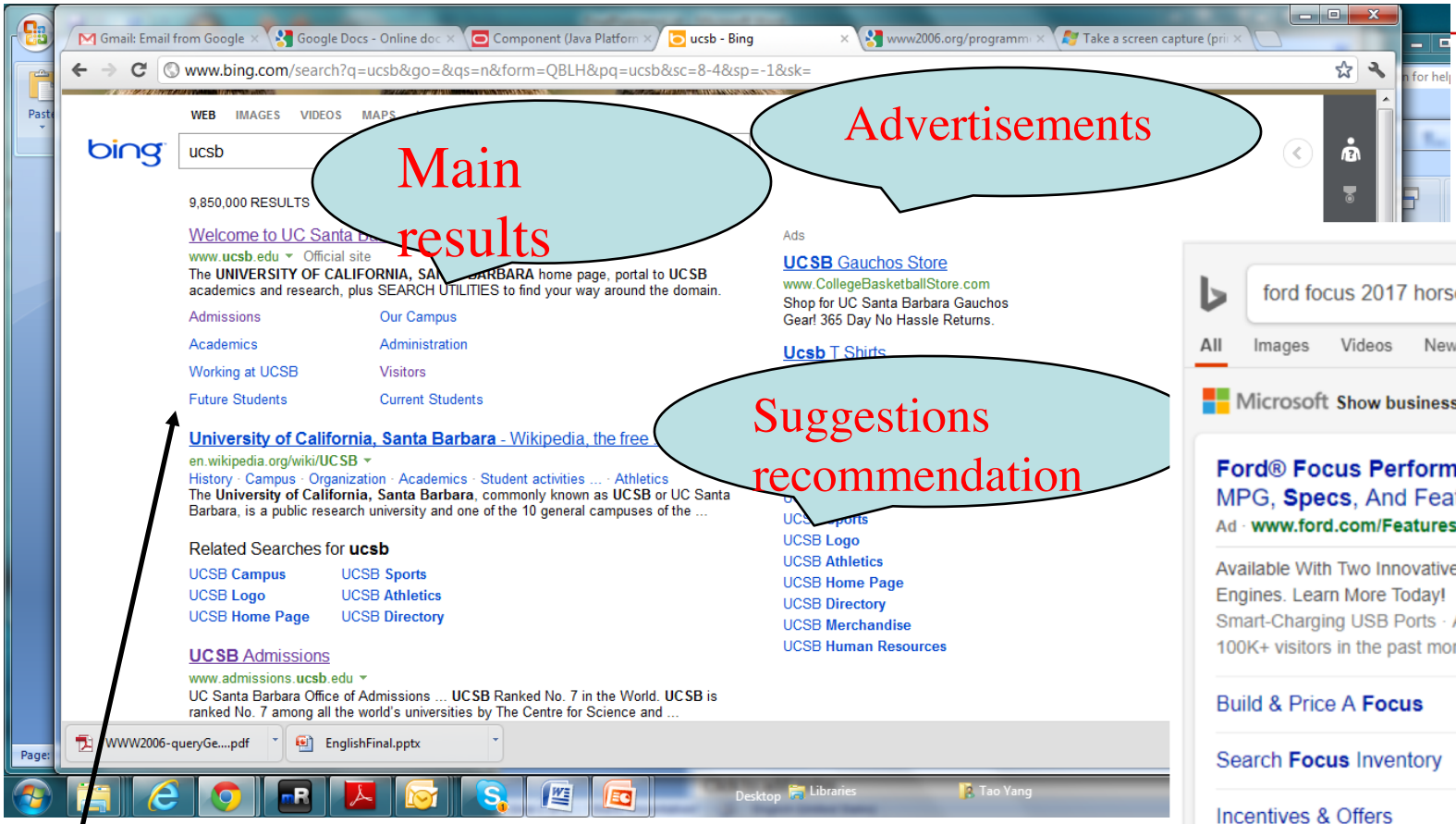
CS293S Summary

2022 Tao Yang

Result Reply Pages of Search Engines

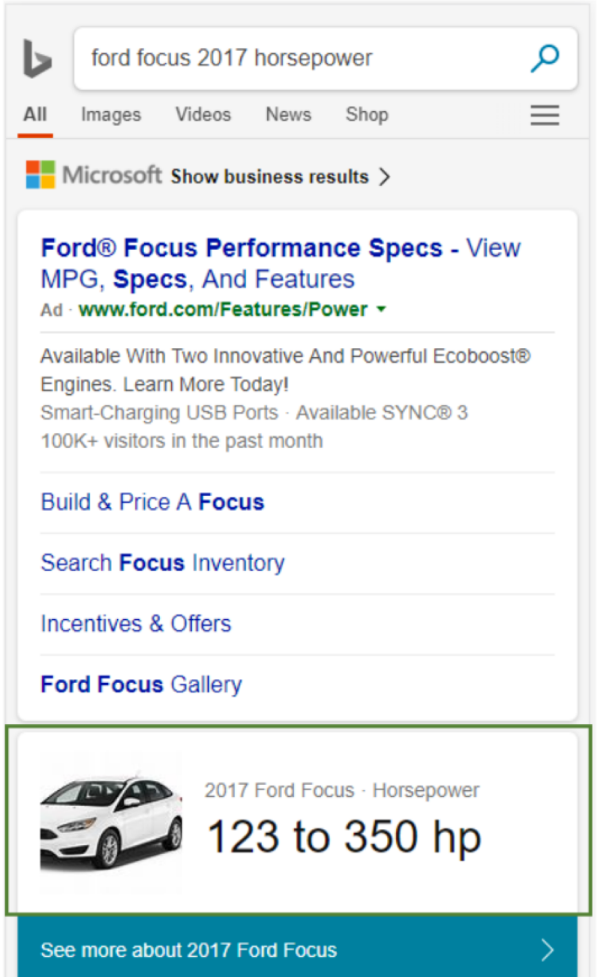


