Evaluierung von Algorithmen der MajorClust-Familie

Denis Kreis

Web Technology & Information Systems
Bauhaus-Universität Weimar

7. Mai 2009

Outline

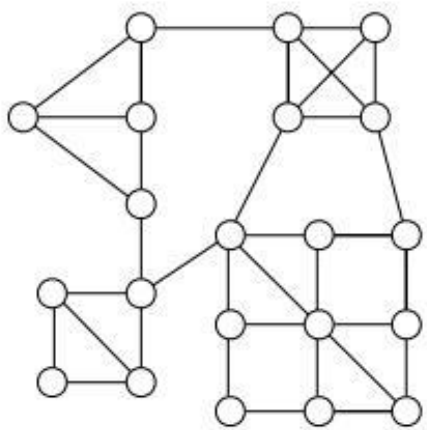
Outline

Clusteranalyse

Clusteranalyse ist ein strukturentdeckendes Verfahren zur Ermittlung von Gruppen (Clustern) von Objekten, die Ähnlichkeiten zueinander aufweisen.

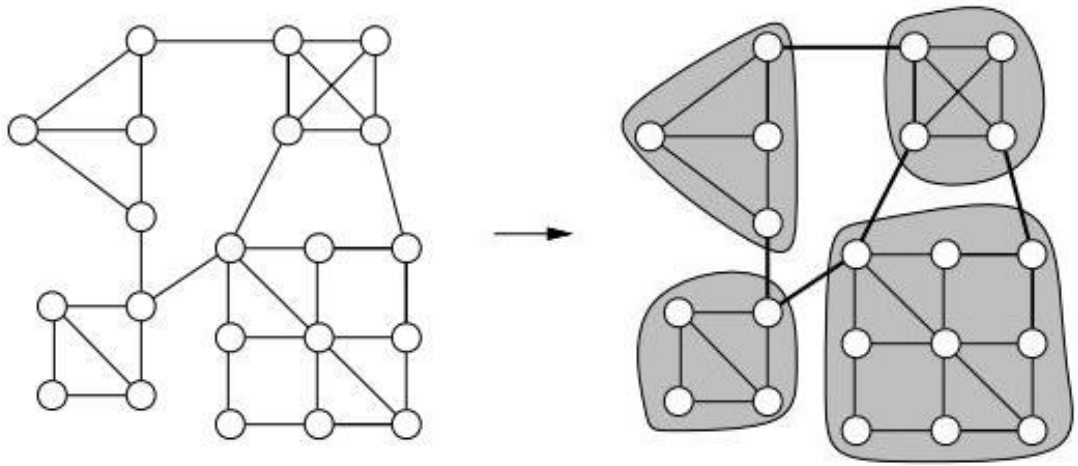
Clusteranalyse

Clusteranalyse ist ein strukturentdeckendes Verfahren zur Ermittlung von Gruppen (Clustern) von Objekten, die Ähnlichkeiten zueinander aufweisen.



Clusteranalyse

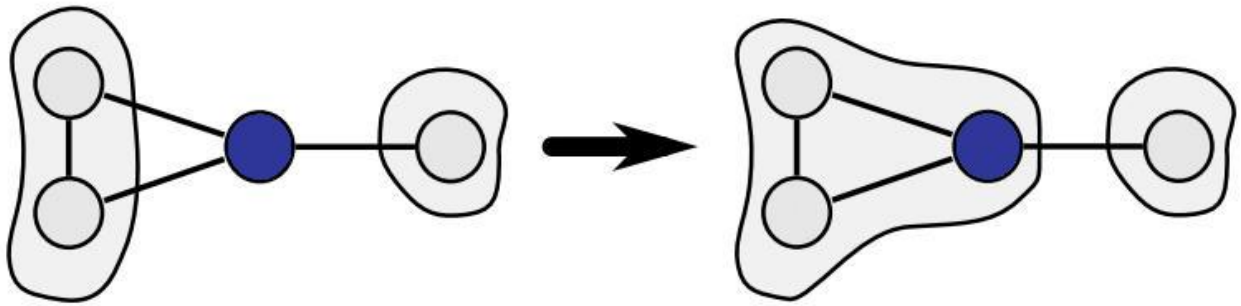
Clusteranalyse ist ein strukturentdeckendes Verfahren zur Ermittlung von Gruppen (Clustern) von Objekten, die Ähnlichkeiten zueinander aufweisen.



