



Introduction to Scalable Machine Learning with Apache Mahout

Grant Ingersoll

February 15, 2010

Thinking Lucene ▼ Think Lucid.

Introduction

▼ You

- ▼ Machine learning experience?
- ▼ Business Intelligence?
- ▼ Natural Lang. Processing?
- ▼ Apache Hadoop?

▼ Me

- ▼ Co-founder Apache Mahout
- ▼ Apache Lucene/Solr committer
- ▼ Co-founder Lucid Imagination

- ▼ **What is Machine Learning?**
- ▼ **ML Use Cases**
- ▼ **What is Mahout?**
- ▼ **What can I do with it right now?**
- ▼ **Where's Mahout headed?**

What is Machine Learning?



McCain the show horse: Way off track

Seattle Post Intelligencer - 36 minutes ago

By JOEL CONNELLY ABOARD THE now-jettisoned "Straight Talk" with reporters about go-to heroes in history, none more than The Candidates scramble to prepare for debate amid bailout crisis McCain Decides to Participate in Debate New York Times BBC News - Voice of America - Washington Post - AFP all 5,183 news articles »

WaMu's Bank Split From Holding Company, Spal

Bloomberg - 46 minutes ago

By Linda Shen Sept. 26 (Bloomberg) -- Washington Mutual Inc.'s branches and deposits when JPMorgan Chase & Co. Video: Wall Street watches Washington Reuters Video Update: JPMorgan takes over WaMu after snapping up assets Los Angeles Times - CNNMoney.com - Wall Street Journal - Ma all 2,791 news articles »

Russia warship heads to Africa after pirate attac

The Associated Press - 1 hour ago

MOSCOW (AP) - A Russian warship on Friday rushed to intercede and a hoard of ammunition that was seized by pirates off the Horn of Africa, heightening fears about surging piracy ... Russian Navy ship sent to combat pirates ABC Online Somali pirates grab Ukrainian ship loaded with tanks Reuters International Herald Tribune - Voice of America - CNN - Bloomberg all 561 news articles »



[Introduction to Data Mining](#)

by Pang-Ning Tan

★★★★★ (10) \$87.97

mazon.com



Really it's...

- ▼ **“Machine Learning is programming computers to optimize a performance criterion using example data or past experience”**
 - ▼ *Intro. To Machine Learning* by E. Alpaydin
- ▼ **Subset of Artificial Intelligence**
- ▼ **Lots of related fields:**
 - ▼ Information Retrieval
 - ▼ Stats
 - ▼ Biology
 - ▼ Linear algebra
 - ▼ Many more

Common Use Cases

- ▼ Recommend friends/dates/products
- ▼ Classify content into predefined groups
- ▼ Find similar content based on object properties
- ▼ Find associations/patterns in actions/behaviors
- ▼ Identify key topics in large collections of text
- ▼ Detect anomalies in machine output
- ▼ Ranking search results
- ▼ Others?

Useful Terminology

- ▼ **Vectors/Matrices**
 - ▼ Weights
 - ▼ Sparse
 - ▼ Dense
 - ▼ Norms
- ▼ **Features**
 - ▼ Feature reduction
- ▼ **Occurrences and Cooccurrences**

Getting Started with ML

- ▼ **Get your data**
- ▼ **Decide on your features per your algorithm**
- ▼ **Prep the data**
 - ▼ Different approaches for different algorithms
- ▼ **Run your algorithm(s)**
 - ▼ Lather, rinse, repeat
- ▼ **Validate your results**
 - ▼ Smell test, A/B testing, more formal methods

Apache Mahout

<http://dictionary.reference.com/browse/mahout>

▼ An Apache Software Foundation project to create scalable machine learning libraries under the Apache Software License

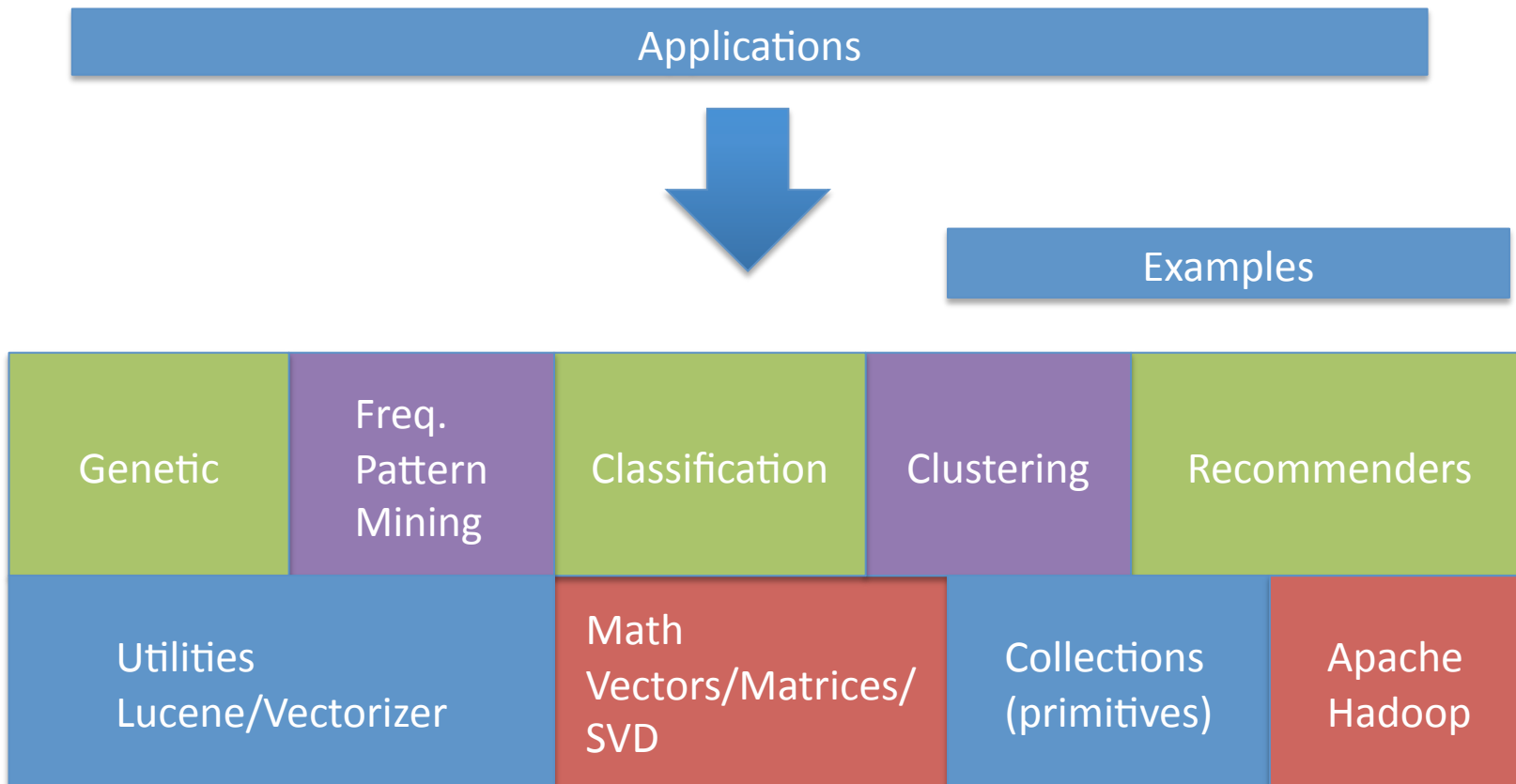
▼ Why Mahout?