Main results

Advertisements

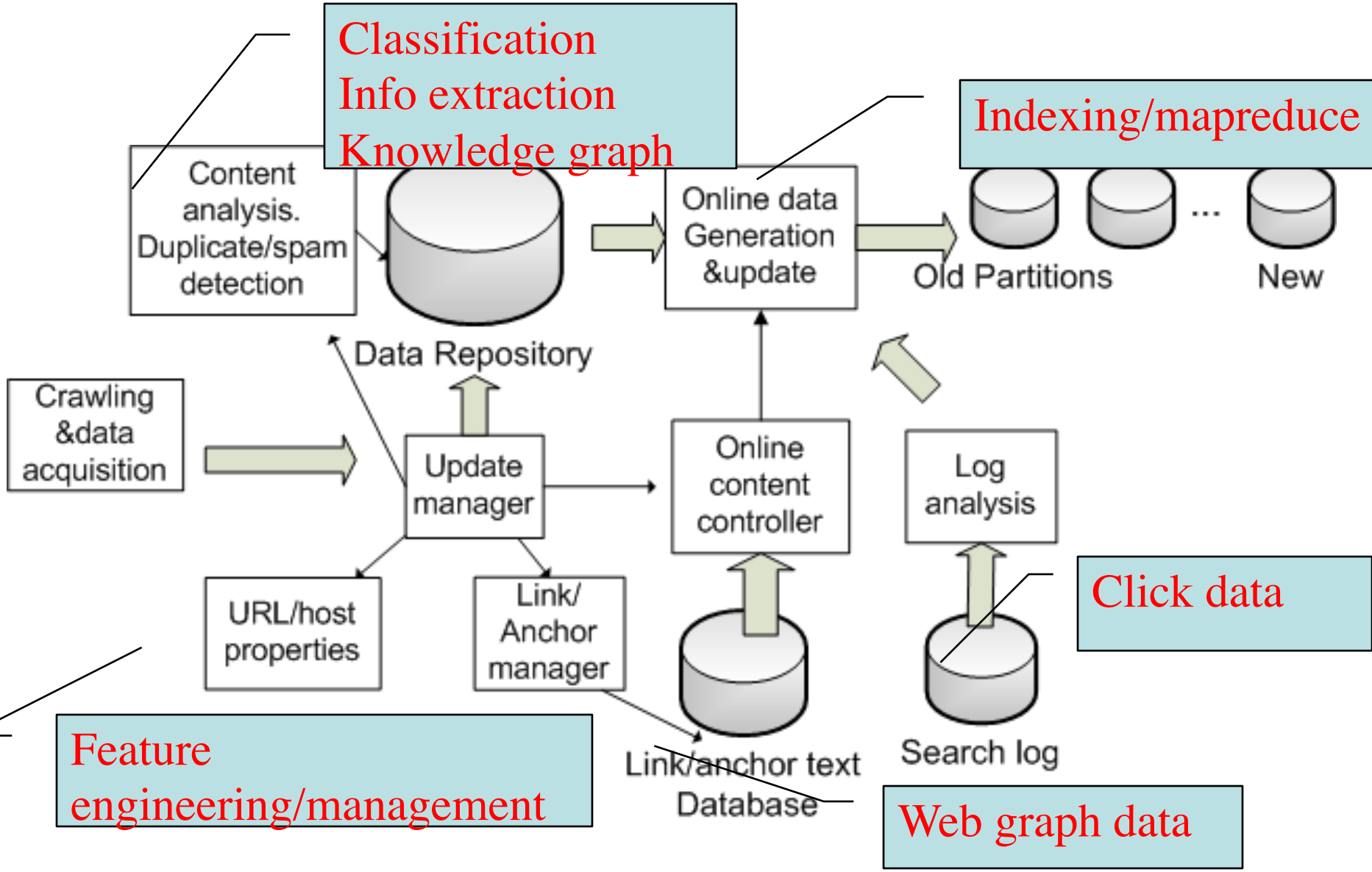
Suggestions recommendation

Popular links of top site



