

APACHE
LUCENE
EUROCON



Configuring mahout Clustering Jobs

Frank Scholten, DutchWorks
frank.scholten@dutchworks.nl, 19 october 2011

Presented by

lucid
IMAGINATION

Lucene

Apache
Solr 

My Background

Frank Scholten



@Frank_Scholten

Software Developer at

dutchworks[™]
formerly known as JTeam

Blogger at

 SearchWorkings



user & contributor

Agenda

What is clustering?



Intro to



Clustering



Clustering introduction

So much
data...

How to get a nice
overview?



What is clustering?

Grouping & summarizing data

Unsupervised machine learning

“...the assignment of a set of observations into subsets so that observations in the same clusters are similar in some sense...”

Source: Wikipedia

Applications

Market segmentation

Species identifications

Machine vision

Information retrieval & search

...and many more!

Example - Google news

Google news

open source

Search News

[Advanced news search](#)

News

[Top Stories](#)

[More sections ▾](#)

All news

[Images](#)

[Blogs](#)

Any recent news

[Past hour](#)

[Past day](#)

[Past week](#)

[Past month](#)

[Archives](#)

Sorted by relevance

[Sorted by date](#)

Follow "open source" news

[Dreaming of an Open-Source CES 2012](#) ☆

PC World - [Katherine Noyes](#) - 5 hours ago

Exciting as this year's debuts have been, however, I can't help but think ahead with fresh hopes for next year--hopes for a bigger presence for **open-source** ...



iPhone FAQ

[No GPL Apps for Apple's App Store](#) ☆

ZDNet (blog) - [Steven J. Vaughan-Nichols](#) - 7 minutes ago

VLC media player is free software licensed solely under the terms of the **open source** GNU General Public License (aka GPL). Those terms are contradicted by ...

[Apple pulls VLC from App Store over open-source DRM dispute](#) SlashGear

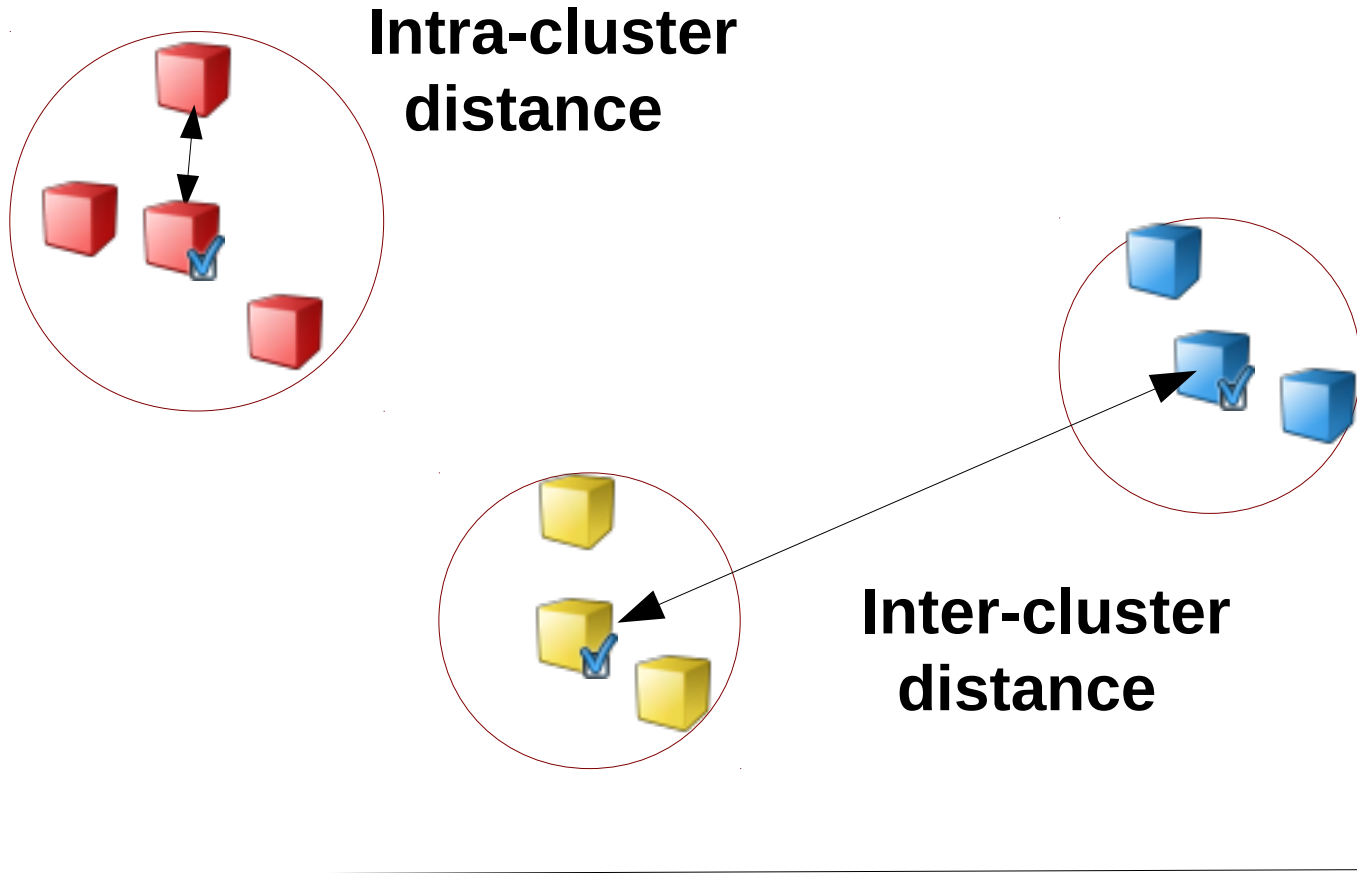
[VLC for iOS removed from the App Store](#) 9 to 5 Mac

