"An Efficient HC Method for Large Datasets with Map-Reduce"

Relevance:
→ mentions applications to astronomical datasets.

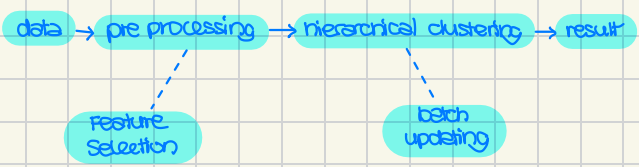
Introduction:
→ large datasets pose challenges for data-mining algorithms to efficiently process data within given constraints (e.g. memory, execution time).

→ to overcome constraints, data mining algos can be implemented with Map-Reduce:
- Map Reduce breaks large datasets into small chunks and processes them in parallel on multiple cluster nodes and scales easily to mine hundreds of TBs of data.

→ low efficiency of HCA stems from two aspects:
i) HCA of large datasets consists of many successive iterations of clustering processes in which feature matrix merging & updating & similarity value modification are common operations.
- invokes file operations of distributed file system and constant input-output (IO) operations.

ii) large dimension of feature vectors demands high memory usages.

→ two proposed optimization techniques:
i) co-currence based feature selection at the pre-processing stage
ii) batch updating to reduce IO overhead, batching as many IO and communications operations as possible in one iteration.



Overview
→ mining application consists of two major phases:
i) preprocessing of raw data
ii) HC of user groups
} specific data used in this study are web access logs.

heavy IO overhead

old Approach: