Cloud Computing
CS 15-319

Apache Mahout
Feb 13, 2012

Shannon Quinn
MapReduce Review

- Scalable programming model
- Map phase
- Shuffle
- Reduce phase
- MapReduce Implementations
  - Google
  - Hadoop
MapReduce Review

+ Scalable programming model
+ Map phase
+ Shuffle
+ Reduce phase
+ MapReduce Implementations
  + Google
  + Hadoop

This is our focus!
Apache Mahout

+ A scalable machine learning library
Apache Mahout

+ A scalable machine learning library
+ Built on Hadoop
Apache Mahout

+ A scalable machine learning library
+ Built on Hadoop
+ Philosophy of Mahout (and Hadoop by proxy)
What does Mahout do?
Recommendation

Here's a daily sample of items recommended for you. Click here to see all recommendations.

- **Programming in Python 3: A Gentle Introduction** (Paperback) by Mark Summerfield
  - Rating: 4.5 stars (11) $28.98
  - Fix this recommendation

- **Algorithm Design** (Hardcover) by Jon Kleinberg
  - Rating: 4.5 stars (23) $103.24
  - Fix this recommendation

- **Econometrics** (Hardcover) by Fumio Hayashi
  - Rating: 4.5 stars (19) $74.55
  - Fix this recommendation

- **Mathematical Statistics, Basic Principles and Applications** (Paperback) by Peter J. Bickel
  - Rating: 4.5 stars (61) $68.34
  - Fix this recommendation

---

**Critically-acclaimed Goofy Comedies**

- **Young Frankenstein**
  - Top Rated

- **Shaolin Soccer**
  - Most Popular

- **Airplane!**

---

Carnegie Mellon Qatar
Classification
Clustering
Other Mahout Algorithms

+ Dimensionality Reduction

+ Regression

+ Evolutionary Algorithms
Mahout

1. Recommendation

2. Classification

3. Clustering
Recommendation Overview

+ Help users find items they might like based on historical preferences

Here's a daily sample of items recommended for you. Click here to see all recommendations.

- Programming in Python 3: A C... (Paperback) by Mark Summerfield
  - Rating: 4.5 (11) - $26.98
  - Fix this recommendation

- Algorithm Design (Hardcover) by Jon Kleinberg
  - Rating: 4.5 (25) - $163.24
  - Fix this recommendation

- Econometrics (Hardcover) by Fumio Hayashi
  - Rating: 4.5 (19) - $74.55
  - Fix this recommendation

- Mathematical Statistics, Bas... (Paperback) by Peter J. Bickel
  - Rating: 4.5 (4) - $58.84
  - Fix this recommendation
Recommendation Overview

+ Mathematically
Recommendation Overview

*based on example by Sebastian Schelter

Carnegie Mellon Qatar
Recommendation Overview

*based on example by Sebastian Schelter
Recommendation Overview

Bob

5 1 4

? 2 5

4 3 2

*based on example by Sebastian Schelter
Recommendation Overview

Based on example by Sebastian Schelter
Recommendation in Mahout

+ 1st Map phase: process input

- (Alice, Matrix, 5) → <Alice, (Matrix, 5)>
- (Alice, Alien, 1) → <Alice, (Alien, 1)>
- (Alice, Inception, 4) → <Alice, (Inception, 4)>
- (Bob, Alien, 2) → <Bob, (Alien, 2)>

*based on example by Sebastian Schelter
Recommendation in Mahout

1st Map phase: process input

- (Alice, Matrix, 5)
- (Alice, Alien, 1)
- (Alice, Inception, 4)
- (Bob, Alien, 2)

1st Reduce phase: list by user

- <Alice, (Matrix, 5)>
- <Alice, (Alien, 1)>
- <Alice, (Inception, 4)>
- <Bob, (Alien, 2)>

- <Alice, (Matrix, 5); (Alien, 1); ...>
- <Bob, (Alien, 2); (Inception, 5); ...>

*based on example by Sebastian Schelter
Recommendation in Mahout

+ 2\textsuperscript{nd} Map phase: Emit co-occurred ratings

\begin{itemize}
  \item \texttt{<Alice, (Matrix, 5);(Alien, 1);…>}
  \item \texttt{<Bob, (Alien, 2);(Inception, 5);…>}
  \item \texttt{<Matrix;Alien, (5;1)>}
  \item \texttt{<Matrix;Inception, (5;4)>}
  \item \texttt{<Alien;Inception, (1;4)>}
  \item \texttt{<Alien;Inception, (2;5)>}
\end{itemize}

*based on example by Sebastian Schelter
Recommendation in Mahout

2nd Map phase: Emit co-occurred ratings

2nd Reduce phase: Compute similarities

*based on example by Sebastian Schelter
Mahout

1. Recommendation

2. Classification

3. Clustering
Classification Overview

+ Assigning data to discrete categories
Classification Overview

+ Assigning data to discrete categories
+ Train a model on labeled data

Spam    Not spam
Classification Overview

- Assigning data to discrete categories
- Train a model on labeled data
- Run the model on new, unlabeled data

Spam  Not spam
Naïve Bayes Example
Naïve Bayes Example

\[
\text{Prob (token | label)} =
\]
Naïve Bayes Example

Not spam
Naïve Bayes Example

President Obama’s Nobel Prize Speech

Not spam
Naïve Bayes Example

Spam
Naïve Bayes Example

Spam email content
Naïve Bayes Example
Naïve Bayes Example

“Order a trial Adobe chicken daily
EAB-List new summer savings, welcome!”
Naïve Bayes in Mahout

+ Complex!
Naïve Bayes in Mahout

+ Complex!
+ Training
  1. Read the features
Naïve Bayes in Mahout

+ Complex!
+ Training
  1. Read the features
  2. Calculate per-document statistics
Naïve Bayes in Mahout

+ Complex!
+ Training
  1. Read the features
  2. Calculate per-document statistics
  3. Normalize across categories
Naïve Bayes in Mahout

+ Complex!
+ Training
  1. Read the features
  2. Calculate per-document statistics
  3. Normalize across categories
  4. Calculate normalizing factor of each label
Naïve Bayes in Mahout

+ Complex!
+ Training
  1. Read the features
  2. Calculate per-document statistics
  3. Normalize across categories
  4. Calculate normalizing factor of each label
+ Testing
+ Classification
Other Classification Algorithms

+ Stochastic Gradient Descent
Other Classification Algorithms

+ Stochastic Gradient Descent
+ Support Vector Machines
Other Classification Algorithms

+ Stochastic Gradient Descent
+ Support Vector Machines
+ Random Forests
Mahout

1. Recommendation
2. Classification
3. Clustering
Clustering Overview

+ Grouping
  unstructured data
Clustering Overview

+ Grouping unstructured data
+ Small intra-cluster distance
Clustering Overview

+ Grouping unstructured data
+ Small intra-cluster distance
+ Large inter-cluster distance
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example
K-Means Clustering Example

Cats

Dogs
K-Means Clustering in Mahout

Figure from lecture 6: MapReduce
K-Means Clustering in Mahout

+ Assume: # clusters <<<< # points
K-Means Clustering in Mahout

+ Assume: # clusters <<< # points
K-Means Clustering in Mahout

+ Assume: # clusters <<< # points
K-Means Clustering in Mahout

+ Map phase: assign cluster IDs

\[(x_1, y_1), \text{ centroid1, centroid2, } \ldots \rightarrow \text{<cluster}_n, (x_1, y_1)\text{>}\]
\[(x_2, y_2), \text{ centroid1, centroid2, } \ldots \rightarrow \text{<cluster}_m, (x_2, y_2)\text{>}\]
\[(x_3, y_3), \text{ centroid1, centroid2, } \ldots \rightarrow \text{<cluster}_i, (x_3, y_3)\text{>}\]
\[(x_4, y_4), \text{ centroid1, centroid2, } \ldots \rightarrow \text{<cluster}_k, (x_4, y_4)\text{>}\]
K-Means Clustering in Mahout

+ Map phase: assign cluster IDs

(x1, y1), centroid1, centroid2, ...  →  <cluster_n, (x1, y1)>
(x2, y2), centroid1, centroid2, ...  →  <cluster_m, (x2, y2)>
(x3, y3), centroid1, centroid2, ...  →  <cluster_i, (x3, y3)>
(x4, y4), centroid1, centroid2, ...  →  <cluster_k, (x4, y4)>

+ Reduce phase: reset centroids

<cluster_n, (x1, y1)>  →  <cluster_n, centroid_n>
<cluster_m, (x2, y2)>  →  <cluster_m, centroid_m>
<cluster_i, (x3, y3)>  →  <cluster_i, centroid_i>
<cluster_k, (x4, y4)>  →  <cluster_k, centroid_k>
K-Means Clustering in Mahout

+ Important notes
  + --maxIter
  + --convergenceDelta
  + method
Other Clustering Algorithms

+ Latent Dirichlet Allocation
+ Topic models
Other Clustering Algorithms

- Latent Dirichlet Allocation
  - Topic models
- Fuzzy K-Means
  - Points are assigned multiple clusters
Other Clustering Algorithms

+ Latent Dirichlet Allocation
  + Topic models
+ Fuzzy K-Means
  + Points are assigned multiple clusters
+ Canopy clustering
  + Fast approximations of clusters
Other Clustering Algorithms

+ Latent Dirichlet Allocation
  + Topic models
+ Fuzzy K-Means
  + Points are assigned multiple clusters
+ Canopy clustering
  + Fast approximations of clusters
+ Spectral clustering
  + Treat points as a graph
Other Clustering Algorithms

+ Latent Dirichlet Allocation
  + Topic models

+ Fuzzy K-Means
  + Points are assigned multiple clusters

+ Canopy clustering
  + Fast approximations of clusters

+ Spectral clustering
  + Treat points as a graph

K-Means & Eigencuts
Mahout in Summary
Mahout in Summary

+ Scalable library
Mahout in Summary

+ Scalable library
+ Three primary areas of focus
Mahout in Summary

+ Scalable library
+ Three primary areas of focus
+ Other algorithms
Mahout in Summary

- Scalable library
- Three primary areas of focus
- Other algorithms
- All in your friendly neighborhood MapReduce
Mahout in Summary

+ Scalable library
+ Three primary areas of focus
+ Other algorithms
+ All in your friendly neighborhood

MapReduce

http://mahout.apache.org/
Thank you!

Why, yes! This IS relevant to my interests!