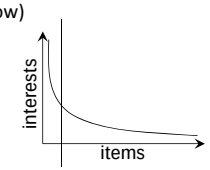
Recommendation Systems

Kai Shen

Recommendation Systems

- Advertising on traditional medias
  - Can't be customized to each user; many ad showings do not reach intended targets
- Big data on user behaviors tell what a user may like
  - Allow online, targeted advertising
  - Broaden the ad market (low-interest items are appropriate targets now)



  - Make online advertising much more effective ⇨ for the first time allow very large businesses to thrive on advertising alone

Applications

- Search ads
- Product recommendations at online retailers
- Movie recommendations
- YouTube suggestions
- Recommendations for news articles, blogs
- … …

Methods for Recommendation

- Identify user/item profiles and match them for recommendation
  - Content-based recommendation
- Link similar users and identify preferred items by similar users as recommendation
  - Collaborative filtering
  - *Into Thin Air* and *Touching the Void*

Content-Based Recommendation

- Identify user/item profiles and match them for recommendation
- Search ads
  - Item profiles are categories and keywords for ads
  - User profiles are the keywords user provided for search

Discover Features of Documents

- Too much information: a document contains many words (terms)
  - A term appearing many times tends to describe the document better
  - But not always the case ⇨ if it occurs in every document (“the”), then it isn’t really descriptive
- TF.IDF: term frequency–inverse document frequency
  - Raw count, boolean, logarithmic
- Similarity of documents
  - Jaccard distance: intersection size divided by the union’s size
  - Cosine distance: normalize the vectors and then compute inner products

Obtain Item Features from Tags

- Too little descriptive information: images
  - Tagging
- How to tag images?
  - Computer vision
    - Mapreduce?
  - Web crawling and mining
  - Image tagging from computer games

Obtain User Profiles

- Probably the most valuable data are those that contain user activities or behaviors
- Direct: search keywords, filing out profiles/surveys
- Indirect inference:
  - blogposts, tweets
  - browsing history (offer free content to attract users, infer their profiles, and show targeted recommendation)
  - offer free entertainment (games) to attract users, …
  - … …

Making Recommendation

- Similarity of user/item profiles
  - Jaccard distance, cosine distance, …
- Machine learning: train a decision tree
  - http://en.wikipedia.org/wiki/Decision_tree_learning
  - How to train? Iterative Dichotomiser 3.
  - Not so great result, but fast decision

Collaborative Filtering

- Find similar users for a given target user and identify preferred items by similar users as recommendation
- Clustering users

Big Data Recommendations

- A large amount of data, many users (so data can still be sparse)
- Similarity matching against a large dataset
  - Given a user profile (set of keywords), find the best matches with existing items (products, ads, movies, …)
  - Mapreduce: each task finds matches within a partition of the items
  - Not scalable: N users, M items ⇨ total efforts proportional to N×M

Big Data Recommendations

- Clustering (k-means)
  - Each iteration of K-means: given K centers, each sample is grouped into the nearest center
  - Can be done in MapReduce
    - Map: find the nearest center for one or a few samples
    - Reduce: aggregate the results

Big Data Recommendations

- Decision tree (ID3)
  - Iterative process to identify the most descriptive parameter and partition the dataset
  - Each iteration of ID3: within the current dataset, find the parameter to produce most information gain
  - Can be done in MapReduce?
    - Map: find the information gain of using each parameter to partition the dataset
    - Reduce: find the best parameter
  - Problem: each task needs the whole dataset ⇨ poor locality
    - Map task within a data partition?

The NetFlix Challenge

- Movie recommendation

Offline/online Recommendation

- Offline recommendation
  - Have data about items and users
  - Identify recommendations and mail them out
- Online recommendation
  - Have data about items and past users
  - For a new user with a bit of usage data, identify recommendation quickly and show on screen
- Optimized online recommendation
  - Optimize certain goal (profit) with additional conditions (advertiser constraints and offer)

Disclaimer

- Preparation of this class was helped by materials in the online book “Mining of Massive Datasets” by Rajaraman, Leskovec, and Ullman.