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 churks and processes them in parallel on multiple diuster nodes and scales easily to mine hundreds of TBs of data. > low efficiency of HCA stems from two aspects: i) HiCA of large distasets consists of many successive iterations of clustering processes in which feature matrix merging & updating & similarity value modification are common operations. · involves file operations of distributed file system and constant input-output (10) 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) betch updating to reduce 10 overhead, batching as many 10 and communications operations as possible in one iteration. data - one processing - nierarchical clustering - result **Feature** updeling Selection Overview -> mining application consists of two major phases: near 10 overhead i) preprocessing of you data specific data used in this ii) HC of user groups Study are web access lops. Old Approach: HCA Stage: Preprocessing Stage: · prunes web access logs · users grouped based on · extracts web page topics xintem browysky - nestu using bottom - up clustering. generates user-keyword matrix eliminating less user-keyword clustering is clusters merge matrix peneraled ntiw basilallared useful access records until termination with uses as Map-Reduce condition is rows & words words in title & met Keyword metadata as columns in site headers totalias ere

Cocumence Based Feature Selection

- one method to improve efficiency is to reduce the dimension of the user- Keyword matrix.
 this can be done with CBFS.
- → motivations for CBFS:
- f · keywords in title vs. metadata are not equally important in weight; some are even notay.

 U · summay of keywords in weighage can be given by sample keywords.
 - CBFs reduces dim. of feature vectors by summarizing a user's interested topics with the most representative keywords.
 - 'for any two keywords, their co-cumence frequency is calculated to reflect the relationant in semantics between Key words.
 - · higher element values attention degrees (d.m.) of feature vectors is given to more related key words.

HCA with Balth updating

Phase 1: Batch choosing pairs of user groups

Top N pairs of user groups with highest similarity values

For every pair of user groups

Phase 2: In-memory batch updating"

(ii) Adding a user group pair (iii) Batch Merges and Updates a user group pair

(ii) Removing a user group pair

Phase 3: Batch merging and updating

Updating user-matrix file | Modify similarity values

Figure 2. Hierarchical Clustering with Batch Updating

Table II NUMBERS OF ITERATIONS OF BATCH UPDATING AND EXECUTION TIMES WITH DIFFERENT SIZE N OF C-QUEUE

N	# of iter.	Execution time(seconds)
500	1678	46962
100	1680	28603
50	1680	21412
10	1684	26615

Table III
NUMBERS OF UPDATES AND EXECUTION TIMES OF HIERARCHICAL
CLUSTERING WITH AND WITHOUT BATCH UPDATING

	# of iter.	Execution time(seconds)
With BU	1680	21412
Without BU	27939	287258

feature selection ranks input variables in terms
 of how useful they are to predict the texpet value.

Betch updating is proposed to improve HCA efficiency:

- combines several iterations of clustering into one so that the updating mathrifle and modifying similarity values are processed in batch mode.
- Uses two new diale structures for efficient clustering:
- i) C-queue: a queue that stores the

top N pairs of user groups
with the highest similarity
values.
C.1 J desrotes the pair contains
C.1 J user groups ut & uj

ii) Batch-queue: a queue that stores

Cij to be batch
processed in one
iteration, in order.

Results:

- -> CBTR reduced dim. of feature
- → Batch updating decreases iteration of clustering to very

Takeaways:

Y15.

- > consider as an alternative to PCA?
 - * PCA tries to reduce dimensionality by exploring linear departiency.
 - · PCA is unsupervised.