▼ Many Open Source ML libraries either:

- Lack Community
- Lack Documentation and Examples
- Lack Scalability
- Lack the Apache License ;-)
- Or are research-oriented



Focus: Machine Learning



See <http://cwiki.apache.org/confluence/display/MAHOUT/Algorithms>

Focus: Scalable



- ▼ **Goal: Be as fast and efficient as the possible given the intrinsic design of the algorithm**
 - ▼ Some algorithms won't scale to massive machine clusters
 - ▼ Others fit logically on a Map Reduce framework like Apache Hadoop
 - ▼ Still others will need other distributed programming models
 - ▼ Be pragmatic
- ▼ **Most Mahout implementations are Map Reduce enabled**
- ▼ **Work in Progress**

Prepare Data from Raw content

- ▼ **Data Sources:**
 - ▼ Lucene integration
 - bin/mahout lucenevector ...
 - ▼ Document Vectorizer
 - bin/mahout seqdirectory ...
 - bin/mahout seq2sparse ...
 - ▼ Programmatically
 - See the Utils module in Mahout
 - ▼ Database
 - ▼ File system

Recommendations

- ▶ Extensive framework for collaborative filtering
- ▶ Recommenders
 - ▶ User based
 - ▶ Item based
- ▶ Online and Offline support
 - ▶ Offline can utilize Hadoop
- ▶ Many different Similarity measures
 - ▶ Cosine, LLR, Tanimoto, Pearson, others

Customers Who Bought This Item Also Bought



Clustering

Document level

- ▶ Group documents based on a notion of similarity
- ▶ K-Means, Fuzzy K-Means, Dirichlet, Canopy, Mean-Shift
- ▶ Distance Measures
 - Manhattan, Euclidean, other

McCain the show horse: Way off track

Seattle Post Intelligencer - **36 minutes ago**
 By JOEL CONNELLY ABOARD THE now-jettisoned "Straight Talk Express," Sen. John McCain loved to talk with reporters about go-to heroes in history, none more than Theodore Roosevelt and Winston Churchill.
[Candidates scramble to prepare for debate amid bailout crisis](#) CNN International
[McCain Decides to Participate in Debate](#) New York Times
[BBC News - Voice of America - Washington Post - AFP](#)
[all 5,183 news articles »](#)



BBC News

WaMu's Bank Split From Holding Company, Sparing FDIC (Update1)

Bloomberg - **46 minutes ago**
 By Linda Shen Sept. 26 (Bloomberg) -- Washington Mutual Inc.'s holding company was detached from its branches and deposits when JPMorgan Chase & Co.
[Video: Wall Street watches Washington](#) ReutersVideo
[Update: JPMorgan takes over WaMu after snapping up assets](#) Bizjournals.com
[Los Angeles Times - CNNMoney.com - Wall Street Journal - MarketWatch](#)
[all 2,791 news articles »](#)



Telegraph.co.uk

Russia warship heads to Africa after pirate attack

The Associated Press - **1 hour ago**
 MOSCOW (AP) - A Russian warship on Friday rushed to intercept a Ukrainian vessel carrying 33 battle tanks and a hoard of ammunition that was seized by pirates off the Horn of Africa - a bold hijacking that again heightened fears about surging piracy ...
[Russian Navy ship sent to combat pirates](#) ABC Online
[Somali pirates grab Ukrainian ship loaded with tanks](#) Reuters
[International Herald Tribune - Voice of America - CNN - Bloomberg](#)
[all 561 news articles »](#)



The Southern Ledger

Topic Modeling

- ▶ Cluster words across documents to identify topics
- ▶ Latent Dirichlet Allocation

