Apache Mahout

Scalable Machine Learning

Speaker: Isabel Drost
Isabel Drost

Nighttime:
Co-Founder Apache Mahout.
Organizer of Berlin Hadoop Get Together.

Daytime:
Software developer @ Berlin
Agenda

- Motivation.

- Introduction to Mahout.
News aggregation

Today: Read news papers, Blogs, Twitter, RSS feed.

Wish: Aggregate sources and track emerging topics.
Go to cinema

Today: IMDB, zitty, movie review pages, twitter, blogs, ask friends.

Wish: Reviews, sentiment detection, recommendations.

March 22, 2008 by Crystian Cruz
http://www.flickr.com/photos/crystiancruz/2353895708
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

- Margin
- Separating hyperplane
- Class +1: \( w^*x + b > 1 \)
- Class -1: \( w^*x + b < -1 \)

\[ r_k \]
The challenge
- Large amounts of data.
- Structured and unstructured data.
- Diverse tasks.
Mission

Provide scalable data mining algorithms.
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.
Identify dominant topics

- Given a dataset of texts, identify main topics.

  Algorithms: Parallel LDA

- Examples:
  - Dominant topics in set of mails.
  - Identify news message categories.
Assign items to defined categories.

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

- Examples:
  - Spam mail classification.
  - Discovery of images depicting humans.
Recommendation mining.

- Recommend items to users.

- Examples:
  - Find books related to the book I am buying.
  - Find movies I might like.
Recommending places
Recommending people
Recommendation mining.

- Integrated Taste.
- Mature Java library.
- Java-based, web service / HTTP bindings.
- Batch mode based on EC2 and Hadoop.
Frequent pattern mining

- Given groups of items, find commonly co-occurring items.

- Examples:
  - In shopping carts find items bought together.
  - In query logs find queries issued in one session.
Upcoming

- More algorithms.
- Optimization of existing implementations.
- More examples.
- Release 0.2
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.
mahout-user@lucene.apache.org
mahout-dev@lucene.apache.apache.org

Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

Bug reports, patches, features.
Documentation, code, examples.
Dec., 16th 2009: Hadoop* Get Together in Berlin

- Richard Hutton (nugg.ad): “Moving from five days to one hour.”
- Jörg Möllenkamp (Sun): “Jörg Möllenkamp (Sun): "Hadoop on Sun."
- Nikolaus Pohle (nurago): "M/R for MR - Online Market Research powered by Apache Hadoop. Enable consultants to analyze online behavior for audience segmentation, advertising effects and usage patterns."

http://upcoming.yahoo.com/event/4842528/

* UIMA, Hbase, Lucene, Solr, katta, Mahout, CouchDB, pig, Hive, Cassandra, Cascading, JAQL, ... talks welcome as well.
Lucene/ Mahout booth next door

Isabel Drost
Simon Willnauer
mahout-user@lucene.apache.org
mahout-dev@lucene.apache.apache.org

Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

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

July 9, 2006 by trackrecord
http://www.flickr.com/photos/trackrecord/185514449
No, you are not the only one... Many a sleepless night spent on it... :-)

I usually try to refer to it as Lucene Java, but old habits die hard and often times I just call it Lucene. I think the name has a good brand at this point and is very strongly associated with the Java library. I seem to recall when they were forming the TLP, that the original proposal was search.a.o, but then changed b/c the ASF didn't like generic names (or at least that is how I recall it.) And, of course, with Hadoop and the potential for Tika/Lius, it isn't just search anymore. I have often thought about an Apache "Text" project, that could eventually hold a whole family of text based tools like Lucene, Tika, Hadoop, Solr, etc. plus things like part of speech taggers, clustering/classification algorithms, UIMA, etc. all under one roof. But that is just my two cents and I don't know if it fits with what other people have in mind. There are a lot of OSS tools out there for these things, but none bring together a whole suite under a brand like Apache.

-Grant
Going parallel: k-Means
Until stable:
Data intensive.
Output: Cluster assignment.
Pre-Compute centers.

Done in Map.
Data intensive.
Output: Cluster assignment.
Pre-Compute centers.

Done in Map.  

Until stable.

Done in Reduce.
Make searching the web easier

Find job postings on the internet.

Group by categories.

Automatically extract: Place, title, date...