HICAP
HIerarchial **C**lustering with **PA**ttern **P**reservation

RIT2013061 Soumya Agrawal

RIT2013062 Gaurav Goyal

RIT2013063 Vamsi Sangam

RIT2013064 Kumar Abhishek

RIT2013086 Viswakanth Korada

Problem statement

To introduce a new approach in clustering algorithms, which is **HI**erarchial **C**lustering with P**A**ttern **P**reservation (HICAP) .

Hyperclique Pattern

- Hyperclique pattern is a type of association pattern that contains items that are highly affiliated with each other. By high affiliation, we mean that the presence of an item in a transaction strongly implies the presence of every other item that belongs to the same hyperclique pattern.
- h-confidence measure is specifically designing need to measure the strength of association .

The **h-confidence** of the itemset P = $\{i_1, i_2, \ldots i_n\}$ denoted as hconf(P), is a measure that reflects the overall affinity among items within the itemset. This measure is defined as:

$$hconf(P) = \mathbf{min}\ \{ conf\ \{ i_1 \rightarrow i_2, i_3 \ldots, i_n \}, conf\ \{ i_2 \rightarrow i_1, i_3, \ldots i_n \} \ldots, conf\ \{ i_n \rightarrow i_1, i_2, \ldots i_{n-1} \}\},$$

where conf is the conventional definition of association rule confidence.

- Given a transaction database and the set of all item set $I = \{ i_1, i_2, \ldots i_n \}$ of an item set P is a hyperclique pattern if and only if

  1. $P \subseteq I$ *and* $|P| > 0$.
  2. $hconf(P) \geq h_c$, *where* h_c *is the minimum h-confidence threshold.*

- A Hyperclique pattern is a **maximal** hyperclique pattern if no superset of this pattern is a hyperclique pattern.

ALGORITHM

- The algorithm consist of two phases :

1st **phase**: HICAP finds maximal hyperclique pattern which we want to preserve in HICAP algorithm.

2nd **phase** : HICAP conducts hierarchical clustering and output the clustering results.

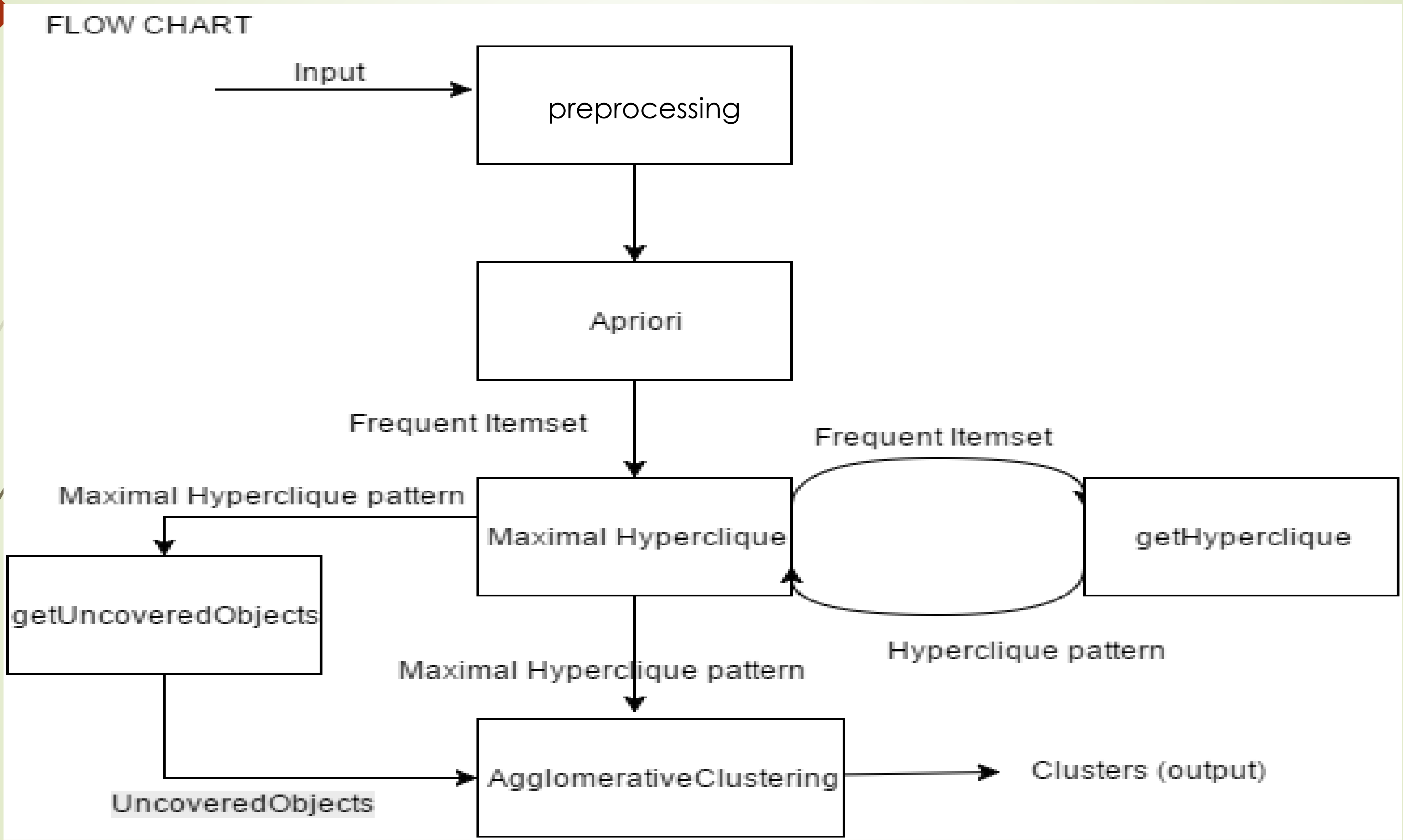
Maximal hyperclique patterns cover only 10% - 20 % of all objects and thus HICAP also includes uncovered objects as a separate initial cluster.

