From Data to Information

Apache Mahout

Speaker: Isabel Drost
Isabel Drost

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

Daytime:
Software developer @ Berlin
Hello ApacheCon visitors!
Agenda

- Motivation.
- HowTo: A path from data to information.
- 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.
HowTo: From data to information.
From data to information.

- Start collecting and storing data.
- Analyse and understand data.
- Answer more complex questions.
Collecting and storing data.
Data storage options

- Structured, relational.
  - Customer data.
  - Bug database.
Data storage options

• Continuous files.
  – Log data.
Data storage options

- Semi-structured data:
  - Documents.
  - Independent rows.
From data to information.

✔ Start collecting and storing your data.

• Analyse and understand your data.

• Answer more complex questions.
Understanding your data

• Data profiling.

• Goals:
  – Identify usual behaviour.
  – Find exceptional cases.

• Exact questions depend on domain.
Example: Shopping sessions

- Average amount of money spent.
- Number of customers per state.
- Min/Max age of customers.
- Number of shopping sessions.
- Words associated with product.
Example: Access Logs

- Average session length.
- Entry-/ exit-pages.
- Average number of hits/ day.
- Clean data.
Example: Textual documents

- Average length of documents.
- Distribution of document topics.
- Distribution of authors.
Understanding your data

- Analysing data in HDFS/ HBase/ CouchDB:
  - Write analysis code as Map/Reduce jobs.
  - Use higher level language.
From data to information.

- Start collecting and storing your data.
- Analyse and understand your data.
  - Answer more complex questions.
Analyse shopping lists

By tanakawho, http://www.flickr.com/photos/28481088@N00/349049527/
Interactive web search

• Input:
  - User votings.
  - Clicks on search results.
  - Query refinement logs.

• Task:
  - Create a perfect ranking.
Show most relevant ads
Show most relevant ads

1. Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60€
   Neu kaufen: EUR 5,99  2 Angebote ab EUR 5,15
   Lieferung bis Samstag, 22. März: Bestellen Sie innerhalb der nächsten 23 Stunden
   Kostenlose Lieferung möglich.
   ★★★★★ (1)
   Drogerie & Bad: Alle 13 Artikel ansehen

2. Clearasil Ultra Anti-Pickel Reinigungspads, 65 Stück von Clearasil (Ba
   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
Show most relevant ads

Lucene in Action

~ Otis Gosподnijc (Author), Erik Hatcher (Author)

List Price: $44.95
Price: $29.67 & this item ships for FREE with Super Saver Shipping. Details
You Save: $15.28 (34%)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

23 new from $19.99  15 used from $17.11

Also Available in: List Price: Our Price: Other Offers:
Paperback (2) $44.99 $29.69

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

Price For Both: $70.12
Add both to Cart Add both to Wish List
Show availability and shipping details

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 Ahuja
Find emerging news topics

Olympics

Johnson ends gold drought with Olympic win in balance beam
Kansas City Star - 51 minutes ago
US gymnast Shawn Johnson had a big smile and a gold medal at the end of Tuesday's balance beam. B EIJING | Jerry Seinfeld always says that the silver medal at the Olympics has to be the worst one.

Video: Olympics '08: Johnson, Liukin Take More Medals

The Making Of China's Olympic Golden Age
CBS News - 43 minutes ago
China is obsessed with Olympic gold that it is training 200,000 handpicked kids in state-run sports boarding schools. (CBS) CBS News staffers file insider impressions and share their experiences throughout the day.

Americans' necks full of Olympic medals, but precious few are gold
Dallas Morning News
Machine learning – what's that?
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
The challenge
• Large amounts of data.

• Structured and unstructured data.

• Diverse tasks.
Mission

Provide scalable data mining algorithms.
• Commercially friendly license.

• Scalable to large amounts of data.

• Well documented.

• Healthy community.

• Targeted to developers.
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.
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.

Obama vows no rush on Afghanistan

BBC News - 3 hours ago

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

Al-Qaeda linked group claims Baghdad bomb attacks

Reuters - Andrew Hammond - 2 hours ago

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.

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.
Discover groups of similar items

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

- Example: Synthetic Control
  - Example Job: `<MAHOUT_HOME>/examples`
  - Outputs clusters

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

- Naïve Bayes.

- Complementary naïve bayes.

- Random forests.

- Others upcoming.
Assign items to defined categories

- Examples based on “standard” datasets:
  - 20 Newsgroups
    http://cwiki.apache.org/MAHOUT/twentynewsgroups.html
  - Wikipedia
    http://cwiki.apache.org/MAHOUT/wikipediabayesexample.html
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
Upcoming

- “TU Winter of Code”
  - Crawl and store blog postings.
  - Group posts and identify emerging topics.
  - Index results with Solr.

Database Systems and Information Management

Prof. Dr. Volker Markl
TU Winter of Code

- 6 students, 5 months.
- http://github.org/MaineC/Playground
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.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.
March 2009: Hadoop* Get Together in Berlin

- Dragan Milosevic ( ): TBA

- YOU!

newthinking store Berlin
Tucholskystr. 48

* UIMA, Hbase, Lucene, Solr, katta, Mahout, CouchDB, pig, Hive, Cassandra, Cascading, JAQL, ... talks welcome as well.
Mahout Meetup this evening
Interest in solving hard problems.
Being part of lively community.
Engineering best practices.

Bug reports, patches, features.
Documentation, code, examples.
<table>
<thead>
<tr>
<th>From</th>
<th>Grant Ingersoll <a href="mailto:gsing...@apache.org">gsing...@apache.org</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Re: Lucene Branding: the TLP, and &quot;Lucene Java&quot;</td>
</tr>
<tr>
<td>Date</td>
<td>Wed, 11 Apr 2007 01:13:36 GMT</td>
</tr>
</tbody>
</table>

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 w/ 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.

Until stable:

Done in Map.

Done in Reduce.
Make searching the web easier

Find job postings on the internet.

Group by categories.

Automatically extract: Place, title, date...