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 Devoxx!
Hello Devoxx!
Hello Devoxx!
Hello Devoxxx!
Hello Devoxx!
Hello Devoxx!

Machine learning background?
Hello Devoxx!
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?
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath
Archimedes model of nature

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

\[
\frac{Weight}{Weight - \text{Apparent immersed weight}}
\]
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.
The Colorful History of the Internet and its currently Problematic Future

The 1960s, 1970s, and 1980s. As the 1960s and 1970s, when computers were made small, IBM created the first personal computer and the first portable computer. The IBM computer was the first to use a magnetic disk as its main storage device. In 1971, the first personal computer, the Altair 8800, was introduced. The Altair 8800 was a 4-bit microcomputer that could be used as a personal computer or as a professional computer. The Altair 8800 was designed by Ken Olsen and his group at the Massachusetts Institute of Technology (MIT). The Altair 8800 was a revolutionary product, as it was the first personal computer that could be used for personal use.

In 1981, the first personal computer, the IBM PC, was introduced. The IBM PC was a 16-bit microcomputer that could be used as a personal computer or as a professional computer. The IBM PC was designed by Jim Turin, who was the chief engineer at the IBM Personal Computer Division. The IBM PC was a revolutionary product, as it was the first personal computer that could be used for personal use.

In 1986, the first personal computer, the Apple II, was introduced. The Apple II was a 8-bit microcomputer that could be used as a personal computer or as a professional computer. The Apple II was designed by Steve Wozniak, who was the co-founder of Apple Computer. The Apple II was a revolutionary product, as it was the first personal computer that could be used for personal use.

In 1990, the first personal computer, the Macintosh II, was introduced. The Macintosh II was a 16-bit microcomputer that could be used as a personal computer or as a professional computer. The Macintosh II was designed by John Sculley, who was the chief executive officer of Apple Computer. The Macintosh II was a revolutionary product, as it was the first personal computer that could be used for personal use.

In 1995, the first personal computer, the IBM ThinkPad, was introduced. The IBM ThinkPad was a 32-bit microcomputer that could be used as a personal computer or as a professional computer. The IBM ThinkPad was designed by DavidCross, who was the chief executive officer of IBM. The IBM ThinkPad was a revolutionary product, as it was the first personal computer that could be used for personal use.
HowTo: From data to information.
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**.
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 lauschende 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 versammelten sich unter anderem vor der Volkszählungsverwaltung der Stadt Berlin. Sie protestierten gegen die Einführung der Gesundheitskarte und forderten eine stärkere Schutz des Arbeitnehmerdatenschutzes.

http://www.flickr.com/photos/redux/409356158/in/photostream/

http://www.flickr.com/photos/topsy/204929063/
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: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungstaat.

In Berlin demonstrieren Tausende Demokraten für mehr Datenschutz.

Rund 7500 Demonstranten nahmen an der Protestkundgebung unter dem Motto "Freiheit statt Angst – Stopp des Überwachungswahn" 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 Überwachungsgast.
Tausende demonstrieren für Bürgerrechte im Netz
Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 150 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsgastast.

Freiheit statt Angst
In Berlin demonstrieren tausende Demonstranten für den Datenschutz.

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 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.

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

- 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
Euclidean

Cosine
Step 3: Clustering
Until stable.
Reality

- Seed selection.
- Choice of initial k.
- Continuous updates.
- Regular addition of clusters.
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.
Recommendation mining.

- Collaborative filtering.
Show most relevant ads

1. Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60€
   - Neu kaufen: EUR 5,99
   - Bewertung: (1) ★★★★★
   - Drogerie & Bad: Alle 13 Artikel ansehen

2. Clearasil Ultra Anti-Pickel Reinigungspads, 65 Stück von Clearasil (Bad)
   - Neu kaufen: EUR 7,99
   - Gewöhnlich versandfertig in 1 bis 3 Wochen.
   - Kostenlose Lieferung möglich.
   - Drogerie & Bad: Alle 13 Artikel ansehen

3. World of WarCraft: Wrath of the Lich King (Add-on) von Vivendi Unive Vista / XP
   - Neu kaufen: EUR 39,99
   - Vorbestellbar
   - Kostenlose Lieferung möglich.
   - Games: Alle Artikel ansehen
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
Accepted devrooms

We were very pleased with the devroom proposals we received, as many were collaborative proposals with ample opportunity for cross-pollination between projects. As a side effect, this did make the decision process even harder.

- Data Analytics
  Open Source tools for Data Analytics. This includes Machine Learning, Text Mining and Natural Language Processing, Computer Vision, Recommender Engines, Visualisation, Data Mining, Large Scale and distributed data processing, etc.
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.
Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

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

Image by: Patrick McEvoy