Finally the similarity between the cluster is calculated using average of the pairwise cosine similarity.

```
HICAP Algorithm
Input:      D: a document data set.
            θ: a minimum h-confidence threshold.
            α: a minimum support threshold.
Output:     CR: the hierarchical clustering result.
Variables:  S: the hyperclique pattern set.
            MS: the maximal hyperclique pattern set.
            PD: The output set of preprocessing
            LS: a set of objects which are not covered by
                identified maximal hyperclique patterns
            CS: a set containing target clustering objects
Method
Phase I: Maximum Hyperclique Dattern Discovery
 1.         S = hyperclique_miner(θ, α, D)
 2.         MS = maximal_hyperclique_pattern(S)

Phase II: Hierarchical Clustering
 3.         PD = preprocessing(D)
 4.         LS = uncovered_objects(MS, D)
 5.         CS = LS ∪ MS
 6.         for i=1 to |CS|-1
 7.             find the pair of elements with max group
                average cosine value from the set CS,
 8.             merge the identified pair, and update
                CS and CR accordingly
 9.         endfor
 10.        OUTPUT CR
 11.        End
```

HICAP


FLOW CHART
Input
preprocessing
Apriori
Frequent Itemset
Frequent Itemset
Maximal Hyperclique pattern
Maximal Hyperclique
getHyperclique
getUncoveredObjects
Hyperclique pattern
Maximal Hyperclique pattern
AgglomerativeClustering
Clusters (output)
UncoveredObjects

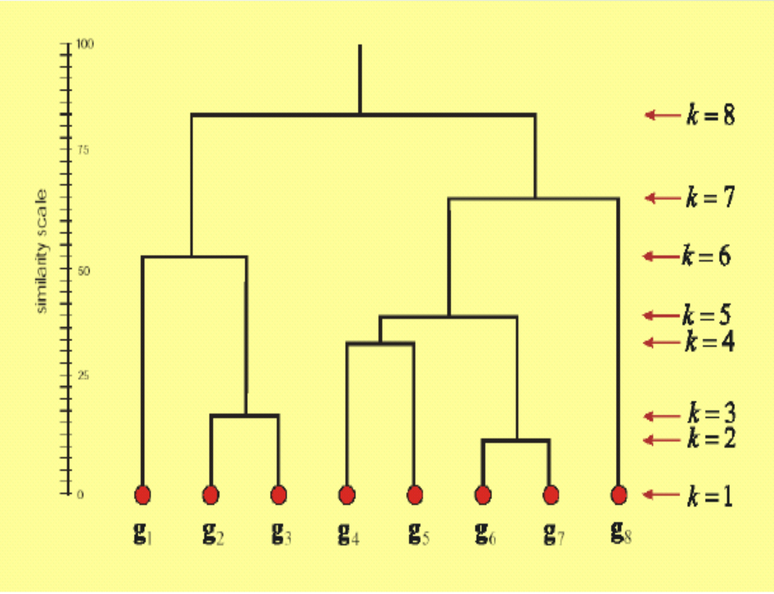
Hyperclique pattern v/s Frequent Itemset

- Hyperclique pattern include objects which are strongly similar to each other with respect to the cosine measure ,in contrast many pair of objects from a frequent item set may have very poor cosine measure.
- Hyperclique pattern have better performance at low level of support than the frequent item set .

Finally , the size of the maximal hyperclique pattern is significantly smaller than the size of the maximal frequent item sets.

Hierarchical clustering

- There are two styles of hierarchical clustering algorithms to build a tree from the input set S:
  - **Agglomerative (bottom-up)**:
    - Beginning with singletons (sets with 1 element)
    - Merging them until S is achieved as the root.
    - It is the most common approach.
  - **Divisive (top-down)**:
    - Recursively partitioning S until singleton sets are reached.



DataSet

Data	
# Transactions in Input Data	9835
# Columns in Input Data	32
# Items in Input Data	169

- There are 9835 transaction records. There were atmost 32 items purchased on one of its transactions. The total number of unique items is 169.

1	citrus fruit	semi-finished bread	margarine	ready soups	
2	tropical fruit	yogurt	coffee		
3	whole milk				
4	pip fruit	yogurt	cream cheese	meat spreads	
5	other vegetables	whole milk	condensed milk	long life bakery product	
6	whole milk	butter	yogurt	rice	abrasive cleaner
7	rolls/buns				
8	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)
9	pot plants				
10	whole milk	cereals			
11	tropical fruit	other vegetables	white bread	bottled water	chocolate
12	citrus fruit	tropical fruit	whole milk	butter	curd
13	beef				
14	frankfurter	rolls/buns	soda		
15	chicken	tropical fruit			
16	butter	sugar	fruit/vegetable juice	newspapers	
17	fruit/vegetable juice				
18	packaged fruit/vegetables				
19	chocolate				
20	specialty bar				
21	other vegetables				

Thank you…