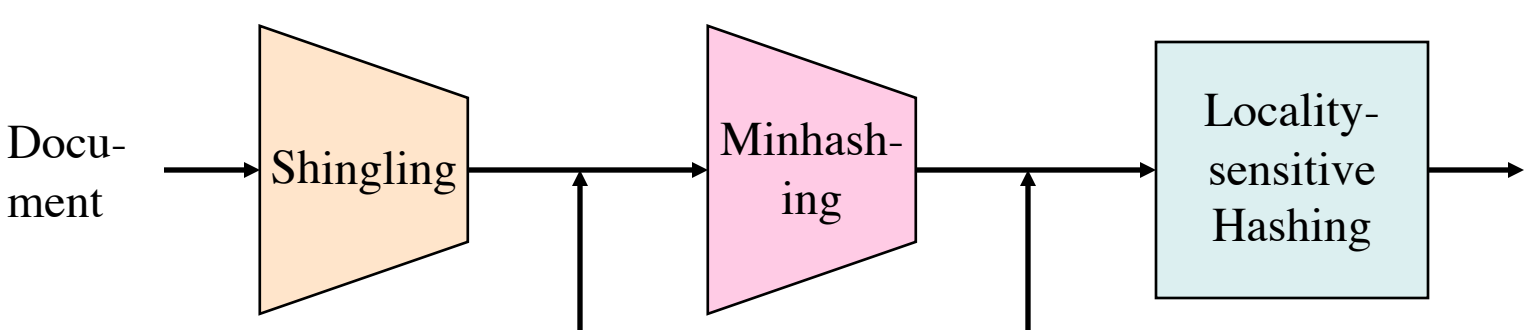
Direct answer

Offline Architecture with Continuous Updating: Crawling, Classification/Mining, and Indexing

