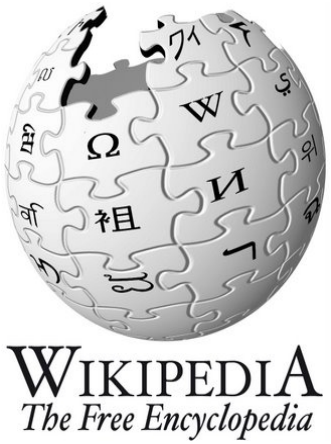
Apache Mahout. Применение модели MapReduce для задач машинного обучения.

Кузнецов Виталий
vitty@altlinux.ru

Apache Mahout

- **Масштабируемая библиотека алгоритмов машинного обучения, поддерживающая большие объёмы данных**

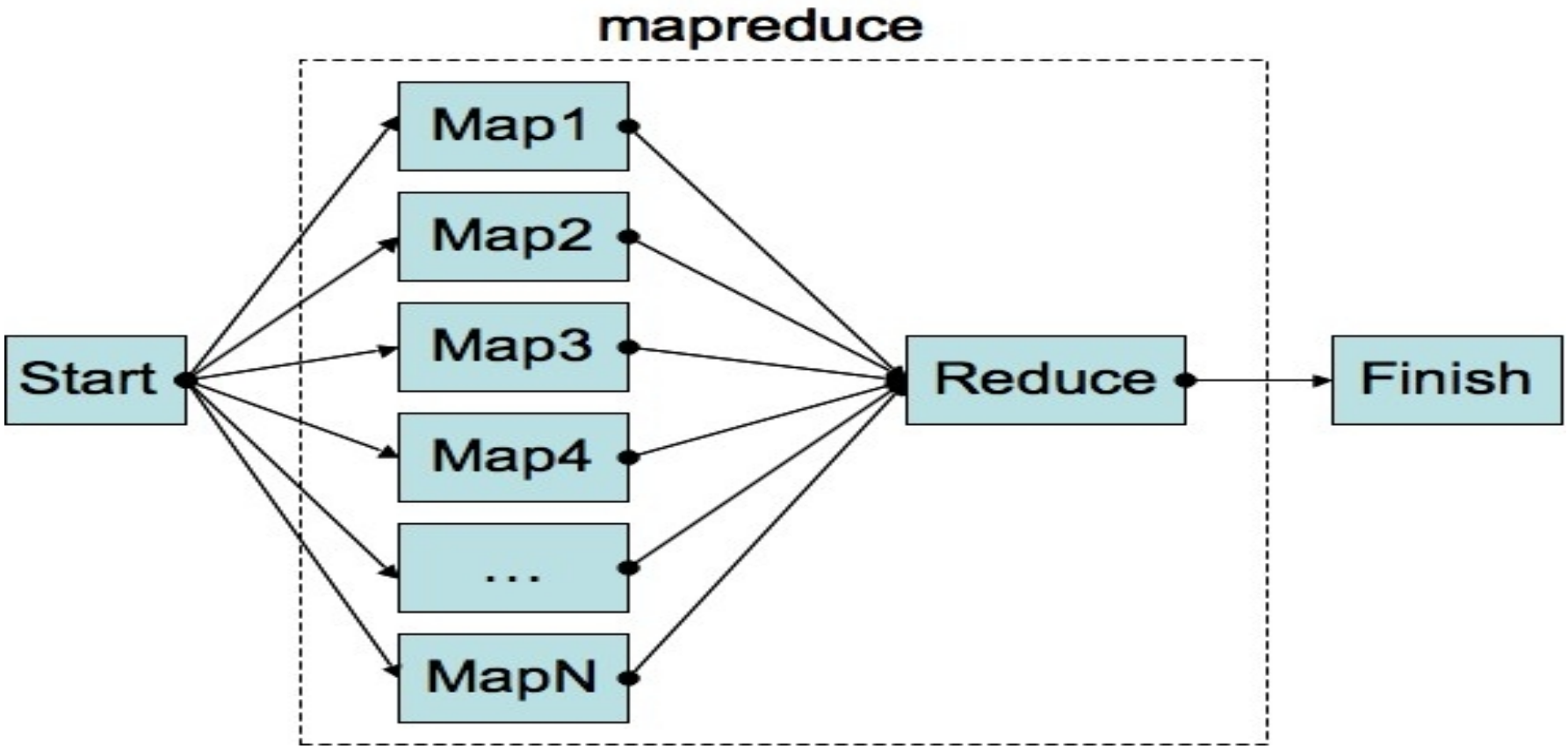


- **Wikipedia:** A ***mahout*** is a person who drives an elephant. The word mahout comes from the Hindi words *mahaut* and *mahavat*.

Apache Hadoop

- Распределённое хранение (HDFS)
