HICAP: Hierarchical Clustering with Pattern Preservation

Hui Xiong

Department of Computer Science & Engineering
University of Minnesota - Twin Cities

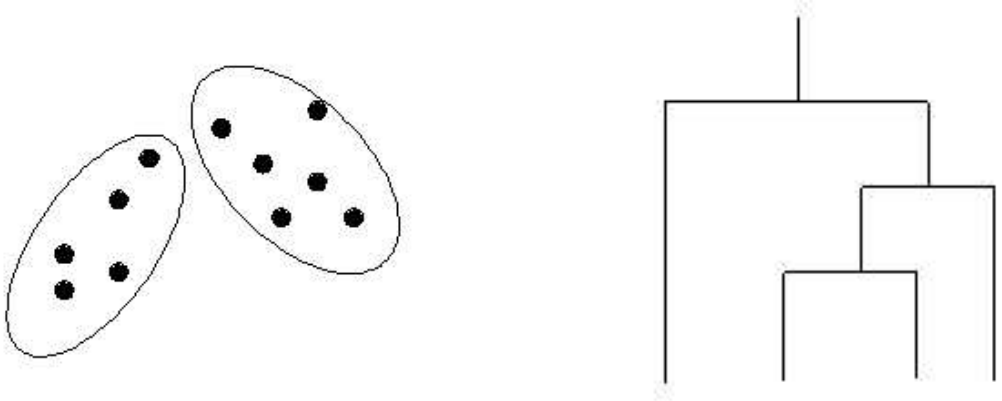
Overview

⇒ Introduction

- The Hyperclique Pattern
- HICAP: Approach and Algorithm
- Experimental Results
- Conclusions and Future Work

Clustering

- Partitional clustering partitions objects into disjoint groups
  - Such as K-means, DBSCAN



- Hierarchical (nested) clustering produces a nested set of clusters
  - Each level is equivalent to a partitional clustering
  - Start with each point as a cluster and merge clusters according to some scheme, e.g. Single link, complete link, group average.

Group Average Hierarchical Clustering

- Similarity between clusters is based on the pairwise average similarity between the objects to be clustered.
- Also called UPGMA: Unweighted pair-group method using arithmetic averages

$$sim(Cluster_i, Cluster_j) = \frac{\Sigma_{O_i \in Cluster_i, O_j \in Cluster_j} sim(O_i, O_j)}{|Cluster_i| \cdot |Cluster_j|}$$

Pattern Preserving Clustering: Motivation

- In many domains, there are groups of objects that are involved in strong patterns that are key for understanding this domain.
  - In text mining, collections of words that form a topic.
  - In genomics, sequences of nucleotides that form a functional unit.
- We want to design a clustering schema that preserves these patterns, i.e. that puts the objects or attributes of these patterns in the same cluster.
  - Otherwise, the resulting clusters will be harder to understand since they must be interpreted solely in terms of objects instead of well-understood patterns.
  - The value of a data analysis is greatly diminished for end users.

Overview

- Introduction

⇒ The Hyperclique Pattern

- HICAP: Approach and Algorithm
- Experimental Results
- Conclusions and Future Work

The Hyperclique Pattern

- [**The H-confidence Measure:**] The **h-confidence** of an itemset $P = \{i_1, i_2, \cdots, i_m\}$ is defined as $hconf(P) = \min [conf\{i_1 \rightarrow i_2, \ldots, i_m\}, conf\{i_2 \rightarrow i_1, i_3, \ldots, i_m\}, \ldots, conf\{i_m \rightarrow i_1, \ldots, i_{m-1}\}]$, where $conf$ follows from the conventional definition of association rule confidence.

- [**Hyperclique Pattern:**] An itemset $I = \{i_1, i_2, \ldots, i_m\}$ is a hyperclique pattern if $hconf(P) \geq h_c$, where h_c is a user-specified minimum h-confidence threshold.

Hyperclique: Example

- For a hyperclique pattern $P = \{A, B, C\}$, assume that:
  - $supp(\{A\}) = 0.1$, $supp(\{\mathsf{B}\}) = 0.1$, $supp(\{C\}) = 0.06$, $supp(\{A, B, C\}) = 0.06$.

$\Rightarrow$

  - $conf\{A \rightarrow B, C\} = supp(\{A, B, C\})/supp(\{A\}) = 0.6$
  - $conf\{B \rightarrow A, C\} = 0.6$
  - $conf\{C \rightarrow A, B\} = 1$.

