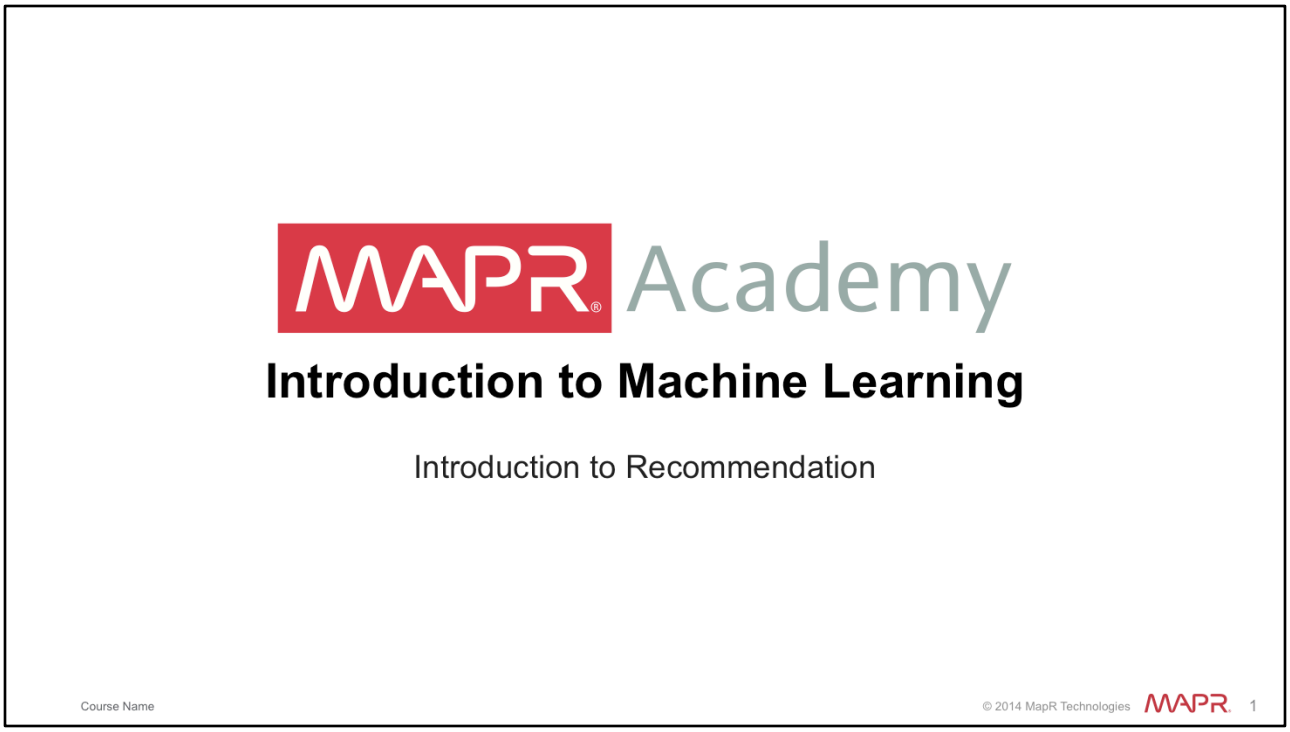
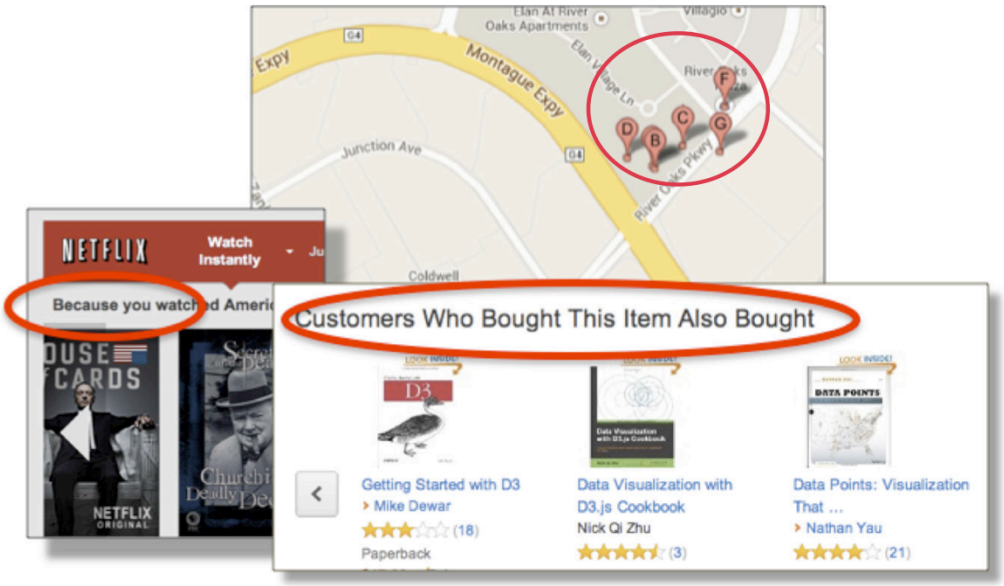

