Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

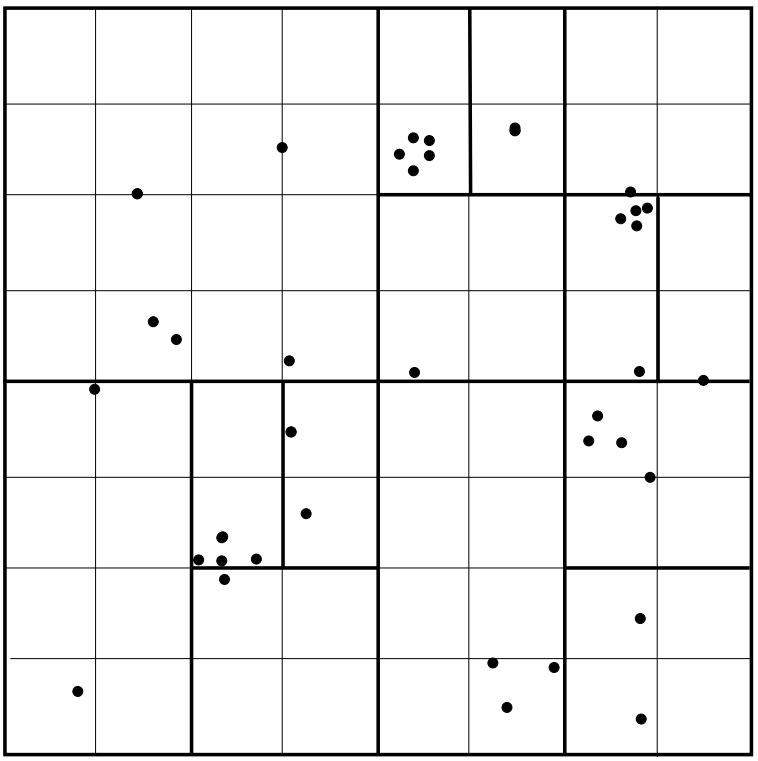
4.12.2008

6. Übung „Knowledge Discovery“

Wintersemester 2008/2009

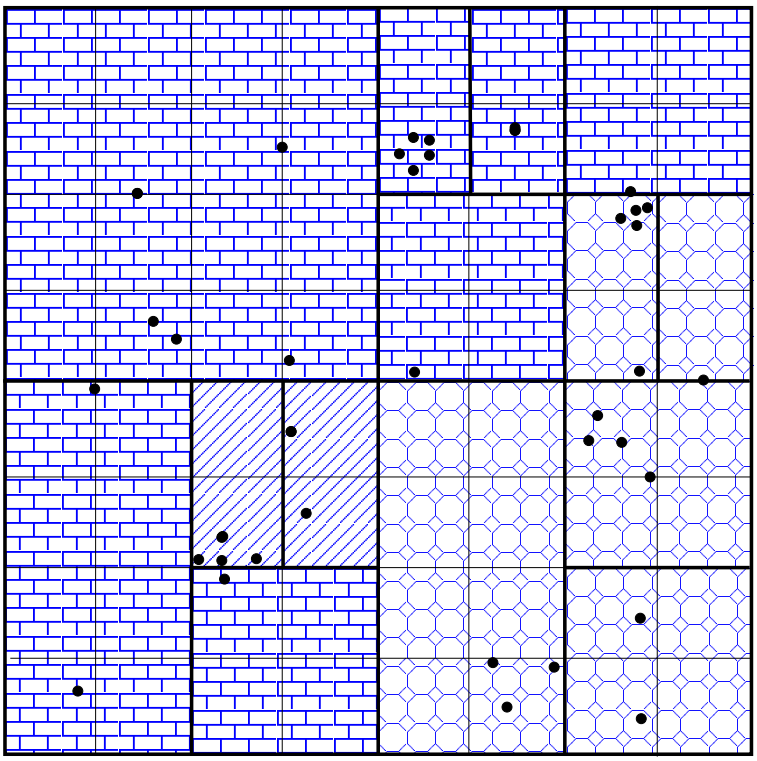
1 GRID-Clustering

Die folgende Abbildung zeigt Punkte der Ebene, die mittels eines *Gridfiles* mit einer Anzahl von maximal $n = 5$ Punkten pro Cluster vorgeclustert wurden (die dünnen Rasterlinien dienen nur der Orientierung).