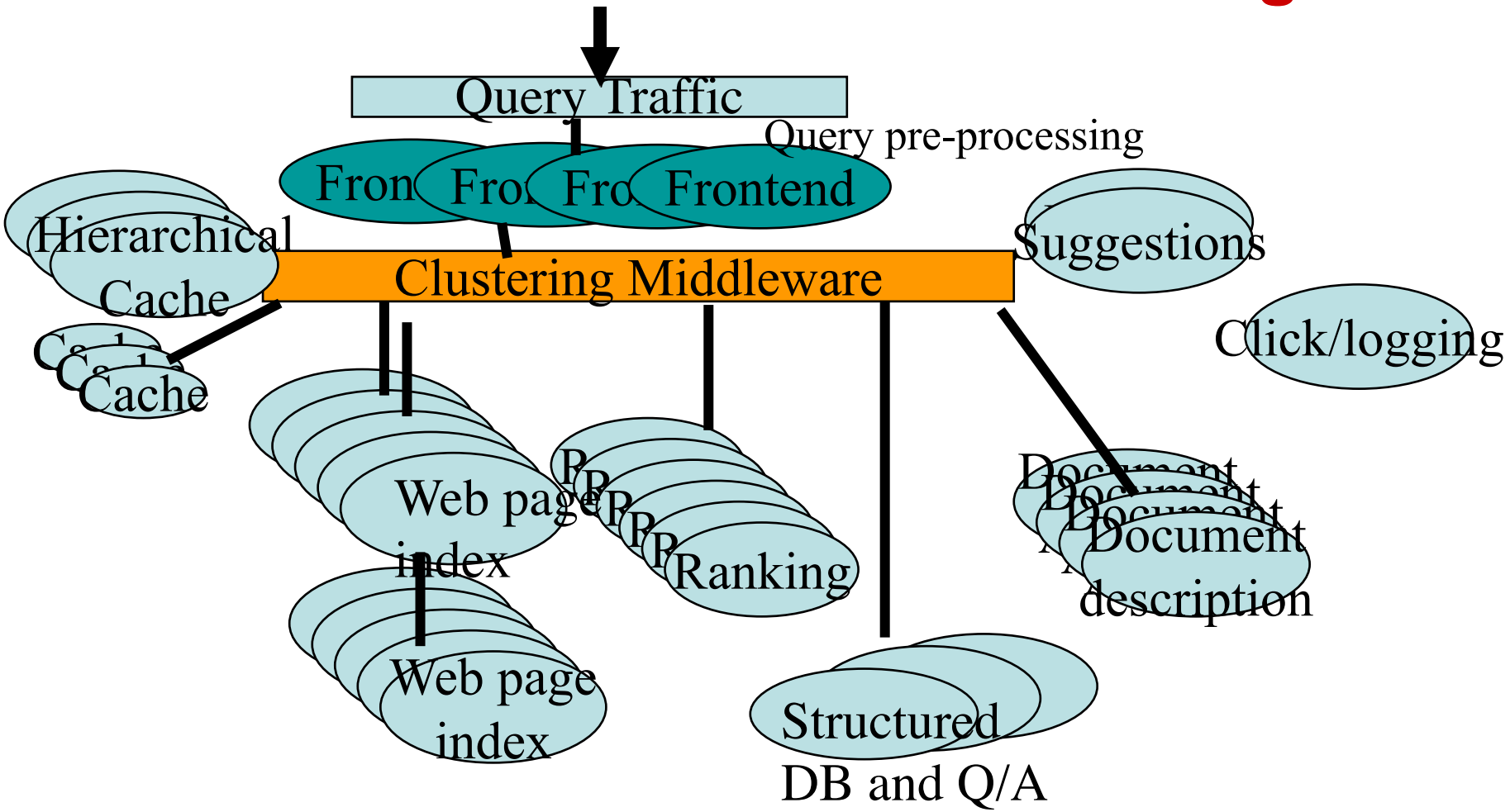


Near Duplicate Detection and Similarity Analysis

- ~15% of web pages crawled are duplicated
- ~30% are near duplicate



Online Search Engine Architecture with Document Retrieval and Ranking



Document Ranking with Text, Quality, and Click Features

- **Text features**
  - TFIDF, BM25 in 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
- **Neural features with word embeddings for semantic similarity**