MAPR Academy
Introduction to Machine Learning
Introduction to Recommendation
Course Name
© 2014 MapR Technologies
MAPR
1

Introduction to Recommendation

- Define recommendation terms and concepts
- Describe the K-nearest neighbors algorithm
- Evaluate a recommendation model
- Discuss the Mahout implementations for recommendation

Recommender examples in the wild

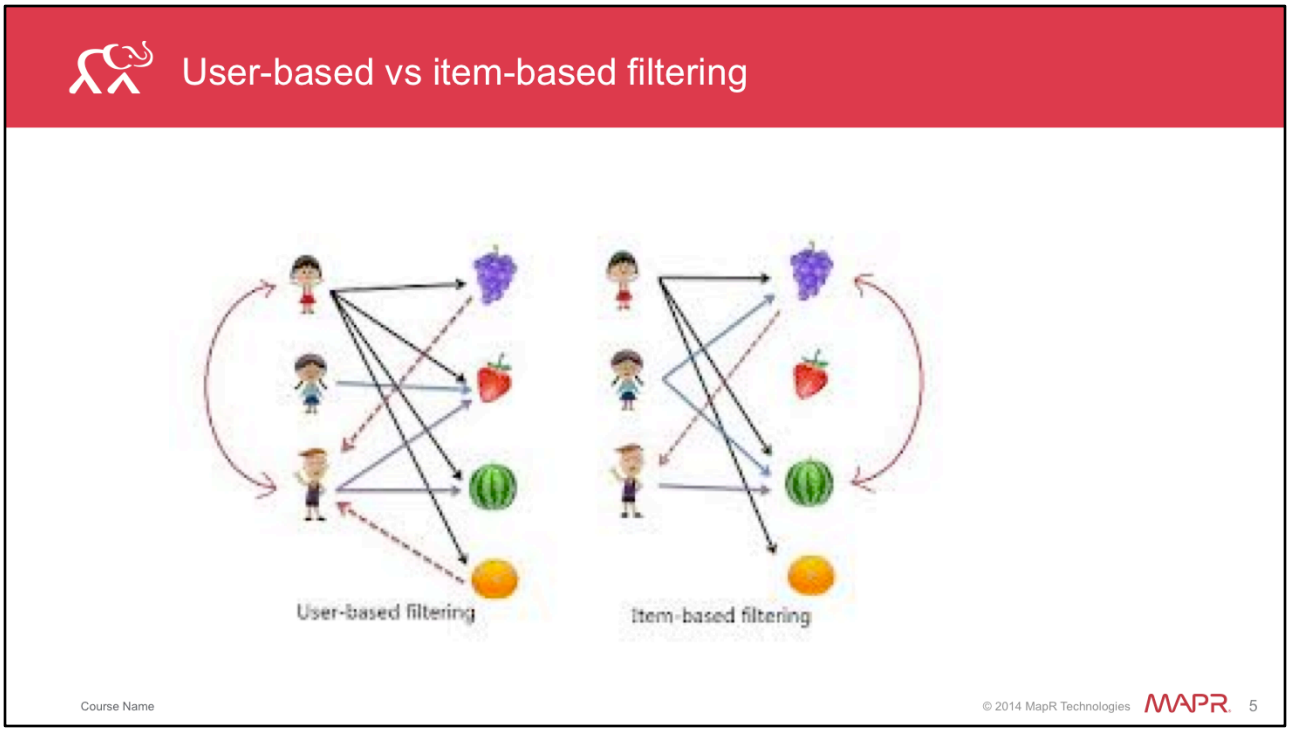


What is recommendation?

- **Recommendation** is a class of ML that seeks to predict a user's *preference* for or *rating* of an item

- **Recommender systems are used in industry to recommend:**
  - Books and other products (e.g Amazon)
  - Music (e.g. Pandora)
  - Movies (e.g. Netflix)
  - Restaurants (e.g. Yelp)
  - Jobs (e.g. LinkedIn)
  - . . . LOTS . . .