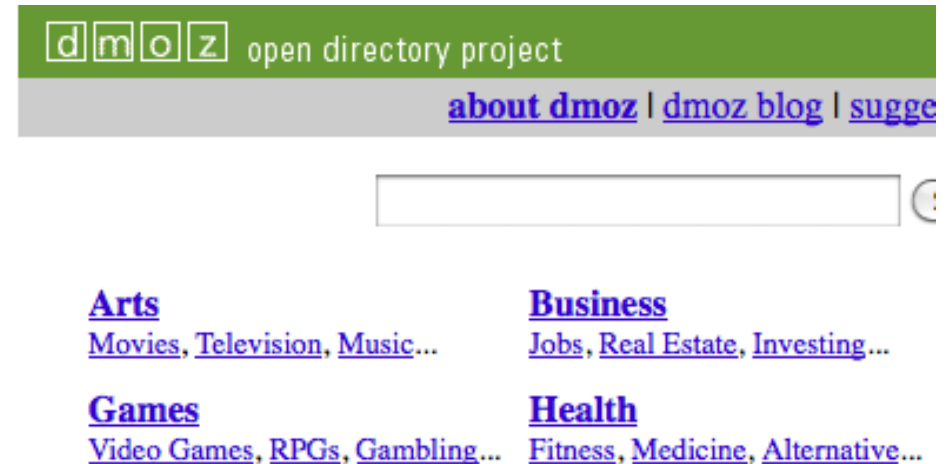
Categorization

Place new items into predefined categories:

- ▼ Sports, politics, entertainment

Mahout has several implementations

- ▼ Naïve Bayes
- ▼ Complementary Naïve Bayes
- ▼ Decision Forests



Freq. Pattern Mining

- ▼ Identify frequently co-occurrent items
- ▼ Useful for:
 - ▼ Query Recommendations
 - Apple -> iPhone, orange, OS X
 - ▼ Related product placement
 - “Beer and Diapers”

"apple"

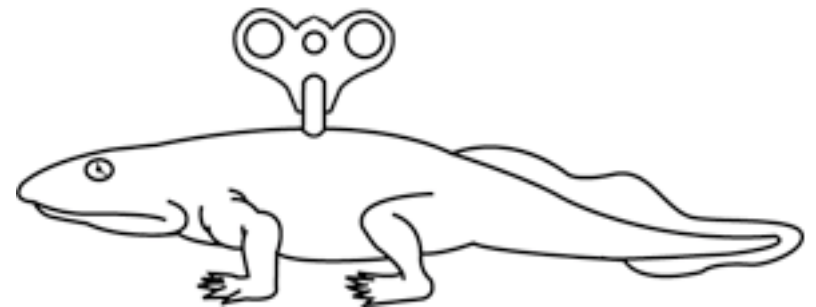
Related Searches: [apple iphone](#), [iphone](#), [ipod](#).

Select Results from All Departments

<http://www.amazon.com>



- ▼ Map-Reduce ready fitness functions for genetic programming
- ▼ Integration with Watchmaker
 - ▼ <http://watchmaker.uncommons.org/index.php>
- ▼ Problems solved:
 - ▼ Traveling salesman
 - ▼ Class discovery
 - ▼ Many others



How To: Recommenders

- ▼ **Data:**
 - ▼ Users (abstract)
 - ▼ Items (abstract)
 - ▼ Ratings (optional)
- ▼ **Load the data model**
- ▼ **Ask for Recommendations:**
 - ▼ User-User
 - ▼ Item-Item

Ugly Demo I

In other words: the reason why I work on servers, not UIs!

- ▼ Group Lens Data: <http://www.grouplens.org>
- ▼ <http://lucene.apache.org/mahout/taste.html#demo>
- ▼ <http://localhost:8080/RecommenderServlet?userID=1&debug=true>
- ▼

How to: Command Line

- ▼ **Most algorithms have a Driver program**
 - ▼ Shell script in `$MAHOUT_HOME/bin` helps with most tasks
- ▼ **Prepare the Data**
 - ▼ Different algorithms require different setup
- ▼ **Run the algorithm**
 - ▼ Single Node
 - ▼ Hadoop
- ▼ **Print out the results**
 - ▼ Several helper classes:
 - LDAPrintTopics, ClusterDumper, etc.

Ugly Demo II - Prep

▼ Data Set: Reuters

▼ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

▼ Convert to Text via

<http://www.lucenebootcamp.com/lucene-boot-camp-preclass-training/>

▼ Convert to Sequence File:

▼ `bin/mahout seqdirectory --input <PATH> --output <PATH> --charset UTF-8`

▼ Convert to Sparse Vector:

▼ `bin/mahout seq2sparse --input <PATH>/content/reuters/seqfiles/ --norm 2 --weight TF --output <PATH>/content/reuters/seqfiles-TF/ --minDF 5 --maxDFPercent 90`