Outline

MajorClust

Gegeben: Graph $G = (V, E, \omega)$



Zuweisung eines Knotens zu Clustern:

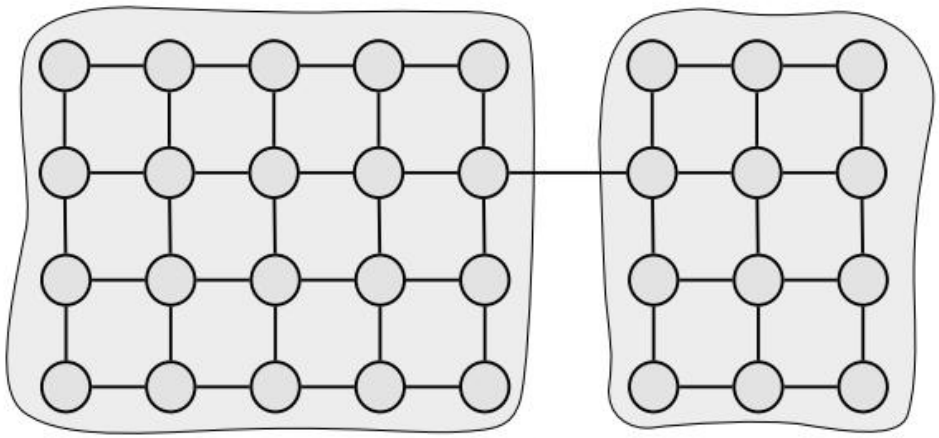
$$\forall v \in V : \; c^*(v) = \arg \max_{i:i\in\{1,\ldots,|V|\}} \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u, v)$$

Terminiert, wenn in einem Durchgang keine Zuordnung stattfindet.

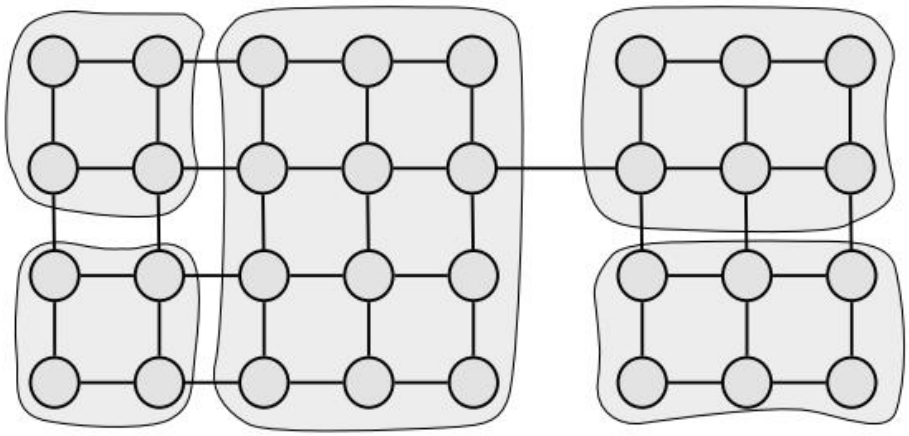
MajorClust

Nachteil:

MajorClust liefert bei homogenen Graphen zu feine Partition.