- Hence, the h-confidence of the hyperclique pattern P is:
  - $hconf(P) = min\{conf\{B \rightarrow A, C\}, conf\{A \rightarrow B, C\}, conf\{C \rightarrow A, B\}\}$= 0.6.

Properties of H-confidence

- Computational Efficency
  - Anti-monotone property
    - h-confidence is non-increasing in the size of the itemset.
  - Cross-support property
- High Affinity Nature
  - The items in an itemset with h-confidence h are guaranteed to have a pairwise cosine similarity of h.
    - For example, items in an itemset with h-confidence 0.5 are guaranteed to have a pairwise cosine similarity of 0.5.
  - Items in a hyperclique pattern are closed associated in a way that agrees well with the goal of clustering to find groups of objects that have high similarity.

Overview

- Introduction
- The Hyperclique Pattern

⇒ HICAP: Approach and Algorithm

- Experimental Results
- Conclusions and Future Work

HICAP: Key Foundations

- Hypercliques seem like a promising pattern on which to base pattern preserving clustering.
  - High affinity nature.
  - Efficency of finding hypercliques vs. frequent patterns.
  - Smaller number of patterns.
- Hierarchical clustering approaches such as group average are promising techniques to use for pattern preserving clustering.
  - Group average never splits groups of points.

HICAP Algorithm

- Find maximal hyperclique patterns
  - Non-maximal hypercliques will tend to be absorbed by their corresponding maximal hyperclique pattern and not affect the clustering process.
  - Thus, using all hyperclique patterns would cause a great deal of overhead with little if any gain.
- Perform a group average hierarchical clustering
  - The starting clusters are hyperclique patterns and the points not covered by hyperclique patterns.
  - Except for the starting point, the clustering algorithm is the same as the group average approach.
  - Since hypercliques are overlapping, resulting clustering may also be overlapping.

Overview

- Introduction
- The Hyperclique Pattern
- HICAP: Approach and Algorithm

$\Rightarrow$ Experimental Results

- Conclusions and Future Work

Experimental Setup

- Data Sets

Data Set	LA1	RE0	WAP
#Documents	3204	1504	1560
#Words	31472	11465	8460
#Classes	6	13	20
Min Class Size	273	11	5
Max Class Size	943	608	341
Min/Max Ratio of Class Size	0.29	0.018	0.015
Source	TREC-5	Reuters	WebAce

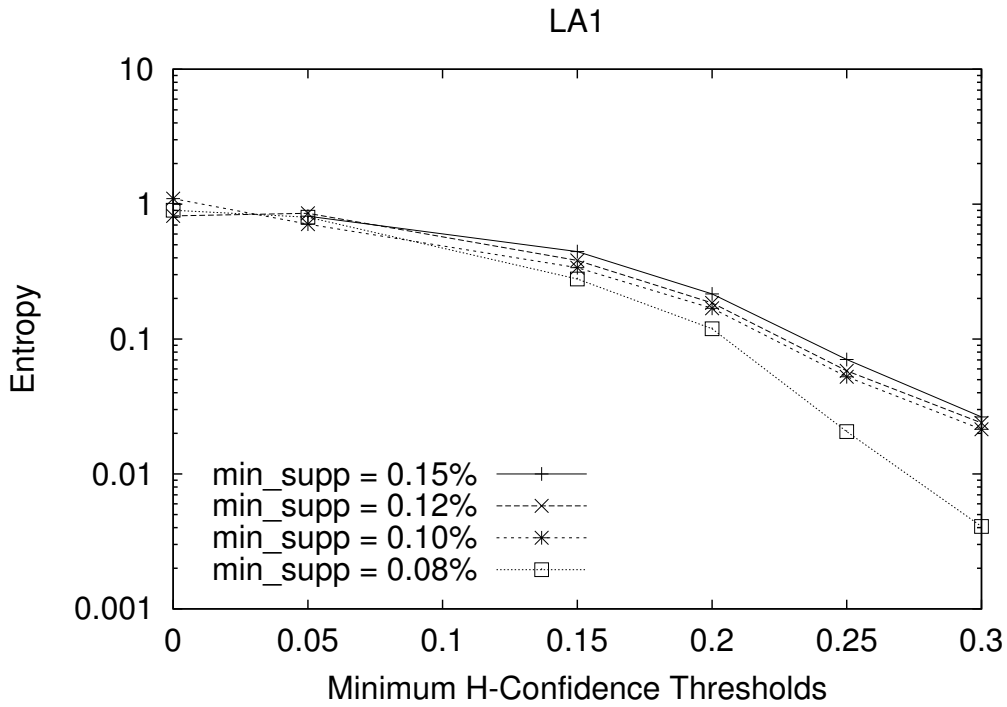
- Entropy:

$$E_j = -\sum_i p_{ij} log(p_{ij}) \quad and \quad E = \sum_{j=1}^{m} \frac{n_j}{n} * E_j$$

p_{ij} is the probability that a member of cluster j belongs to class i.

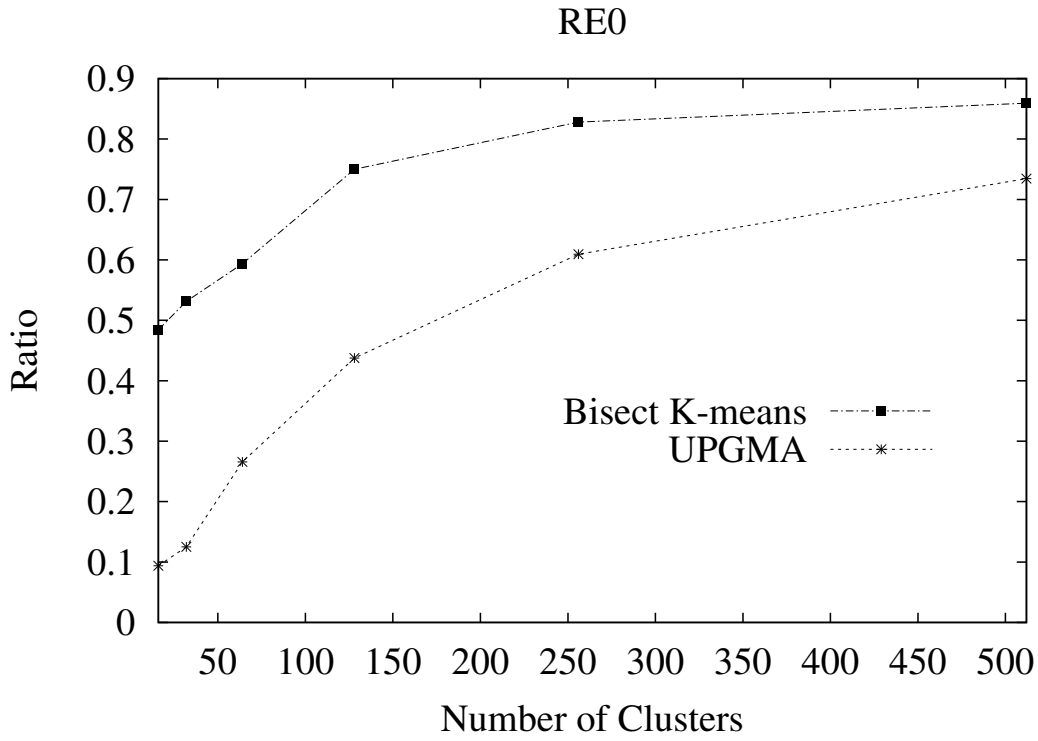
Experimental Results: Entropy of Hypercliques

- Hypercliques tend to contain documents of the same class.
- Entropy of frequent patterns is high.



