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 user-keyword matrix 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