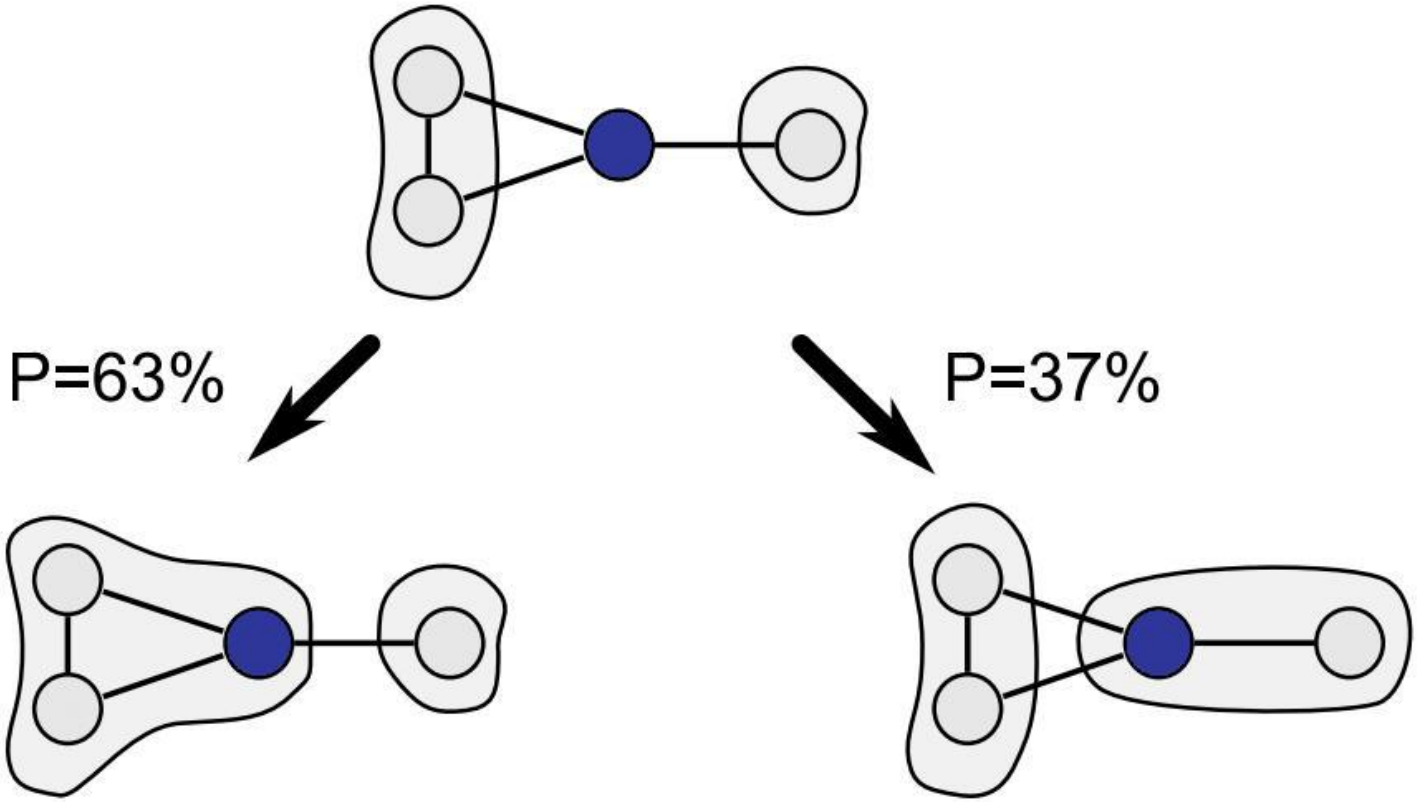
Optimal Clustering



MajorClust

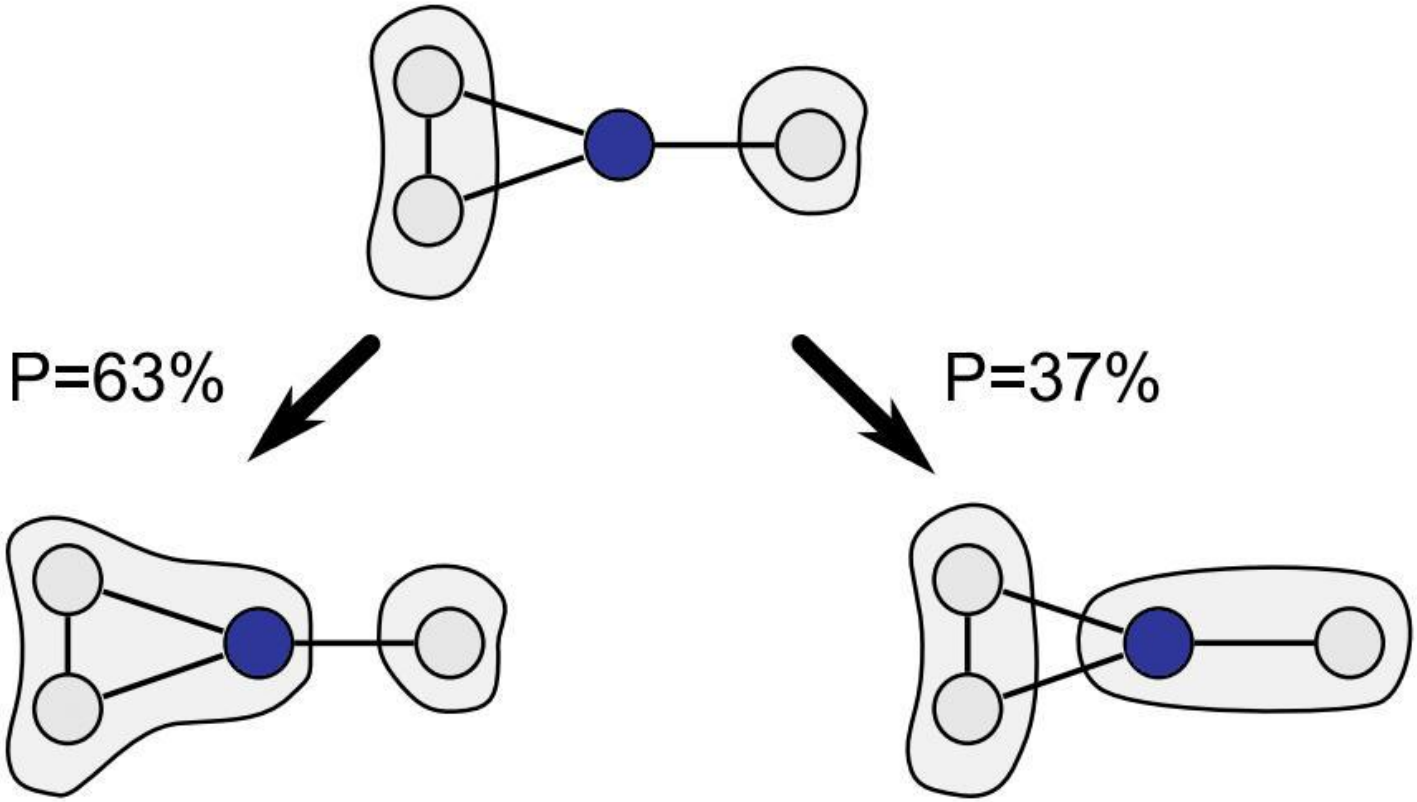
MCProb

Führt eine Randomisierung ein, nach der die Knotenzuordnung stattfindet.



MCProb

Führt eine Randomisierung ein, nach der die Knotenzuordnung stattfindet.



Unterschiedliche Wahrscheinlichkeitsverteilungen sind möglich!