- Распределённая обработка (MapReduce)

MapReduce



Алгоритмы

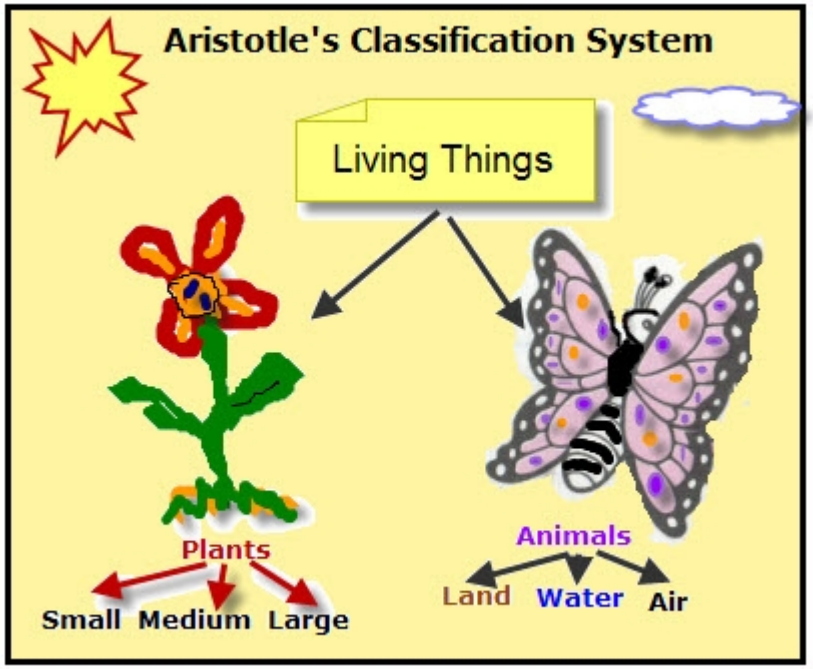
Основные алгоритмы:

- Классификации (Classification)
- Кластеризации (Clustering)
- Рекомендации (Recomenders)

+

- Понижение размерности (Dimension reduction)
- Эволюционные алгоритмы (Evolutionary Algorithms)
- ...

Классификация



Поддерживаются:

- Логистическая регрессия
- Байесовский классификатор
- Случайный лес

В разработке:

- Нейросети

  в т.ч (Perceptron, Winnow, Restricted BoltzmannMachines)