Ugly Demo II: Topic Modeling

▼ Latent Dirichlet Allocation

- ▼ `./mahout lda --input <PATH>/content/reuters/seqfiles-TF/vectors/ --output <PATH>/content/reuters/seqfiles-TF/lda-output --numWords 34000 --numTopics 10`
- ▼ `./mahout org.apache.mahout.clustering.lda.LDAPrintTopics --input <PATH>/content/reuters/seqfiles-TF/lda-output/state-19 --dict <PATH>/content/reuters/seqfiles-TF/dictionary.file-0 --words 10 --output <PATH>/content/reuters/seqfiles-TF/lda-output/topics --dictionaryType sequencefile`
- ▼ Good feature reduction (stopword removal) required

Ugly Demo III: Clustering

▼ K-Means

- ▼ Same Prep as UD II, except use TFIDF weight
- ▼ `./mahout kmeans --input <PATH>/content/reuters/seqfiles-TFIDF/vectors/part-00000 --k 15 --output <PATH>/content/reuters/seqfiles-TFIDF/output-kmeans --clusters <PATH>/content/reuters/seqfiles-TFIDF/output-kmeans/clusters`
- ▼ Print out the clusters: `./mahout clusterdump --seqFileDir <PATH>/content/reuters/seqfiles-TFIDF/output-kmeans/clusters-15/ --pointsDir <PATH>/content/reuters/seqfiles-TFIDF/output-kmeans/points/ --dictionary <PATH>/content/reuters/seqfiles-TFIDF/dictionary.file-0 --dictionaryType sequencefile --substring 20`

Ugly Demo IV: Frequent Pattern Mining

- ▼ Data: <http://fimi.cs.helsinki.fi/data/>
- ▼ `./mahout fpg -i <PATH>/content/freqitemset/accidents.dat -o patterns -k 50 -method mapreduce -g 10 -regex [\]`
- ▼ `./mahout seqdump --seqFile patterns/fpgrowth/part-r-00000`

What's Next?

- ▼ 0.3 release very soon
- ▼ Parallel Singular Value Decomposition (Lanczos)
- ▼ Stabilize API's for 1.0 release
- ▼ Benchmarking
- ▼ Google Summer of Code?
- ▼ More Algorithms
- ▼ <http://cwiki.apache.org/MAHOUT/howtocontribute.html>

- ▼ **Slides and Full Details of Demos at:**
 - ▼ <http://lucene.grantingersoll.com/2010/02/13/intro-to-mahout-slides-and-demo-examples/>
- ▼ **More Examples in Mahout SVN in the examples directory**

Resources

- ▼ <http://lucene.apache.org/mahout>
- ▼ <http://cwiki.apache.org/MAHOUT>
- ▼ mahout-{user|dev}@lucene.apache.org
- ▼ <http://svn.apache.org/repos/asf/lucene/mahout/trunk>
- ▼ <http://hadoop.apache.org>

Resources

- ▼ **“Mahout in Action” by Owen and Anil**
- ▼ **“Introducing Apache Mahout”**
 - ▼ <http://www.ibm.com/developerworks/java/library/j-mahout/>
- ▼ **“Programming Collective Intelligence” by Toby Segaran**
- ▼ **“Data Mining - Practical Machine Learning Tools and Techniques” by Ian H. Witten and Eibe Frank**

References

- ▼ HAL: <http://en.wikipedia.org/wiki/File:Hal-9000.jpg>
- ▼ Terminator: <http://en.wikipedia.org/wiki/File:Terminator1984movieposter.jpg>
- ▼ Matrix: http://en.wikipedia.org/wiki/File:The_Matrix_Poster.jpg
- ▼ Google News: <http://news.google.com>
- ▼ Amazon.com: <http://www.amazon.com>
- ▼ Facebook: <http://www.facebook.com>
- ▼ Mahout: <http://lucene.apache.org/mahout>
- ▼ Beer and Diapers: <http://www.flickr.com/photos/baubcat/2484459070/>
 - ▼ http://www.theregister.co.uk/2006/08/15/beer_diapers/
- ▼ DMOZ: <http://www.dmoz.org>