- **Main approaches to recommendation**
  - Collaborative filtering
  - Content-based filtering



Users are associated with item ***preferences*** through history

Similarity is constructed between users (or items)

Recommendations based on ***similarity to other users (or items)***

Similarity metrics

- **Pearson correlation**
  - Ratio of covariance to product of standard deviations
  - -1 ➔ inversely proportional; 0 ➔ no correlation; 1 ➔ directly proportional
- **Euclidean distance**
  - Coordinates represent item preferences (i1, i2, i3, …)
  - Smaller distance ➔ more similar (so return 1/(1+d))
- **Tanimoto coefficient**
  - Ratio of intersection to union
  - Between 0 and 1 ➔ bigger is more similar

Differences between collaborative and content-based filtering

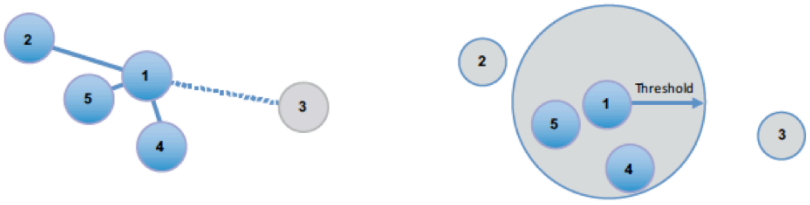
- ***Collaborative-based*** filtering is ***deductive*** (needs user-item history to start with to learn associations)
- ➔ example: last.fm
- ***Content-based*** filtering is ***inductive*** (needs domain knowledge to construct associations between users and items)
- ➔ example: Pandora

Challenges with collaborative filtering

Challenge	Description
Cold start	No user history means no associations day 1
Scale	Huge number of products and users requires a lot of computation
Sparsity	Most users express very little behavior with very few items and no behavior on the vast majority of items

User neighborhood

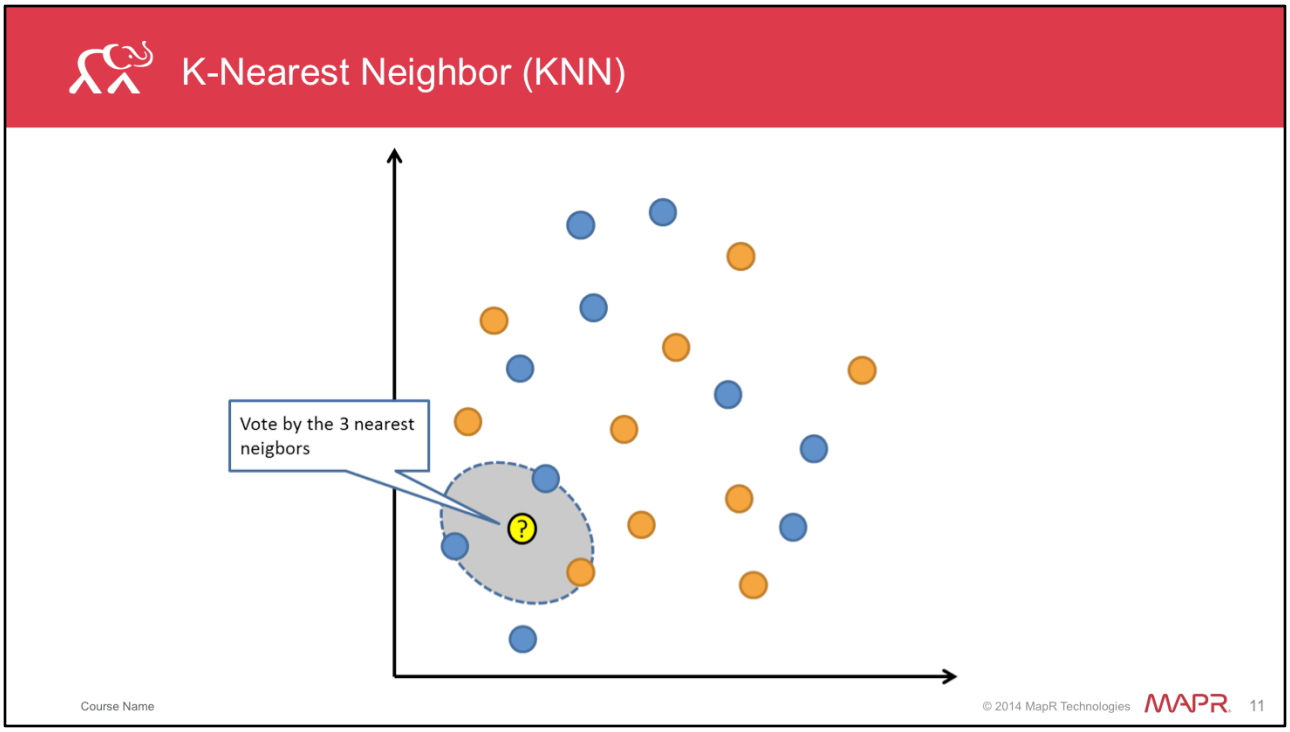
- A neighborhood (of users) is a group of similar users
- 2 types of neighborhoods:
  - ***Fixed-size***: cardinality of neighborhood is fixed a priori (e.g. top ten)
  - ***Threshold-based***: degree of similarity is fixed a priori (e.g. >= 90%)



