Apache Mahout
Making data analysis easy
Isabel Drost

Nighttime:

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

Daytime:

Software developer.
Guest lecturer at TU Berlin.
Co-Organiser Berlin Buzzwords 2010.
“Mastering Data-Intensive Collaboration and Decision Making”

EU funded research project
- Number of partners: 8
- Coordinator: Research Academic Computer Technology Institute (CTI), Greece
Hello Apache Con!
Hello Apache Con!
Hello Apache Con!
Hello Apache Con!

Apache Solr
Hello Apache Con!

Machine learning background?
Hello Apache Con!
Agenda

- Data Mining/ Machine Learning?
- Why is scaling hard?
- Going beyond simple statistics.
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.
Machine learning – what's that?
Archimedes taking a Warm Bath
Archimedes model of nature

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

\[ \frac{\text{Weight}}{\text{Weight – Apparent immersed weight}} \]
An SVM's model of nature
The challenge
Mission

Provide scalable data mining algorithms.
The Colorful History of the Internet and its Increasingly Problematical Future

The rapid development of the Internet has brought forth the need for new regulations and technologies. The development of the Internet took place in the late 1960s and 1970s, when computers were used to connect different networks and communicate data. The invention of the World Wide Web in the late 1980s and early 1990s marked a turning point in the history of the Internet, as it became a global network of information.

The 1940s, 1950s, and 1960s. In the late 1940s and 1950s, when computers were used primarily for calculus and scientific calculations, the first systems were developed to connect them. The development of computer networks was a significant advancement in the field of communication. By the 1960s, computers were beginning to connect to each other, and the concept of the Internet was taking shape.

The 1970s and 1980s. The development of computer networks continued in the 1970s and 1980s, with the creation of the ARPANET and the development of protocols such as TCP/IP. The Internet was becoming more accessible and user-friendly, with the development of graphical user interfaces and web browsers.

The 1990s. In the 1990s, the Internet became a global phenomenon, with the development of the World Wide Web and the widespread adoption of the Internet. The Internet has revolutionized the way we communicate, work, and access information, and it continues to evolve and change our world.
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 Überwachungszustat.

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 Demonstrationen zogen sich unter anderem gegen die Verwaltungen, des Datenschutzes an.
The HDFS filesystem is not restricted to MapReduce jobs. It can be used for other applications, many of which are under way at Apache. The list includes the HBase database, the Apache Mahout machine learning system, and matrix operations.
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 Arbeitnehmer- und gegen die Gesundheitskarte: 150 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

In Berlin demonstrieren Tausende Demonstranten für mehr Datenschutz

Rund 7000 Demonstranten nahmen an dem Protest unter dem Motto "Freiheit statt Angst – Stopp des Überwachungswahns" in Berlin teil.

Experteer
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 an dem Protest gegen die Initiative der Bundesregierung unter der Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

experteer

REM\k noise.
Tausende demonstrieren für Bürgerrechte im Netz

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

Remove noise.

Convert text to vectors.
From texts to vectors
If we looked at two words only:
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 \textit{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.

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

- 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.
Algorithm choice

• Naive Bayes.
  • Cannot reliably indicate how certain its classification is.
Algorithm choice

- Naive Bayes.
  - Cannot reliably indicate how certain its classification is.

- Logistic Regression.
  - Complement. NB.
  - Random Forests.
Algorithm choice

- Do you
  - want to interpret the resulting model?
  - want to update the model in an online fashion?

- The data you are working with
  - lives in high-dim feature space but is sparse?
  - has features that might depend on each other?
  - has outliers?
  - has missing values?
From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.
Goals

• Did I use the best model parameters?

• How well will my model perform in the wild?
Prepare data

Train model

Compute expected performance
Performance

- Use same data for training and testing.

Problem:
- Highly optimistic.
- Model generalization unknown.
Performance

- Use same data for training and testing.

DON'T

- Problem:
  - Highly optimistic.
  - Model generalization unknown.
Performance

- Use just a fraction for training.
- Set some data aside for testing.

Problems:
- Pessimistic predictor: Not all data used for training.
- Result may depend on which data was set aside.
Performance

- Partition your data into n fractions.
- Each fraction set aside for testing in turn.

Problem:
- Still a pessimistic predictor.
Prepare data

Tune model parameters

Train model

Compute expected performance
Performance

- Use just a fraction for training.
- Set some data aside for tuning and testing.

Problems:

- Highly optimistic.
- Parameters manually tuned to testing data.
Performance

- Use just a fraction for training.
- Set some data aside for tuning and testing.

DON'T

- Problems:
  - Highly optimistic.
  - Parameters manually tuned to testing data.
Performance

- Use just a fraction for training.
- Set some data aside for tuning.
- Set another set of data aside for testing.

Problems:
- Pretty pessimistic as not all data is used.
- May depend on which data was set aside.
Performance Measures
Accuracy

\[
ACC = \frac{true \ positive + true \ negative}{true \ positive + false \ positive + false \ negative + true \ negative}
\]

- Problems:
  - What if class distribution is skewed?
Precision/ Recall

\[
\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}
\]

\[
\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}
\]

- Problem:
  - Depends on decision threshold.
ROC Curves
ROC Curves

Orange rate
ROC Curves

True orange rate

False orange rate
ROC Curves

True orange rate

False orange rate
ROC Curves

True orange rate

False orange rate
ROC Curves

True orange rate

False orange rate
AUC – area under ROC

True orange rate

False orange rate
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.
Discover groups of items

• Group items by similarity.

• Examples:
  - Group news articles by topic.
  - Find developers with similar interests.
**Qaeda-linked group claims Baghdad bomb attacks**

Reuters - Andrew Hammond - 2 hours ago

DUBAI (Reuters) - An al Qaeda-linked group has said it carried out the twin suicide bombings that killed 155 people in Baghdad on Sunday and revived doubts about security in the run-up to Iraq's elections in January.

**Obama vows no rush on Afghanistan**

BBC News - 3 hours ago

US President Barack Obama has said he will "never rush" a decision to send more troops to Afghanistan, as he comes under pressure to set out a new policy.

**Karadzic court case due to resume**

BBC News - 1 hour ago

The genocide and war crimes trial of former Bosnian Serb leader Radovan Karadzic is due to resume in The Hague, a day after it was adjourned.
Recommendation mining.

- Collaborative filtering.
Show most relevant ads
Recommending places

Thanks to Falko Menge for the pictures of Brussels.
Recommending people
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.)
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:
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.
Thanks to Tim Lossen et. al for taking amazing pictures of the conf.
I can't recommend this conference enough. Top industry speakers, top developers and fantastic organisation. Mark this event on your sponsoring calendar!" - Scott Robinson, Senior Marketing Manager, neofonie GmbH

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

Great to have this kind of conferences here in Berlin. Enjoyed to get a good and overview about the various NoSQL options.

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great to have this kind of conferences here in Berlin. Enjoyed to get a good and overview about the various NoSQL options.

The conference gave me a good overview on all kinds of scalable open-source projects.

The Buzzwords conference last year put Berlin on the map as Europe's perhaps most important hub for startups and cutting edge web technology today. Already looking forward to the next!

- Eric Wahlforss, Soundcloud

I enjoyed it very much: Very good location, decent-sized auditoriums, very good wifi, practically all talks were very good: deep expertise and mostly very good presenting skills. I will definitely try to attend again if the event is continued next year. Nokia

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. Nokia

The conference gave me a good overview on all kinds of scalable open-source projects.
Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

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

Image by: Patrick McEvoy