Apache Mahout
Large Scale Machine Learning

Speaker: Isabel Drost
Agenda

• Motivation.

• What is machine learning?

• Introduction to Mahout.
January 3, 2006 by Matt Callow
http://www.flickr.com/photos/blackcustard/81680010
Today: Read news papers, Blogs, Twitter, RSS feed.

Wish: Aggregate sources and track emerging topics.
Today: IMDB, zitty, movie review pages, twitter, blogs, ask friends.

Wish: Reviews, sentiment detection, recommendations.
Machine learning – what's that?
Archimedes taking a Warm Bath
Archimedes model of nature

\[
\frac{\text{Density of Object}}{\text{Density of Fluid}} = .
\]

\[
\frac{\text{Weight}}{\text{Weight} - \text{Apparent immersed weight}}
\]
An SVM's model of nature
From data to model.
Gather data
Dear Mary,

Yesterday, though short, was a fine day. And let me say just here, don't write because you have nothing to write. Because if it is a Communicate. Write, if it is. I believe all of you lives.

December 31, 2005, birdfarm
http://www.flickr.com/photos/birdfarm/800522248/
If we looked at two words only:
Reality

- There are a few more words in mails.
- Use all relevant features/ signals available.
  - Words.
  - Header fields.
  - Characteristics of attachments.
  - ...
- Usually pipeline of feature extractors.
- UIMA: Apache project focussed on that task.
Gather data
Extract signals
Algor. choice
Parameters
Train model
Apply model
Use results
Training the model

- No single best algorithm for all tasks.
- No single best parameter setting per algorithm.
- Evaluate constantly.
Gather data -> Extract signals

Algo. choice -> Parameters -> Train model

Apply model -> Use results
Problem: “Nature changes”
Challenges.
Challenges

- Amount of data grows exponentially.
  - User generated content on the web.
  - Sensor data.
  - Customer logs.
- Index and search the data.
- Build models and generalize from raw data.
- How do non-Googlers deal with that?
Once upon a time

- A Java library.
- Index with easy to use API.
Once upon a time

A library alone is not enough.
Once upon a time

- Lucene: Umbrella project for search at Apache.
Once upon a time

- nutch needed a way to scale to the web.
- Functional languages support map/reduce.
- 2004 - MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean and Sanjay Ghemawat.
- 2004 - Initial versions of DFS and Map-Reduce by Doug Cutting & Mike Cafarella
- December 2005 - Nutch ported to framework, 20 nodes.
- January 2006 - Doug Cutting joins Yahoo!
- February 2006 - Apache Hadoop project hived off.
- March 2006 - Formation of the Yahoo! Hadoop team
- April 2007 - Research clusters - 2 clusters of 1000 nodes
Once upon a time
Once upon a time
Once upon a time

And many more inside and outside Apache.
Where does Mahout fit in?

- Amount of data to process is growing.
- Idea: Scale and go parallel.
Where does Mahout fit in?

- Amount of data to process is growing.
- Idea: Scale and go parallel.
Where does Mahout fit in?

- Amount of data to process is growing.
- Idea: Scale and go parallel.
Where does Mahout fit in?
What does Mahout have to offer.
Discover groups of items

- Group items by similarity.

- Examples:
  - Group news articles by topic.
  - Find developers with similar interests.
  - Discovery of groups of related search results.
Discover groups of similar items

- Canopy.
- k-Means.
- Fuzzy k-Means.
- Dirichlet based.
- Others upcoming.
Discover groups of similar items

- Example: Synthetic Control
  - http://archive.ics.uci.edu/ml/datasets/Synthetic+Control
  - Example Job: `<MAHOUT_HOME>/examples`
  - Outputs clusters

- Download the distribution.
- Run the example.
- Have a closer look at the examples.
Assign items to defined categories.

- Given pre-defined categories, assign items to it.

- Examples:
  - Spam mail classification.
  - Discovery of images depicting humans.
Assign items to defined categories.

- Naïve Bayes.
- Complementary naïve bayes.
- Winnow/Perceptron.
- Others upcoming.
Assign items to defined categories.

- Examples based on “standard” datasets:
  - 20 Newsgroups
    - http://cwiki.apache.org/confluence/display/MAHOUT/
  - Wikipedia
    - http://cwiki.apache.org/confluence/display/MAHOUT/
Evolutionary algorithms

• Traveling Salesman
  – http://cwiki.apache.org/confluence/display/MAHOUT/Traveling+Salesman

• Classification rule discovery
  – http://cwiki.apache.org/confluence/display/MAHOUT/Class+Discovery
Collaborative filtering

- Recommend items to users.

- Examples:
  - Find movies I might want to watch.
  - Find books related to the book I am buying.
Collaborative filtering
Collaborative filtering
Recommendation mining.

- Mahout with more Taste.
- Mature Java library.
- Java-based, web service / HTTP bindings.
- Batch mode based on EC2 and Hadoop.
What next?

- More algorithms.
- More examples.
What next?

- 2nd Summer of code.
- Four mentors.
- Three students.
- Two returning students.

Robin Anil: *Online Classification and Frequent Pattern Mining using Map-Reduce.*

David Hall: *Distributed Latent Dirichlet Allocation.*

AbdelHakim: *Implement parallel Random/ Regression Forest.*
Why go for Apache Mahout?
Jumpstart your project with proven code.
Discuss ideas and problems online.

November 16, 2005 [phil h]
http://www.flickr.com/photos/hi-phi/64055296
Become part of the community.
Release: 0.1
Big Thanks to those who made this possible!

October 22, 2008 by e_calamar
http://www.flickr.com/photos/e_calamar/2964991182/
mahout-user@lucene.apache.org
mahout-dev@lucene.apache.apache.org

Interest in machine learning.
Interesting problems.
Hadoop proficiency.

Bug reports, patches, features.
Documentation, code, examples.
Ahem – I do not own a big cluster...
• Mahout runs on top of Amazon EMR.
• Run Mahout on your Hadoop cluster on EC2.
• Committers do get free credits for EC2 ;)
• Set up your own Hadoop cluster.
June, 25th 2009: Hadoop* Get Together in Berlin

- Torsten Curdt: “Data Legacy - the challenges of an evolving data warehouse.”
- Christoph M. Friedrich: “SCAIView - Lucene for Life Science Knowledge Discovery”
- Uri Boness, Bram Smeets: “Solr in production.”

newthinking store
Tucholskystr. 48

September, 29th 2009: Hadoop* Get Together in Berlin featuring a talk on UIMA by Thilo Götz.

* UIMA, Hbase, Lucene, Solr, katta, Mahout, CouchDB, pig, Hive, Cassandra, Cascading, JAQL, ... talks welcome as well.
mahout-user@lucene.apache.org
mahout-dev@lucene.apache.apache.org

Interest in machine learning.
Interesting problems.
Hadoop proficiency.

Bug reports, patches, features.
Documentation, code, examples.

July 9, 2006 by trackrecord
http://www.flickr.com/photos/trackrecord/185514449