Q: why no *item* neighborhood in collaborative filtering?

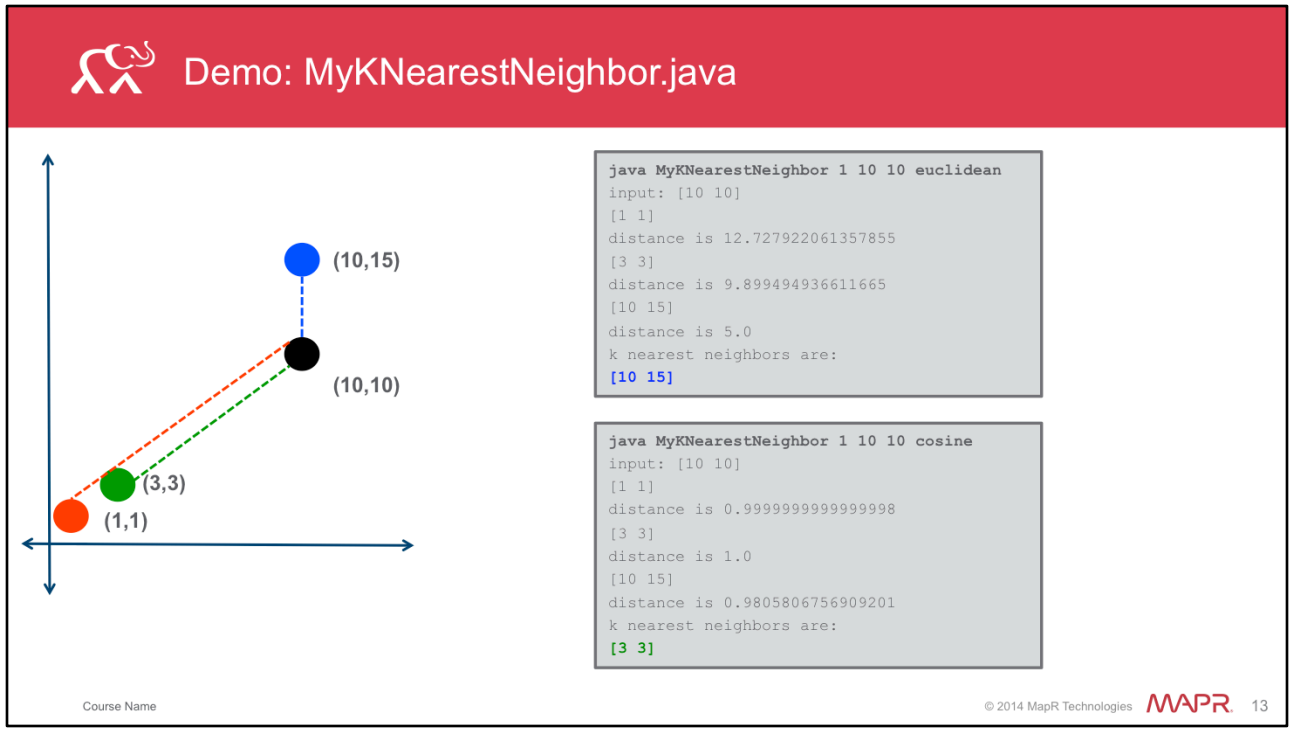
Introduction to Recommendation

- Define recommendation terms and concepts
- Describe the K-nearest neighbors algorithm
- Evaluate a recommender model
- Discuss the Mahout implementations for recommendation


K-Nearest Neighbor (KNN)
Vote by the 3 nearest neigbors
?
Course Name
© 2014 MapR Technologies
11

