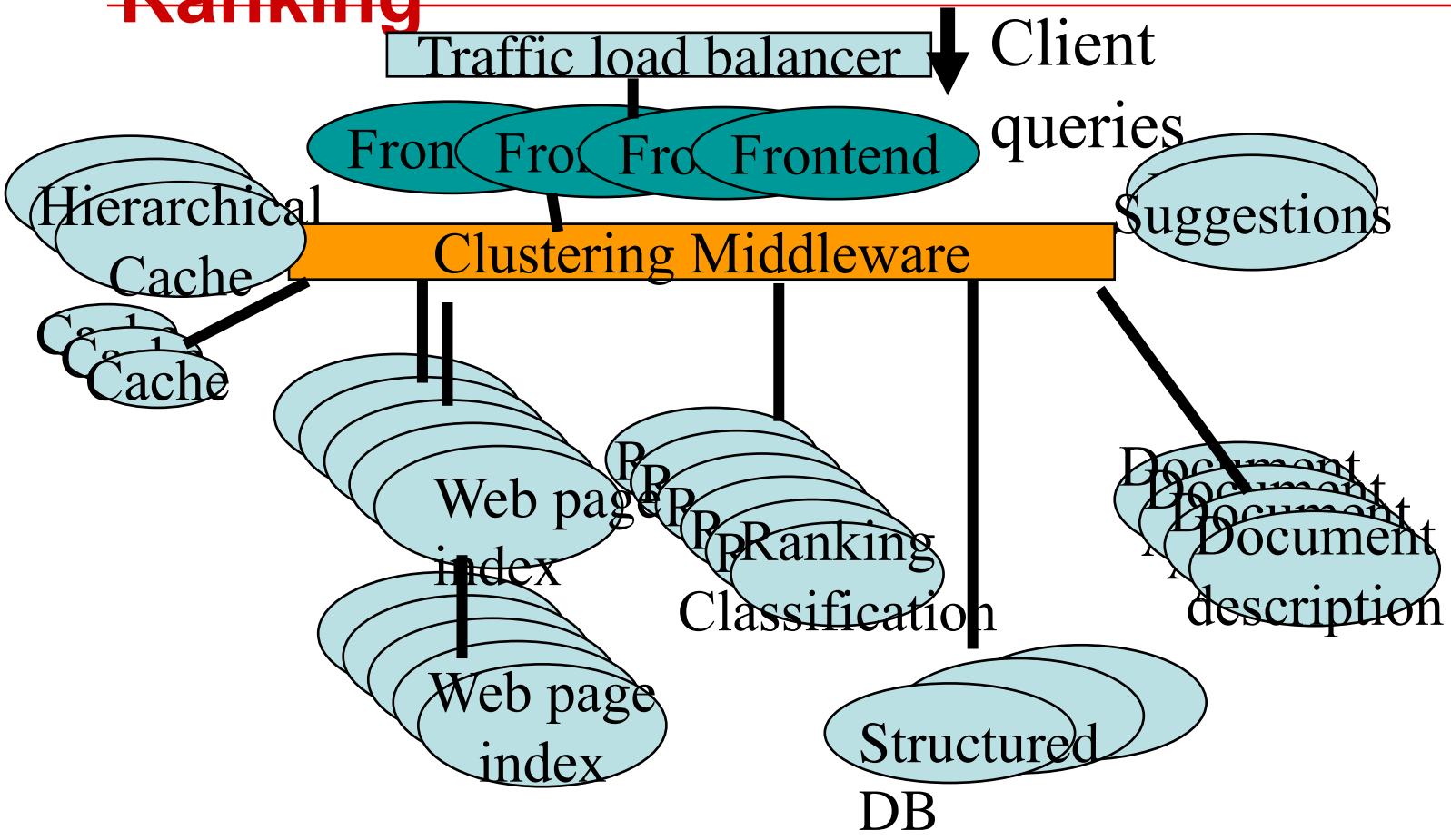


The set of strings of length *k* that appear in the document

Signatures : short integer vectors that represent the sets, and reflect their similarity