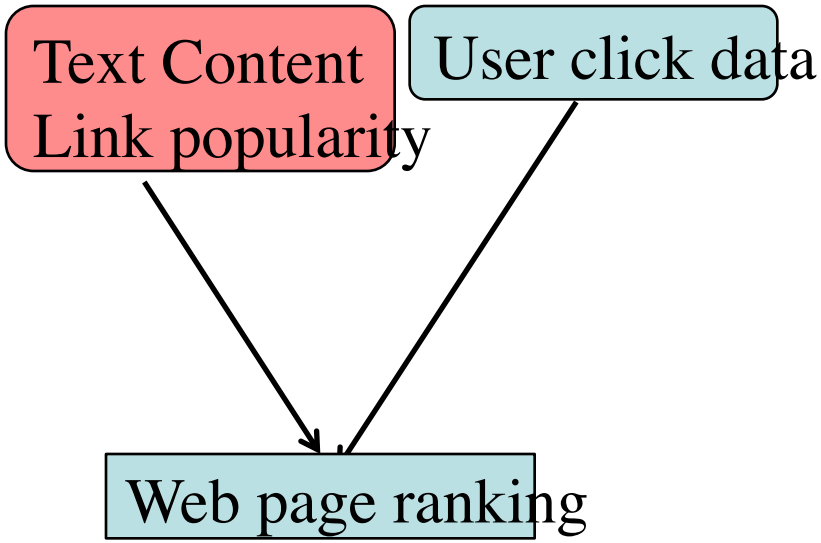
Learning-to-Rank and Classification

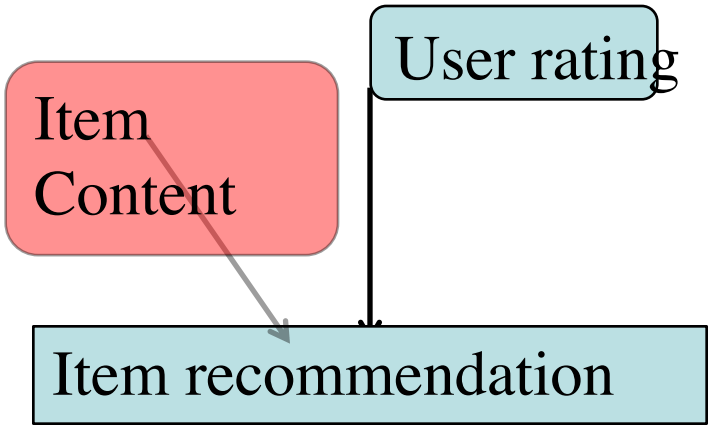
- Convert ranking problem to classification/regression
  - *Point-wise* learning
    - Query-doc optimization
  - *Pair-wise* learning
    - the input is a pair of documents for a query
  - *List-wise learning*
    - Optimize a ranked document list for a query
- Human rules, linear, tree ensembles, similarity-driven search including neural ranking/transformers
  - Learning ensembles with bagging/boosting to combine multiple schemes
- **Metrics:** NDCG, Precision/recall, MAP/MRR

Papers from Group Projects by Wed 29

- **Offline index optimization and data processing**
  - Document expansion by Aman and Krushna for better search
  - Search index compression with quantization by Jiaxin
  - Vertical classification for software bugs by Satyandra, Harsha
- **Online document retrieval**
  - Dense retrieval and FAISS by Zihan Ma
  - Sparse retrieval comparison by Hunter/Ye
  - Zero-shot dense retrieval by Shanxiu
  - Graph neural matching by Zekun and Zheng
  - Online document retrieval by Wenda and Yujie
  - Text/Image search by Weixi
- **Ranking**
  - Counterfactual Online Learning To Rank by Lucas/Gautam
  - Fast passage re-ranking by Arjun and Taanya
  - BERT for queries with typos by Carina and Lianke
  - Explainable search results by Zichen and Qiucheng
- **Document retrieval applications:** Reasoning and acting by Alex
- **More today:**

From Search Ranking to Recommendation & Advertisement





Advertisement

- Match ads to query/context
- Order the ads
- Pricing on a click-through

IR

Econ

Takeaways from 293S

- **Search (Information Retrieval) has many applications**
- **Key characteristics of such info systems**
  - Data intensive. Large-scale. Lots of noise
  - Uncertainty of requirements and metrics
    - Many choices for algorithms and system design.
    - Both quality and computing resource affects choices
- **Data-driven approaches**
  - Develop metrics to narrow requirements.
  - Don't be afraid of challenging algorithms and metrics, and requirements
- **Thrive in uncertainty and large-scale processing**
  - Winner determined by data and evaluation results.

Course project or related IR topic→ master project
Contact me

Course Evaluation https://esci.id.ucsb.edu

- Posted in Piazza
- Online form is open until midnight of this Friday
- Please let us know whether you are learning what I hope you are. Your positive support is greatly appreciated