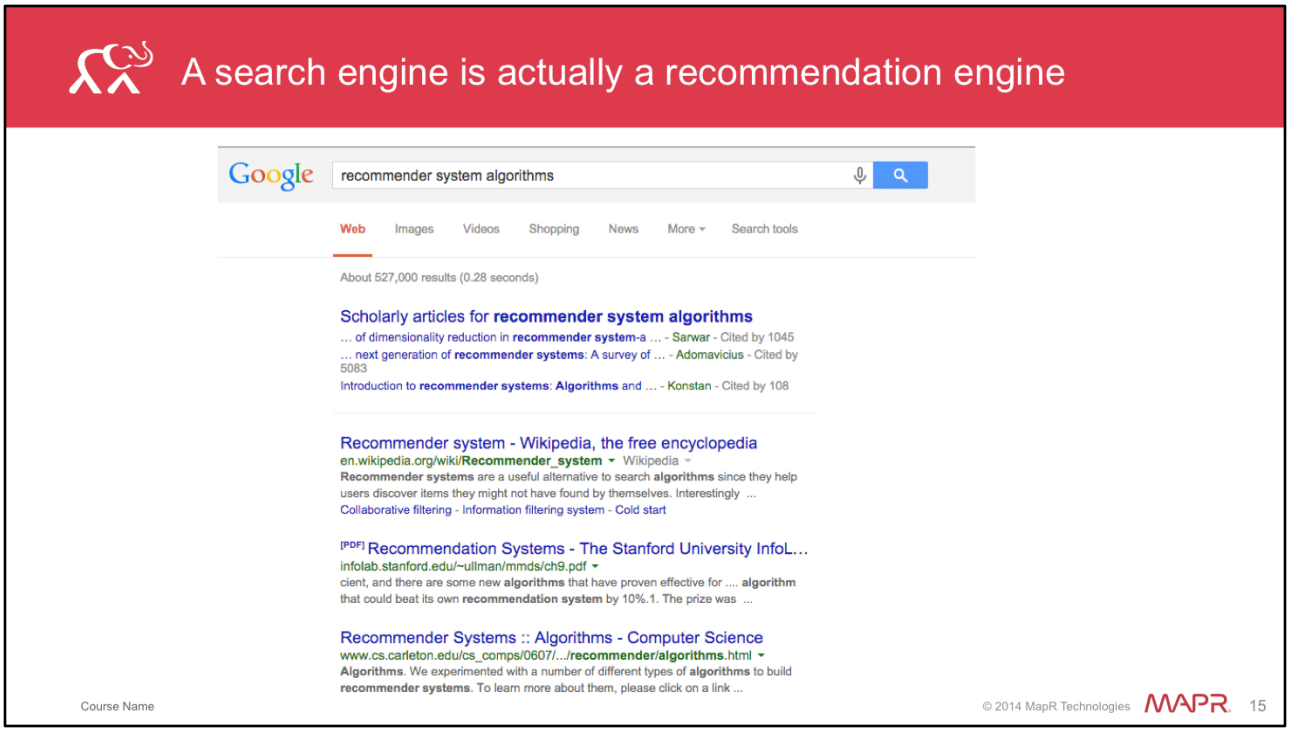
KNN algorithm

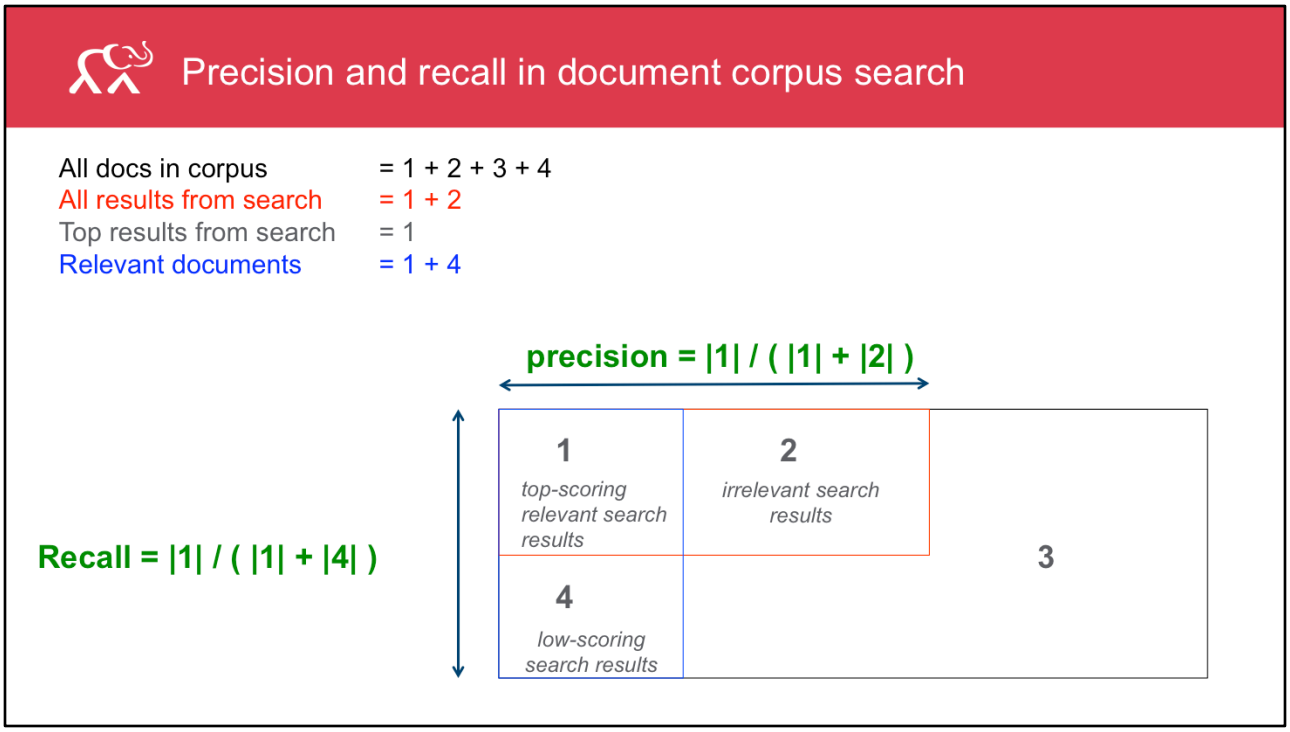
- **Distance**
  - Euclidean
  - Pearson
  - Spearman
- **Majority voting**
  - Problematic if class distribution is skewed
  - Could use distance as weight to compensate
  - Pick odd number for k to avoid ties


Demo: MyKNearestNeighbor.java
(10,15)
(10,10)
(3,3)
(1,1)
java MyKNearestNeighbor 1 10 10 euclidean
input: [10 10]
[1 1]
distance is 12.727922061357855
[3 3]
distance is 9.899494936611665
[10 15]
distance is 5.0
k nearest neighbors are:
[10 15]
java MyKNearestNeighbor 1 10 10 cosine
input: [10 10]
[1 1]
distance is 0.9999999999999998
[3 3]
distance is 1.0
[10 15]
distance is 0.9805806756909201
k nearest neighbors are:
[3 3]
Course Name
© 2014 MapR Technologies
13

Introduction to Recommendation

- Define recommendation terms and concepts
- Describe the K-nearest neighbors algorithm
- Evaluate a recommender model
- Discuss the Mahout implementations for recommendation


A search engine is actually a recommendation engine
Google
recommender system algorithms
Web
Images
Videos
Shopping
News
More
Search tools
About 527,000 results (0.28 seconds)
Scholarly articles for recommender system algorithms
... of dimensionality reduction in recommender system-a ... - Sarwar - Cited by 1045
... next generation of recommender systems: A survey of ... - Adomavicius - Cited by 5083
Introduction to recommender systems: Algorithms and ... - Konstan - Cited by 108
Recommender system - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Recommender_system - Wikipedia
