Apache Mahout
Making data analysis easy
Isabel Drost

Nighttime:

Co-Founder, committer Apache Mahout. Organiser of Berlin Hadoop Get Together.

Daytime:

Hello codebits!
Hello codebits!
Hello codebits!
Hello codebits!
Hello codebits!
Hello codebits!

Machine learning background?
Hello codebits!
Agenda

- Data Mining/ Machine Learning?
- Why is scaling hard?
- Introducing Apache Hadoop.
- Going beyond simple statistics.
Machine learning – what's that?
Data Mining Applications

- Marketing.
- Surveillance.
- Fraud Detection.
- Scientific Discovery.
- Discover items usually purchased together.

= Extracting patterns from data.
Machine Learning Applications

- E-Mail spam classification.
- News-topic discovery.
- Building recommender systems.

= Extracting prediction models from data.
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
- Class +1: \( w^*x + b > 1 \)
- Class -1: \( w^*x + b < -1 \)
- Separating hyperplane: \( w^*x + b = 0 \)
The challenge
Mission

Provide scalable data mining algorithms.
Massive data as in:

- Cannot be stored on single machine.
- Takes too long to process in serial.

Idea: Use multiple machines.
Challenges when scaling out.
Single machines tend to fail:
Hard disk.
Power supply.
...
More machines – increased failure probability.
Typical developer

- Has never dealt with large (petabytes) amount of data.
- Has no thorough understanding of parallel programming.
- Has no time to make software production ready.

September 10, 2007 by .sandemb
http://www.flickr.com/photos/daphid/1354523220/
Go away or I will replace you with a very small shell script.

http://www.flickr.com/photos/cspowers/282944734/ by cspowers on October 29, 2006
Easy distributed programming.
Well known in industry and research.
Scales well beyond 1000 nodes.
Hadoop assumptions
Assumptions:
Moving computation is cheap.
Moving data is expensive.

Ideas:
Move computation to data.
Write software that is easy to distribute.
Assumptions:

Systems run on spinning hard disks. Disk seek >> disk scan.

Ideas:

Improve support for large files. File system API makes scanning easy.
HDFS building blocks
Data

Name Node  Data Nodes

HDFS

(Graphics: Thanks to Thilo.)
Anatomy of a file write

Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly
Anatomy of a file write

Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly
Anatomy of a file write

HDFS client

Client node → Name Node

Create file → Close file

Write packet → Data Node

Ack packet

Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly
Anatomy of a file write

Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly
Map/Reduce by example
$ more data/feeds.opml | grep -o "http://[0-9A-Za-z\-_\.]*" | sort | uniq --count | sort | tail

  3 http://agbs.kyb.tuebingen.mpg.de
  3 http://irgupf.com
  3 http://jeffsutherland.com
  4 http://ml.typepad.com
  4 http://weblogs.java.net
  4 http://www.gridvm.org
  4 http://yaroslavvb.blogspot.com
  5 http://feeds.feedburner.com
  6 http://blogsearch.google.com
 10 http://arxiv.org
pattern="http://[0-9A-Za-z\-\_\.]*"
grep -o "$pattern" feeds.opml | sort | uniq --count
grep -o "pattern" feeds.opml | sort | uniq --count
M A P
| SHUFFLE
| R E D U C E
Jobs

Data

Scheduling

Name Node,
Job tracker

Processing

Data Nodes,
Task trackers

HDFS

Result

(Graphics: Thanks to Thilo.)
Mission

Provide scalable data mining algorithms.
HowTo: From data to information.
Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz.

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppe den Überwachungswahn" in Berlin teil.

Die Demonstranten zogen sich unter anderem gegen die Verteilung der Gesundheitskarte und den Verlust von Selbstbestimmung.
From data to information.

Collect data and define your learning problem.

- Data preparation.
- Training a prediction model.
- Checking the performance of your model.
Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 150 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungstaat.

Rund 7500 Demonstranten nahmen am Protestauftakt unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungsstaat" in Berlin teil.
Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 150 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungstaat.

Suche nach "Bürgerrechte im Netz"

Rund 7000 Demonstranten nahmen an dem Protest unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungsstaat" in Berlin teil.

Die Demonstranten riefen sich unter anderem zum Datenschutz auf.
Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 150 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungstaats.

