Apache Mahout - Overview

Table of contents

1. Apache Lucene Mahout

Mahout's goal is to build scalable machine learning libraries. With scalable we mean:

- Scalable to reasonably large data sets. Our core algorithms for clustering, classfication and batch based collaborative filtering are implemented on top of Apache Hadoop using the map/reduce paradigm. However we do not restrict contributions to Hadoop based implementations: Contributions that run on a single node or on a non-Hadoop cluster are welcome as well. The core libraries are highly optimized to allow for good performance also for non-distributed algorithms.
- Scalable to support your business case. Mahout is distributed under a commercially friendly Apache Software license.
- Scalable community. The goal of Mahout is to build a vibrant, responsive, diverse community to facilitate discussions not only on the project itself but also on potential use cases. Come to the mailing lists to find out more.

Currently Mahout supports mainly four use cases: Recommendation mining takes users' behavior and from that tries to find items users might like. Clustering takes e.g. text documents and groups them into groups of topically related documents. Classification learns from exisiting categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category. Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.

Interested in helping? See the Wiki or send us an email. Also note, we are just getting off the ground, so please be patient as we get the various infrastructure pieces in place.

2. Mahout News

2.1. 29 March 2010 - Google Summer Of Code Projects

Its Summer of Code time again and ASF is accepting proposals from students. Mahout has a number of people willing to be mentors, so if you are a student interested in working on machine learning algorithms using Hadoop or improving the Mahout framework, then please check out our Summer of Code wiki page.

2.2. 17 March 2010 - Apache Mahout 0.3 released

The Apache Lucene project is pleased to announce the release of Apache Mahout 0.3.

Highlights include:

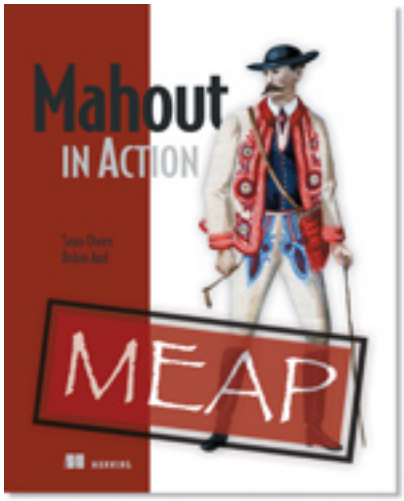
- New: math and collections modules based on the high performance Colt library
- Faster Frequent Pattern Growth(FPGrowth) using FP-bonsai pruning
- Parallel Dirichlet process clustering (model-based clustering algorithm)
- Parallel co-occurrence based recommender
- Parallel text document to vector conversion using LLR based ngram generation
- Parallel Lanczos SVD(Singular Value Decomposition) solver
- Shell scripts for easier running of algorithms, utilities and examples
- ... and much much more: code cleanup, many bug fixes and performance improvements

Details on what's included can be found in the release notes.

Downloads are available from the Apache Mirrors

2.3. 16 January 2010 - Mahout in Action book discount available

Mahout in Action, a forthcoming book on Mahout, is underway. The first 6 chapters, covering recommender engines and collaborative filtering, are available for early access via Manning's "MEAP" program. Until February 28, 2010, get 35% off with discount code "mahout35".



2.4. 17 Nov. 2009 - Apache Mahout 0.2 released

The Apache Lucene project is pleased to announce the release of Apache Mahout 0.2.

Highlights include:

- Significant performance increase (and API changes) in collaborative filtering engine

- K-nearest-neighbor and SVD recommenders
- Much code cleanup, bug fixing
- Random forests, frequent pattern mining using parallel FP growth
- Latent Dirichlet Allocation
- Updates for Hadoop 0.20.x

Details on what's included can be found in the release notes.