Recommender systems are a useful alternative to search algorithms since they help users discover items they might not have found by themselves. Interestingly ...
Collaborative filtering - Information filtering system - Cold start
[PDF] Recommendation Systems - The Stanford University InfoL...
infolab.stanford.edu/~ullman/mmds/ch9.pdf
cient, and there are some new algorithms that have proven effective for .... algorithm that could beat its own recommendation system by 10%.1. The prize was ...
Recommender Systems :: Algorithms - Computer Science
www.cs.carleton.edu/cs_comps/0607/.../recommender/algorithms.html
Algorithms. We experimented with a number of different types of algorithms to build recommender systems. To learn more about them, please click on a link ...
Course Name
© 2014 MapR Technologies
15



Precision and recall are used to ***evaluate search results***

Precision = proportion of top-scoring results that are relevant

Recall = proportion of relevant results that are top-scoring

Precision and recall in collaborative filtering is challenging

- In collaborative filtering, a user history is crucial
  - So the rec evaluator will pick an item from the user history (which isn't necessarily the best rec)
  - Conversely, the rec evaluator should pick an item not in the user history (which would be penalized in the test)
- In boolean recommenders, there's no rating

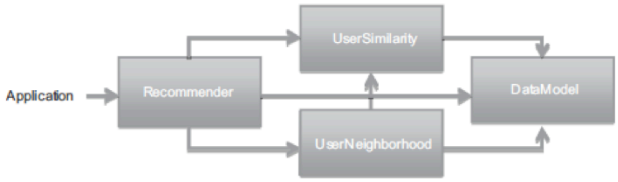
What to tune for better performance?