Wenden Sie die GRID-Clustering-Methode auf den obigen Datensatz an.

Eine mögliche Clusterung stellt die folgende Abbildung dar:



Es sind jedoch auch andere Clusterungen möglich - dies hängt von der Wahl des Startclusters sowie der Vereinigungs-Strategie ab.

2 Datenkompression zum Vor-Clustering

Erzeugen Sie einen CF-Baum aus den Datenpunkten der folgenden Tabelle. Nutzen Sie die Parameter $B = 2, T = 1.5$ und $L = 2$.

1	5	1.5	9	6	6.2	2

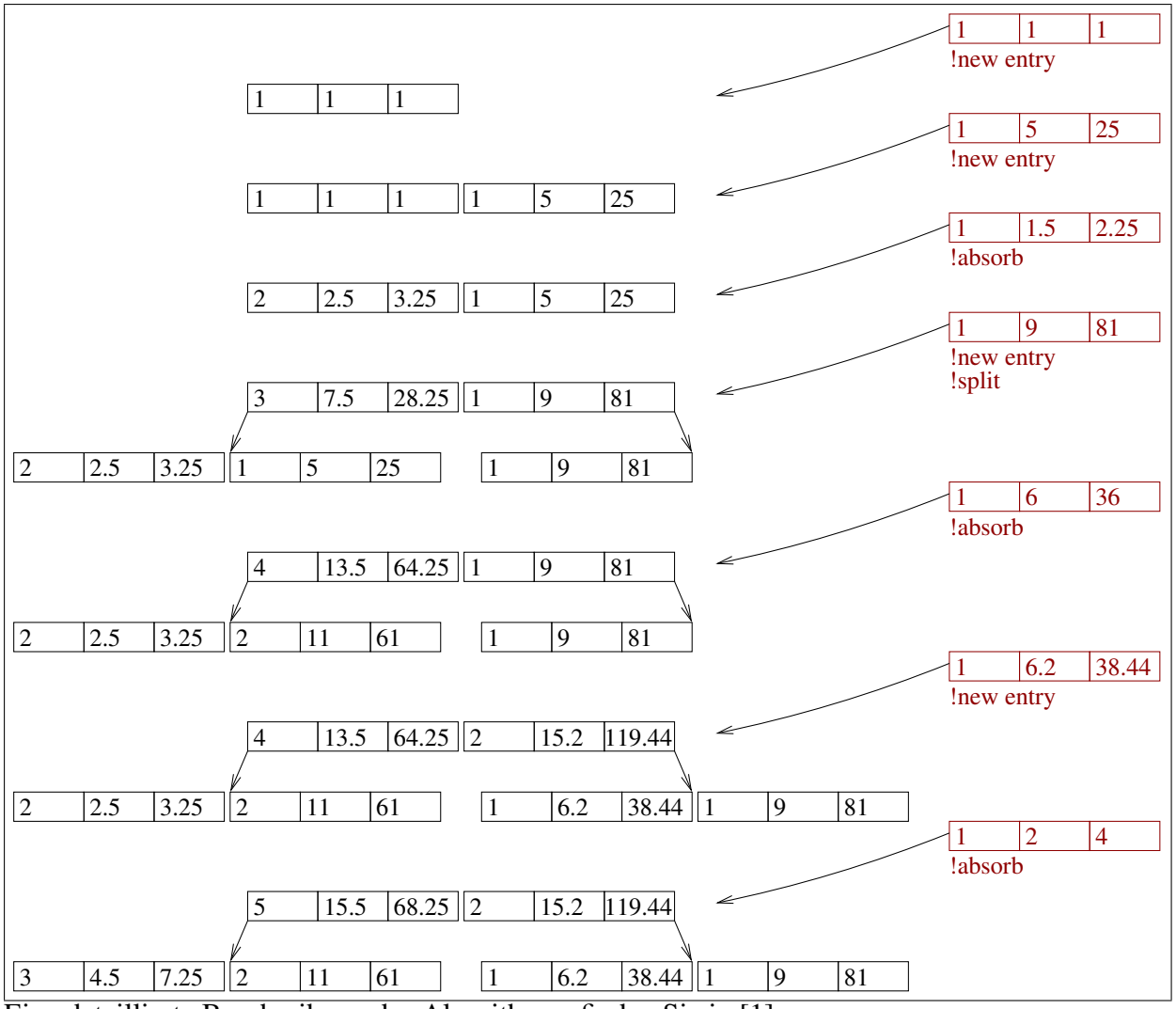
Berechnen Sie den Centroid X_0 für ein Cluster C_i mit dem Clustering-Feature $CF_i = (N, LS, SS)$ mittels der Formel