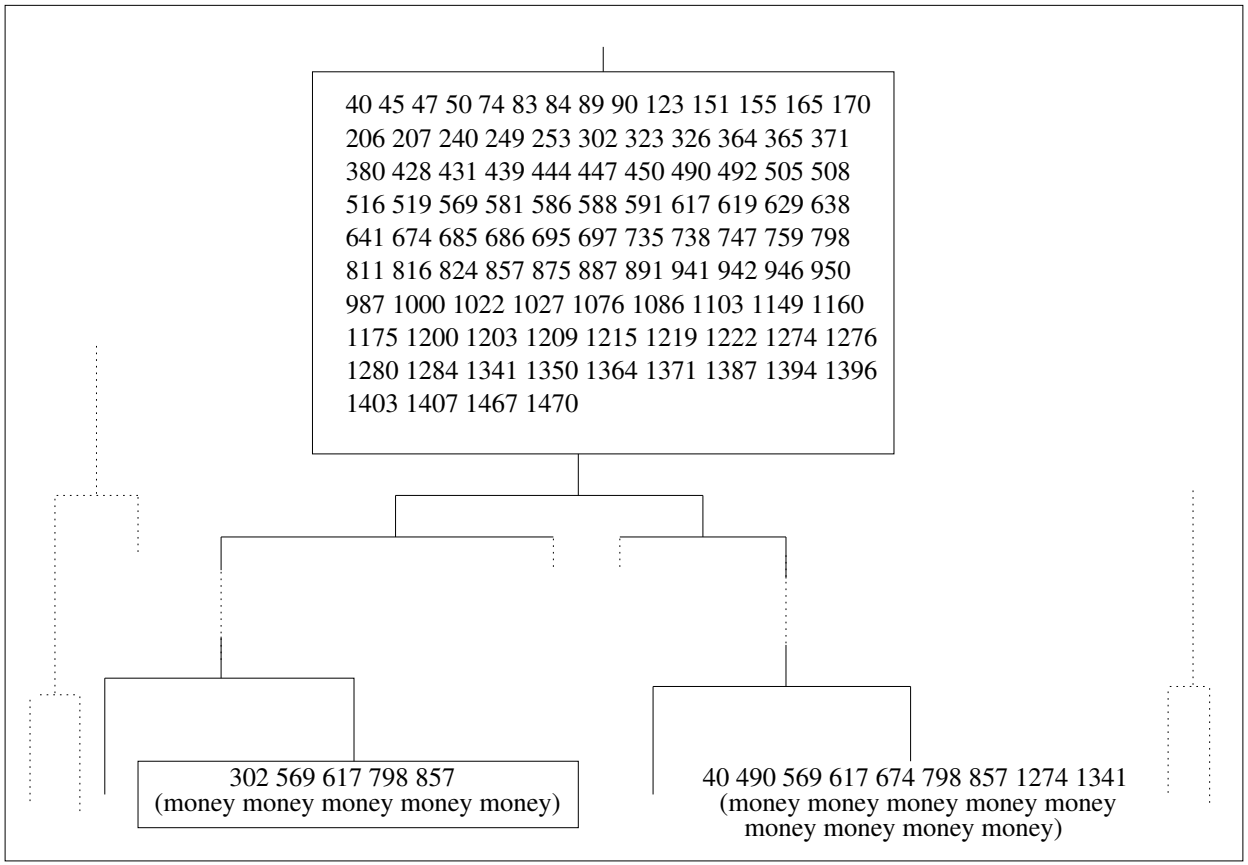
Experimental Results: Pattern Splitting

- HICAP does not split patterns.
- Ratio: the percentage of hyperclique patterns being split.



Analyzing the Nature of Clusters

- Two pure hyperclique patterns are traced (Cluster from RE0).