Downloads are available from the Apache Mirrors

2.5. 14 August 2009 - Lucene at US ApacheCon

ApacheCon US is once again in the Bay Area and Lucene is coming along for the ride! The Lucene community has planned two full days of talks, plus a meetup and the usual bevy of training. With a well-balanced mix of first time and veteran ApacheCon speakers, the Lucene track at ApacheCon US promises to have something for everyone. Be sure not to miss:

Training:

- Lucene Boot Camp - A two day training session, Nov. 2nd & 3rd
- Solr Day - A one day training session, Nov. 2nd

Thursday, Nov. 5th

- Introduction to the Lucene Ecosystem - Grant Ingersoll @ 9:00
- Lucene Basics and New Features - Michael Busch @ 10:00
- Apache Solr: Out of the Box - Chris Hostetter @ 14:00
- Introduction to Nutch - Andrzej Bialecki @ 15:00
- Lucene and Solr Performance Tuning - Mark Miller @ 16:30

Friday, Nov. 6th

- Implementing an Information Retrieval Framework for an Organizational Repository - Sithu D Sudarsan @ 9:00
- Apache Mahout - Going from raw data to Information - Isabel Drost @ 10:00
- MIME Magic with Apache Tika - Jukka Zitting @ 11:30
- Building Intelligent Search Applications with the Lucene Ecosystem - Ted Dunning @ 14:00
- Realtime Search - Jason Rutherglen @ 15:00

2.6. 07 April 2009 - Apache Mahout 0.1 released

The Apache Lucene project is pleased to announce the release of Apache Mahout 0.1. Apache Mahout is a subproject of Apache Lucene with the goal of delivering scalable

machine learning algorithm implementations under the Apache license. The first public release includes implementations for clustering, classification, collaborative filtering and evolutionary programming.

Highlights include:
- Taste Collaborative Filtering
- Several distributed clustering implementations: k-Means, Fuzzy k-Means, Dirchlet, Mean-Shift and Canopy
- Distributed Naive Bayes and Complementary Naive Bayes classification implementations
- Distributed fitness function implementation for the Watchmaker evolutionary programming library
- Most implementations are built on top of Apache Hadoop (http://hadoop.apache.org) for scalability

Details on what's included can be found in the release notes.

Downloads are available from the Apache Mirrors

2.7. 09 February 2009 - Lucene at ApacheCon Europe 2009 in Amsterdam

Lucene will be extremely well represented at ApacheCon US 2009 in Amsterdam, Netherlands this March 23-27, 2009:

- Lucene Boot Camp - A two day training session, March 23 & 24th
- Solr Boot Camp - A one day training session, March 24th
- Introducing Apache Mahout - Grant Ingersoll. March 25th @ 10:30
- Lucene/Solr Case Studies - Erik Hatcher. March 25th @ 11:30
- Advanced Indexing Techniques with Apache Lucene - Michael Busch. March 25th @ 14:00
- Apache Solr - A Case Study - Uri Boness. March 26th @ 17:30
- Best of breed - httpd, forrest, solr and droids - Thorsten Scherler. March 27th @ 17:30
- Apache Droids - an intelligent standalone robot framework - Thorsten Scherler. March

26th @ 15:00

2.8. 22 July 2008 - Lucene at ApacheCon New Orleans

Lucene will be extremely well represented at ApacheCon US 2008 in New Orleans this November 3-7, 2008:

- Lucene Boot Camp - A two day training session, November 3rd & 4th
- Solr Boot Camp - A one day training session, November 4th
- Anentire day of Lucene sessions, including a talk on Mahout by Mahout committer Grant Ingersoll, on November 5th

2.9. 4 April 2008 - Mahout - Now with more Taste!

We are pleased to announce that the Taste Collaborative Filtering (Taste on SourceForge) has donated it's codebase to the Mahout project. In the coming weeks and months we will work to bring it into Mahout and then make it run on Hadoop, bringing truly large scale collaborative filtering capabilities to our users.

2.10. 16 March 2008 - Google Summer Of Code Projects

The ASF is in the process of creating projects for Google's annual Summer of Code Project. Mahout has a number of people willing to be mentors, so if you are a student interested in working on machine learning algorithms using Hadoop, then please check out the ASF Summer of Code wiki page.

2.11. 22 January 2008 - Mahout launches

The Lucene PMC announces the creation of the Mahout subproject.