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

Online Engine: Architecture, Matching, Ranking



Web Search for a Planet: The Google Cluster Architecture
L. Barroso, J. Dean, U. Hölzle, IEEE Micro, vol. 23 (2003)

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

Recommendation vs Search Ranking

- **Collaborative filtering :**

  Similarity-guided recommendation

Sparse

User rating

Content

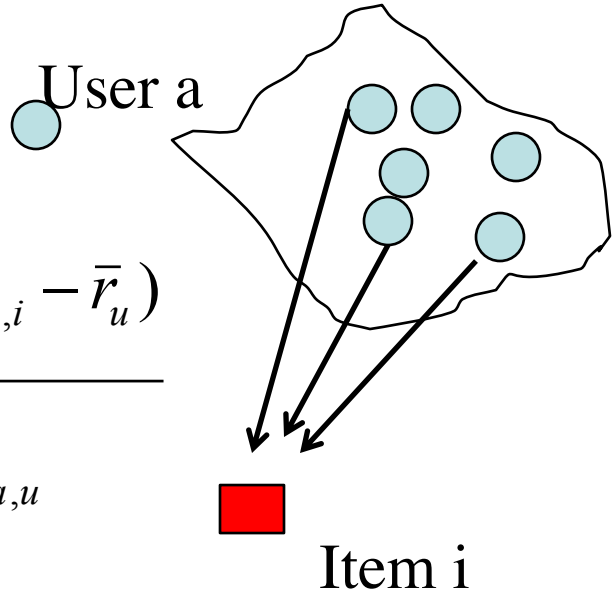
Item recommendation

Text Content
Link popularity

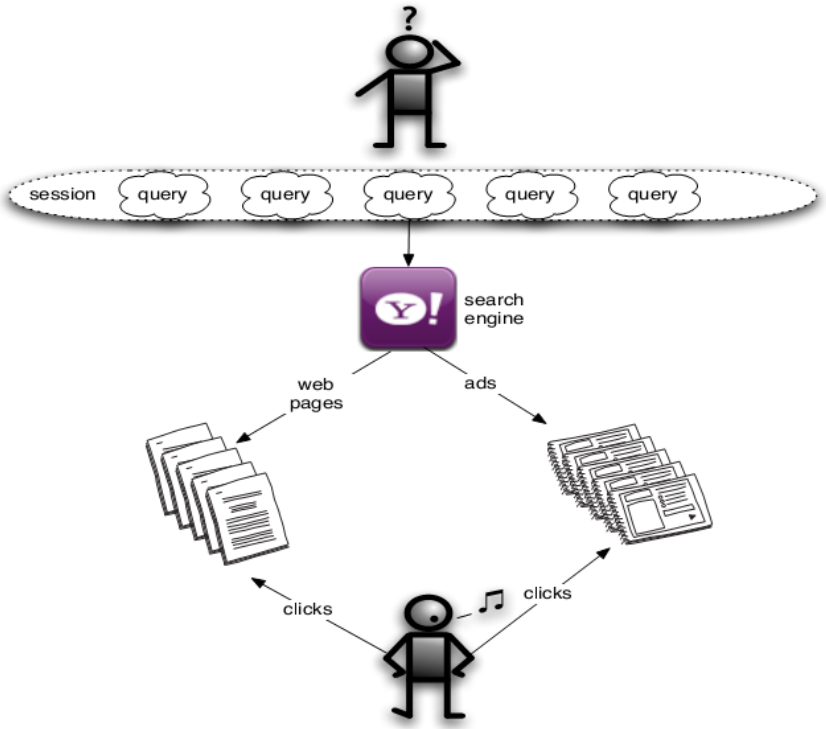
User click data

Web page ranking

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} w_{a,u}(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{n} w_{a,u}}$$



Search Advertisement



Bid phrase: computational advertising
Bid: $0.5

Title { ACL-08:HLT Tutorial

Creative { **Computational Advertising** Tutorial
Columbus, OH - June 15, 2008

Display URL { research.yahoo.com

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