[Not A Good Day For iPad Users As Apple Forced To Pull VLC App](#) Apple Bitch (blog)

[Electronista](#)

[all 16 news articles »](#) AAPL - MSFT

2-D Clustering example



Legend Point  Cluster  Cluster Center 

K-Means algorithm

Select K random vectors

Specify distance measure + threshold

Every iteration

- Add vector closest to cluster
- Recompute center
- Converged if no vectors within threshold

Intro to mahout



Collaborative
Filtering



Is this SPAM?

Classification



Clustering

And much more!



The Project

Apache project started in 2008

Scalable machine learning, often with Hadoop

Steadily growing community

Version 0.6 coming soon



bin/mahout

```
frank@frankthetank:~$ mahout
```

```
no HADOOP_HOME set, running locally
```

```
An example program must be given as the  
first argument.
```

```
Valid program names are:
```

```
arff.vector: : Generate Vectors from an ARFF  
file or directory
```

```
canopy: : Canopy clustering
```

```
cat: : Print a file or resource as the  
logistic regression models would see  
it
```

```
...
```



Help

```
frank@frankthetank:~$ mahout kmeans --help
```

```
usage: <command> [Generic Options]
```

```
[Job-specific Options]
```

```
Generic Options:
```

```
...
```

```
Job-specific Options:
```

```
--input (-i) input Path to job input directory.
```

```
--output (-o) output The directory pathname for  
output.
```

```
...
```



Java Drivers

```
String[] args = new String[] {  
    "--input", input,  
    "--output", output,  
    "--clusters", clusters,  
    "--clustering",  
    "--numClusters", "10"  
};
```

```
ToolRunner.run(conf, new KmeansDriver(), args);
```

Text clustering process



[0.03, 0.95, 0.45, 0.34]
[0.02, 0.98, 0.73, 0.55]

Text files or
Lucene index

Sequence files

Vectors



ABC

Find n-grams



Dictionary



Quick fox



Dog

Cluster labels



K-means



Clusters



(CL-1,23)
(CL-1,37)
(CL-2,45)

Points

(CL-1, [0.32, 0.6, ..]
(CL-1, [0.76, 0.1, ..]
(CL-2, [0.98, 0.2, ..]



Text clustering programs



\$ mahout seqdirectory

**Text files or
Lucene index**

Sequence files



[0.03, 0.95, ..]
[0.29, 0.98, ..]

\$ mahout seq2sparse

Sequence files

Vectors

[0.03, 0.95, ..]
[0.29, 0.98, ..]



\$ mahout kmeans

Vectors

Clusters

Clustering



Publicly available monthly dumps

Posts ~ **5.5 GB** ~ **1.4 M** questions (April 2011)

Let's use  **mahout** to extract a tag cloud!

Clustering



Cluster

Vectorize

Index

Text

[0,1,0,1,1,1,0,0,1,0,1]
[0,1,1,1,1,0,0,0,0,0,1]



Join content
& clusters



Java Git Lucene
Regular expressions
Version control



Post ID
& Title



Pre-process
XML & HTML



Clustering



How to implement?

Pre-process XML  XML & HTML parsing

Vectorize  Custom Analyzer

Cluster  

Index  

Vectorize

```
[ 0,1,0,1,1,1,0,0,1,0,1 ]  
[ 0,1,1,1,1,0,0,0,0,0,1 ]
```

Many options and flags

```
$ mahout seq2sparse  
  --input ..  
  --output ..  
  --analyzerClass ..  
  --maxDFPercent ..  
  --minDF ..
```

Cluster



Run one of the clustering algorithms!

K-means, Fuzzy K-means, Canopy,
Mean-shift, Min-hash, LDA

All with different pros and cons

Index



Custom code to join data at index time

Index clusters

(cluster_id, cluster_name, size)

Index posts

(post_id, post_cluster_id, title)

Demo time!

Conclusions

Clustering is fun!

Vectorization & labeling improvements

Tools for cluster evaluation?

References

Mahout in Action – Just released!

Sean Owen, Ted Dunning, Robin Anil, Ellen Friedman



<mailto:{user|dev}@mahout.apache.org>
<http://jira.apache.org/MAHOUT>



<http://www.searchworkings.org>

Q&A