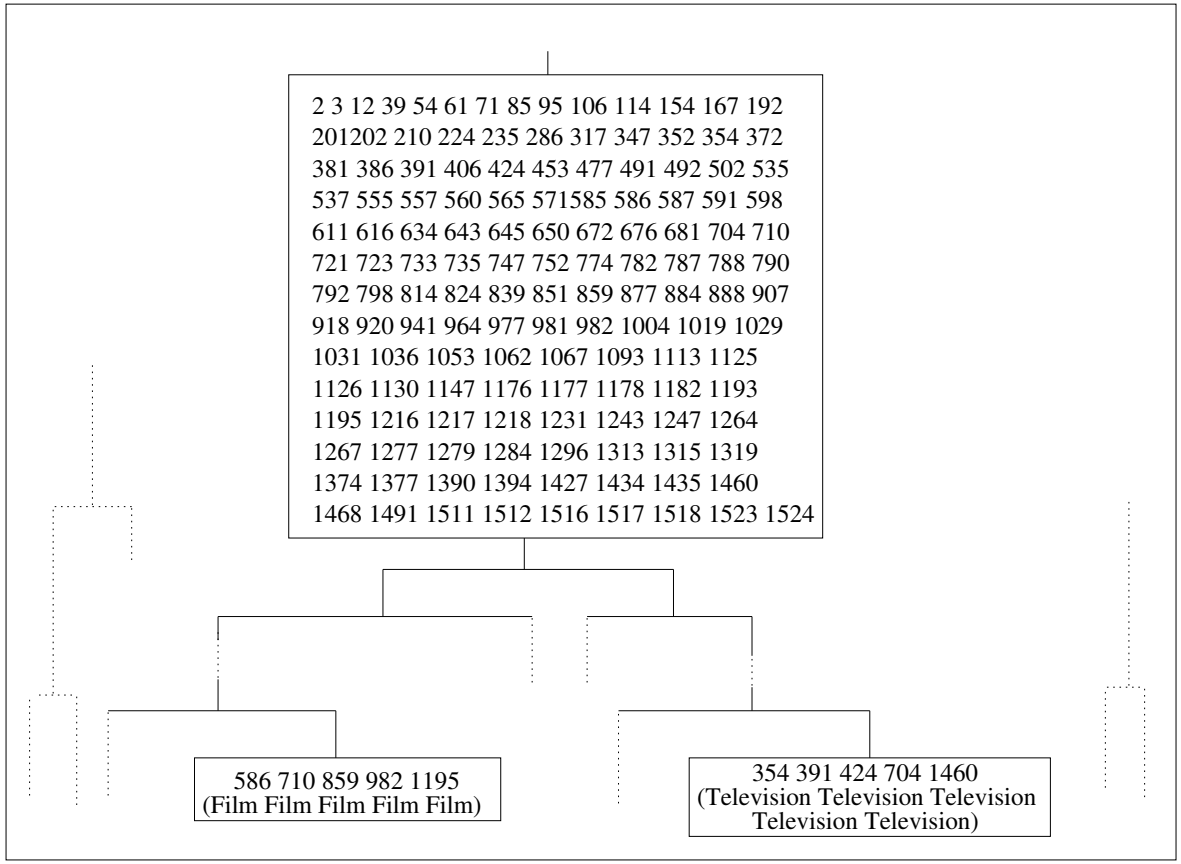
Analyzing the Nature of Clusters ...

- The cluster is also relatively pure.

money money money money money money money money
money money money money money money money money money
money money money money money money money money money
money money money money money money money money money
money money money money money money money money money
interest money money **interest** money money money money
money money money money money money money money money
money money money money money money money money money
money money money money money money money money money
money money money money money money money money money
money money money money money money money money money
money money money money

Analyzing the Nature of Clusters ...

- Two hyperclique patterns are pure, but different (Cluster from WAP).



Analyzing the Nature of Clusters ...

- The cluster is also mixed.

Television Television Television Film Film Television Stage Television Television Television Film Cable Television Television Television Variety Film Television Film Television Stage Television Film Television Film Television Film Television Film Television Television Film Cable Television Stage Film Television People Television Film People Television Cable Film Television Television Media Stage Television Film Television Film Television Television Stage Film Television Film Television Television Stage Film Television Film Television Television Film Film Television Television Cable Television Television People Television Film Television Film Television Television Film Television Variety Variety Television Film Film Cable Film Television Television Film Television Television Film Television Television Television Film Film Television Film Film Television Television Television Film Television Television Television Film Television Film Television Film Television Film Television Film Variety Film Television Film Industry Television Film Television Art Television Television Film Media Industry Stage Television Television Television Television Television

Analyzing the Nature of Clusters ...

- Some statistics from WAP (covered 808 out of 1560 documents).

CNo	size	#unmatch	#hyperclique	Classes of hypercliques
1	49	5	16	People/Online
2	66	0	9	Sports
3	169	59	10	Business/Tech/Politics
4	313	2	49	Health
5	33	3	4	Film
6	61	9	9	Politics
7	18	0	7	Culture
8	44	2	1	Television
9	25	0	7	Sports
10	22	4	1	People
11	8	0	2	Television/Stage
Total	808	84	115	

Overview

- Introduction
- The Hyperclique Pattern
- HICAP: Approach and Algorithm
- Experimental Results

$\Rightarrow$ Conclusions and Future Work

Conclusions and Future Work

- Conclusions
  - HICAP is a pattern preserving clustering technique based on the hyperclique pattern and the group average clustering approach.
  - A key benefit of pattern preserving clustering lies in aiding cluster interpretation
- Future Work
  - Investigating whether other patterns can be used for pattern preserving clustering.
    - What properties of patterns are needed for meaningful results?
  - Applying HICAP to additional fields.
  - Extending pattern preserving clustering to other types of clustering algorithms, such as K-means.

Questions?

- Personal Homepage - http://www.cs.umn.edu/∼huix

Thank You !