- **Model type**
  - User-based
  - Item-based
  - Content-based
- **Distance metric**
  - Euclidean
  - Tanimoto
  - Loglikelihood
- **Neighborhood size**
  - 2
  - 3
  - 200

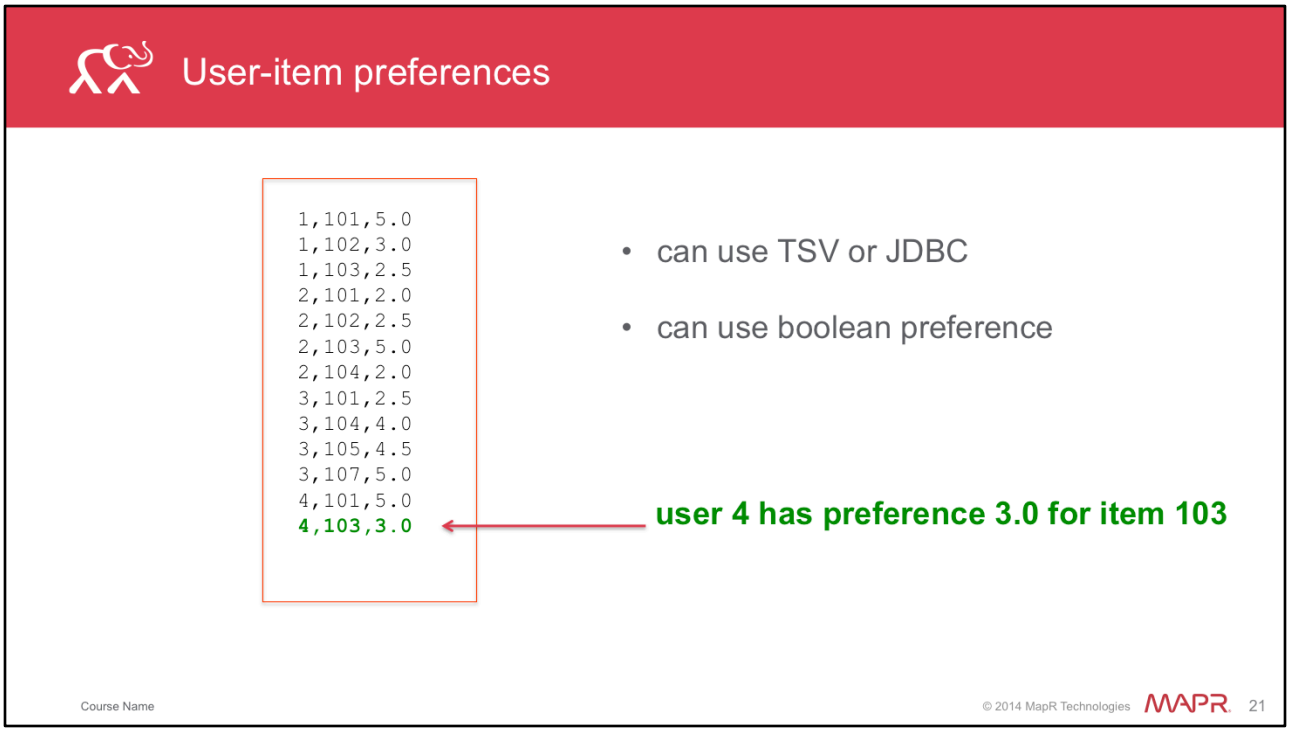
Introduction to Recommendation

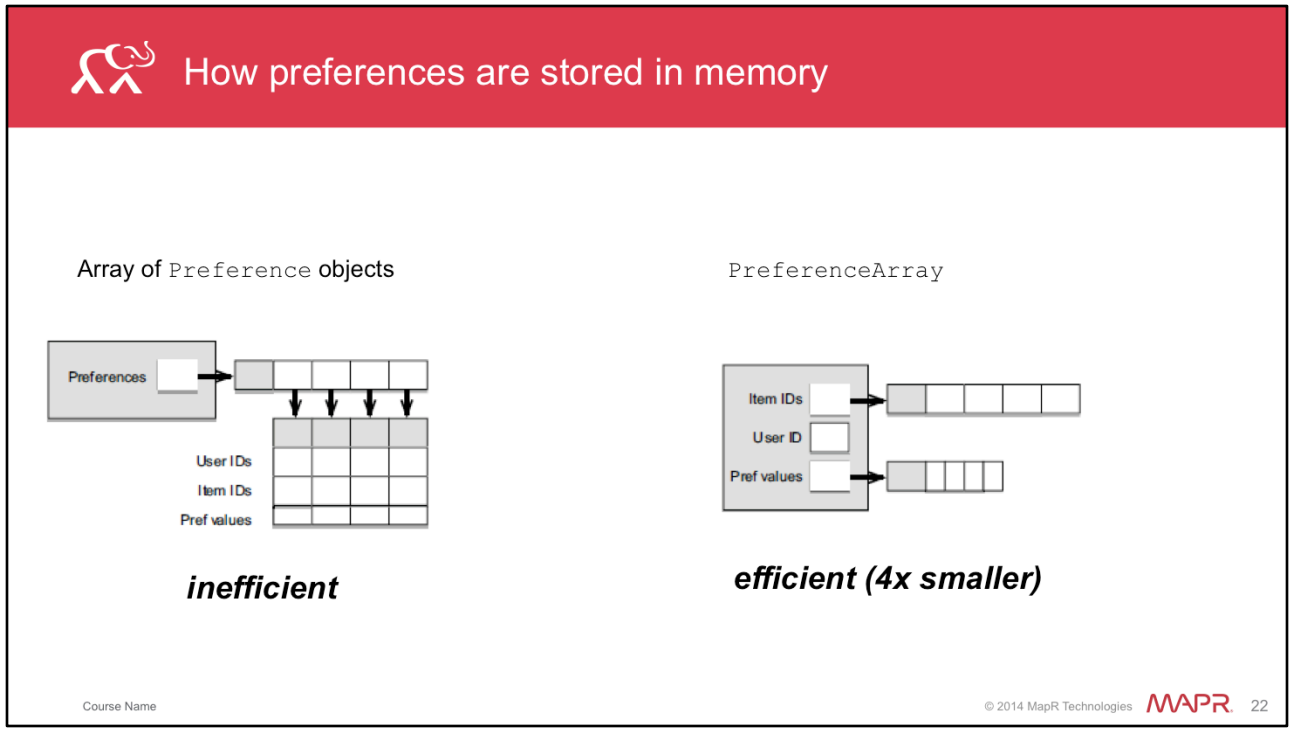
- Define recommendation terms and concepts
- Describe the K-nearest neighbors algorithm
- Evaluate a recommender model
- Discuss the Mahout implementations for recommendation

Mahout Recommender Architecture



- `Recommender`
  - Uses data and algorithms to make recommendation
- `UserSimilarity`
  - Configurable measure of how similar users are to each other
- `UserNeighborhood`
  - Configurable calculation of set of users who are similar to each other
- `DataModel`
  - Provides interface to store and retrieve users, items, and preferences


User-item preferences
1,101,5.0
1,102,3.0
1,103,2.5
2,101,2.0
2,102,2.5
2,103,5.0
2,104,2.0
3,101,2.5
3,104,4.0
3,105,4.5
3,107,5.0
4,101,5.0
4,103,3.0
• can use TSV or JDBC
• can use boolean preference
user 4 has preference 3.0 for item 103
Course Name
© 2014 MapR Technologies
21


How preferences are stored in memory
Array of Preference objects
Preferences
User IDs
Item IDs
Pref values
inefficient
PreferenceArray
Item IDs
User ID
Pref values
efficient (4x smaller)
Course Name
© 2014 MapR Technologies
22

Mahout recommender scale considerations

- Mahout recommenders are primarily ***memory-bound***
- Noisy data can be ***pruned*** to reduce memory requirement
- Data may be ***sampled*** to reduce memory requirement
- Use ***database backend*** to scale up


SUCCESS
IDEA
PRODUCTION
PRODUCT
RESULT
PROCCESS
TEAM
MEDIA
community
GROUP
IDEA
WORKSTATION
WEB
WORK GROUP
STORAGE
SERVER

- What are the two types of collaborative filtering?
- Which one is deductive: collaborative or content-based filtering?
- How are user preferences for items represented to Mahout?


MAPR 26