$$X_0 = \frac{1}{N} LS$$

und den Radius R_i des Clusters mittels

$$R_i = \sqrt{\frac{1}{N}(SS - NX_0^2)}.$$

Dann ergibt sich folgender Aufbau des CF-Trees (dabei wurde als Threshold der Radius $T/2$ benutzt):



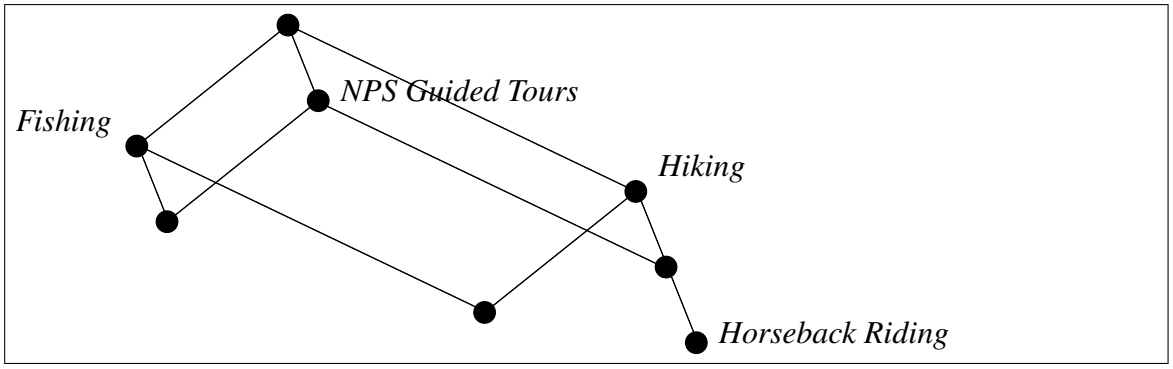
Eine detaillierte Beschreibung des Algorithmus finden Sie in [1].

3 Formale Begriffsanalyse

1. Betrachten Sie das Diagramm des Begriffsverbandes auf Folie 11 im 3. Kapitel, 3. Teil der Vorlesung. Wie lautet der Begriff der zu dem Knoten oberhalb des Knotens „Death Valley“ gehört?

   ({ Death Valley, Devils Postpile, Golden Gate, Lassen Volcanic, Point Reyes, Redwood, Santa Monica Mts, Whiskeytown-Shasta-Trinity, Yosemite }, { Swimming, Hiking, NPS Guided Tours, Horseback Riding })

2. Beschriften Sie die Knoten des Diagramms mit den zugehörigen Support-Werten und zeichnen Sie das Diagramm der Knoten mit mehr als $10/19$ Support.



3. Bestimmen Sie im folgenden Kontext alle Begriffe, indem Sie zu jeder Teilmenge $B \subseteq M$ das Paar (B', B'') bilden.

I	jung	alt	weiblich	männlich
Vater		X		X
Mutter		X	X	
Tochter	X		X	
Sohn	X			X

Die Begriffe lauten:
({}, {jung, alt, weiblich, männlich}),
({Vater}, {alt, männlich}), ({Mutter}, {alt, weiblich}), ({Tochter}, {jung, weiblich}), ({Sohn}, {jung, männlich}),
({Vater, Sohn}, {männlich}), ({Vater, Mutter}, {alt}), ({Mutter, Tochter}, {weiblich}), ({Tochter, Sohn}, {jung}),
({Vater, Mutter, Tochter, Sohn},{})

Literatur

[1] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD'96)*, pages 103–114, 1996. http://citeseer.ist.psu.edu/zhang96birch.html.