- Метод опорных векторов

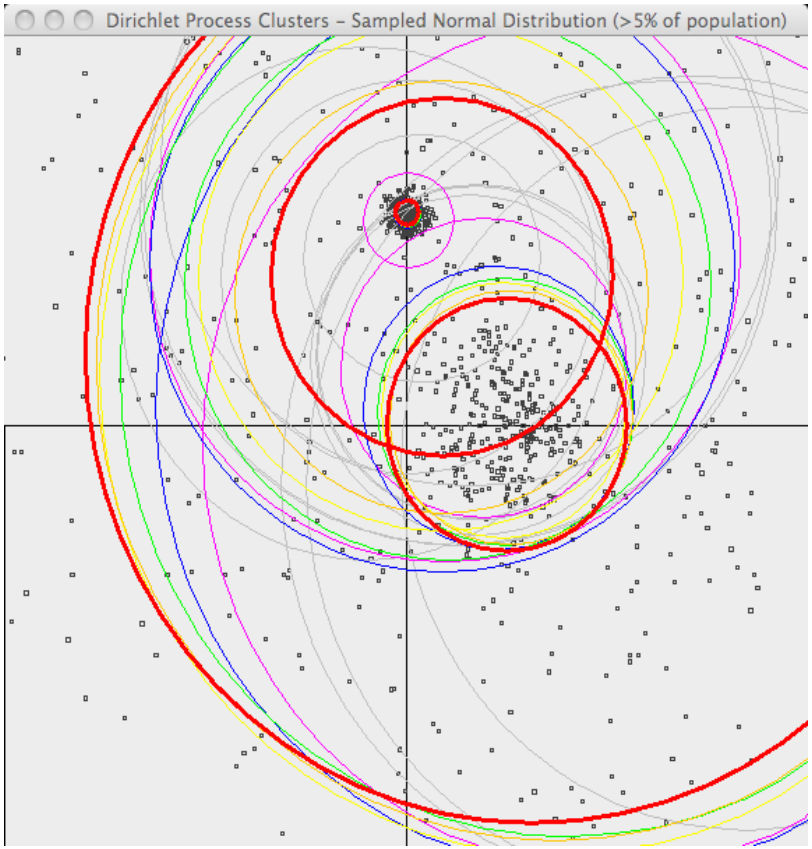
Кластеризация

Поддерживаются:

- Canopy
- K-means,Fuzzy K-means
- Mean Shift,
- Expectation Maximization
- Dirichlet Process,
- Latent Dirichlet Allocation

В разработке:

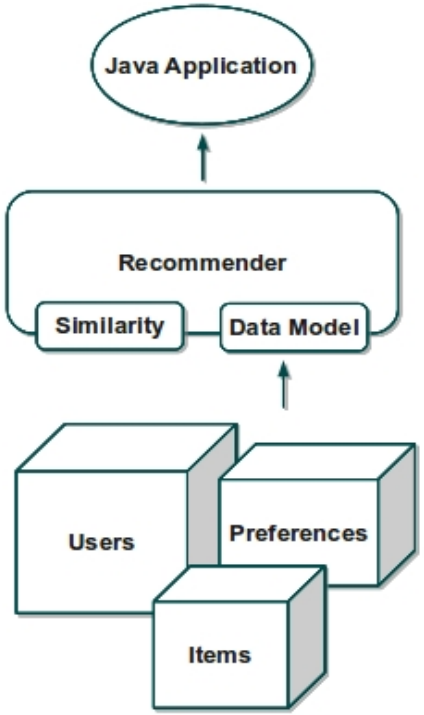
- Hierarchical Clustering



Рекомендации

Поддерживаются:

- Локальные рекомендации
- Распределённые (MapReduce)



Пример задачи

Классификация изменений исходного кода

- Исходные данные: система контроля версий (svn, git, …)
- Параметры:
  - Добавленные/удалённые/изменённые строки кода
  - Сложность добавленного/удалённого/изменённого кода
  - Число добавленных/удалённых/изменённых классов/структур
- Кластеризация (к пример — K-means)
- Экспертная оценка нескольких изменений из кластеров

Ещё о Mahout

Проект верхнего уровня Apache с мая 2010

- **Сайт проекта: http://mahout.apache.org**
- Списки рассылки: https://cwiki.apache.org/confluence/display/MAHOUT/Mailing+Lists
- **Книги (готовятся к публикации):**
  - **Mahout In Action, Май 2011**
  - **Taming Text, Весна 2011**
- **Используется в production:**
  - AOL, YAHOO!, foursquare

Кузнецов Виталий
email: vitty@altlinux.ru

Спасибо за внимание!
Вопросы?