Remove noise.
Convert text to vectors.
From texts to vectors
If we looked at two words only:
Aaron

Zuse
Binary bag of words

• Imagine a n-dimensional space.
• Each dimension = one possible word in texts.
• Entry in vector is one, if word occurs in text.

\[ b_{i,j} = \begin{cases} 
1 & \forall x_i \in d_j \\
0 & \text{else} 
\end{cases} \]

• Problem:
  • Number of word occurrences not accounted for.
Term Frequency

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector equal to the words frequency.

\[ b_{i,j} = n_{i,j} \]

- Problem:
  - Common words dominate vectors.
TF with stop wording

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the words frequency.

\[ b_{i,j} = n_{i,j} \]

- Problem:
  - Common and uncommon words with same weight.
TF-IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.

\[ b_{i,j} = n_{i,j} \times \log \left( \frac{|D|}{|\{d : t_i \in d\}|} \right) \]

- Problem:
  - Long texts get larger values.
Normalized TF-IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.
- Normalize vectors.

$$b_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \left( \frac{|D|}{|\{d : t_i \in d\}|} \right)$$

- Problem:
  - Additional domain knowledge ignored.
Reality

- There are a few more words in news.
- Use all relevant features/ signals available.
  - Words.
  - Header fields.
  - Characteristics of publishing url.
  - ...
- Usually pipeline of feature extractors.
From data to information.

- Collect data and define your learning problem.
- Data preparation.
  - Training a prediction model.
  - Checking the performance of your model.
Step 2: Similarity
Euclidian
Euclidian
Euclidian

Cosine
Step 3: Clustering
Reality

- Seed selection.
- Choice of initial $k$.
- Continuous updates.
- Regular addition of clusters.
Discover groups of similar items

- Canopy.
- k-Means.
- Fuzzy k-Means.
- Dirichlet based.
- Spectral clustering.
- Others upcoming.
From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.

  - Checking the performance of your model.
Evaluation

- Compare against gold standard.
- Use quality measures.
- Manual inspection.
From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.
What else does Mahout have to offer.
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.
Assign items to defined categories.

- Naïve Bayes.
- Complementary naïve bayes.
- Random forests.
- Logistic regression.
- HMM for sequences.
Recommendation mining.

• Collaborative filtering.
Show most relevant ads
Show most relevant ads
Show most relevant ads

Frequently Bought Together
Customers buy this book with Building Search Applications: Lucene, LingPipe, and Gate by Manu Konchady

Price For Both: $70.12

Customers Who Bought This Item Also Bought

| Solr 1.4 Enterprise Search Server by David | Hadoop: The Definitive Guide by Tom White | Hibernate Search in Action by Emmanuel | Introduction to Information Retrieval by | Collective Intelligence in Action by Satnam Alag |
Recommending places

Thanks to Falko Menge for the pictures of Brussels.
Recommending people
Recommendation mining.

- Online collaborative filtering on single machine.
- Offline Map/Reduce based version.
- Content similarity can be integrated.

- Based on former Taste project.
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.
Requirements to get started
Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment. Amazon EC2 reduces the time required to obtain

Amazon Elastic MapReduce

Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Using Amazon Elastic MapReduce, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial
(Thanks to Thilo for helping set up the cluster, Thanks to packet and masq for two of the three machines.)
Requirements to get started
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 a committer.
Become a committer: Of Apache Mahout

Sebastian Schelter
Jake Mannix
Benson Margulies
Robin Anil
David Hall
AbdelHakim Deneche
Karl Wettin
Sean Owen
Grant Ingersoll
Otis Gospodnetic
Drew Farris
Jeff Eastman
Ted Dunning
Isabel Drost

Emeritus:
Niranjan Balasubramanian
Erik Hatcher
Ozgur Yilmazel
Dawid Weiss
Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

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

Image by: Patrick McEvoy
Thanks to Tim Lossen et. al for taking amazing pictures of the conf.
Berlin Buzzwords 2011

Search/ Store/ Scale

May/ June 2011

Thanks to Tim Lossen et. al for taking amazing pictures of the conf.
*user@hadoop.apache.org
*dev@hadoop.apache.org

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

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

Image by: Patrick McEvoy