MCProb

Problem: Finden eines geeigneten Abbruchkriteriums

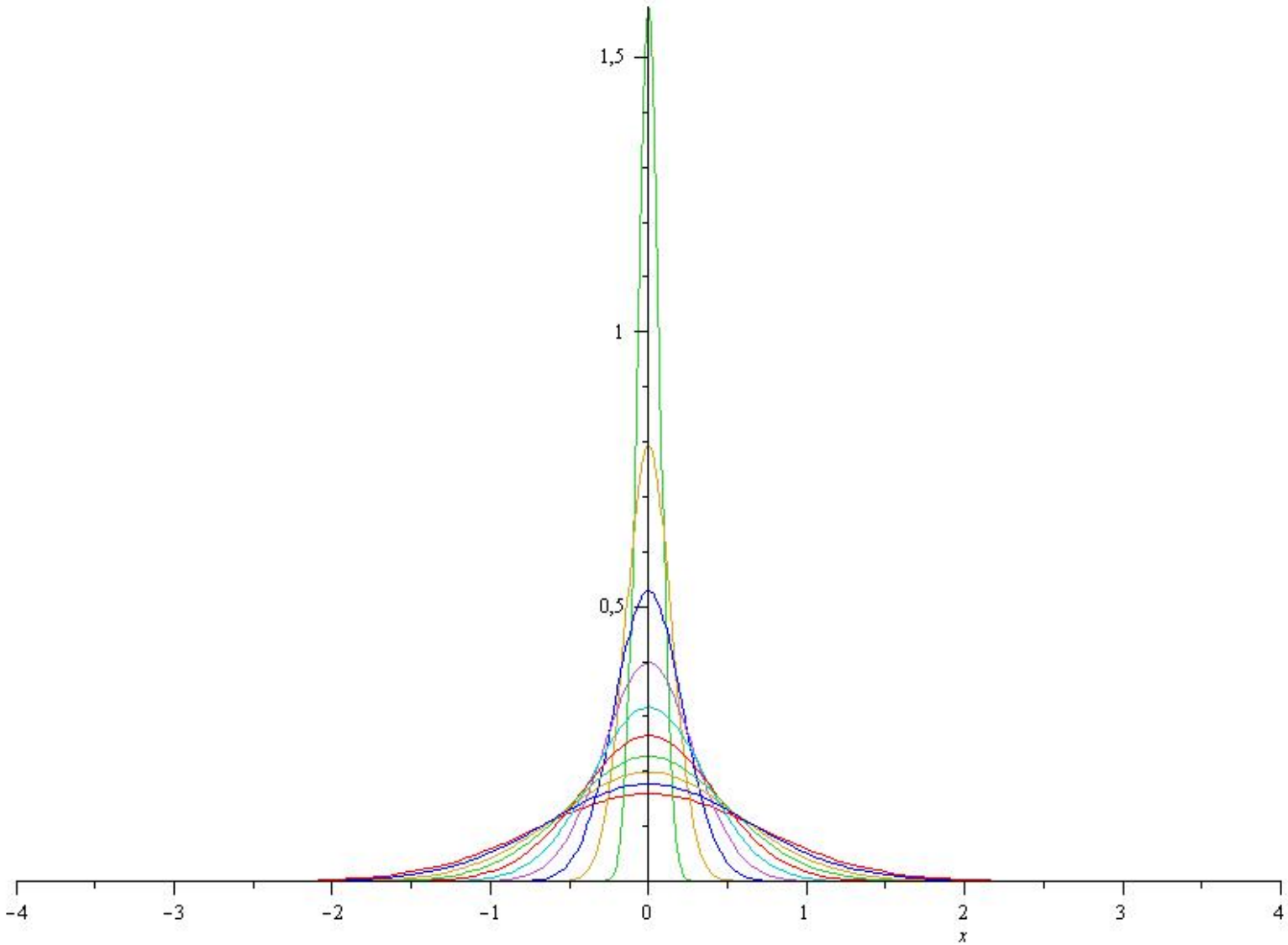
MCProb

Problem: Finden eines geeigneten Abbruchkriteriums

ClusterCounter: Terminiere, wenn die Anzahl an Cluster sich in einem Durchgang nicht geändert hat.

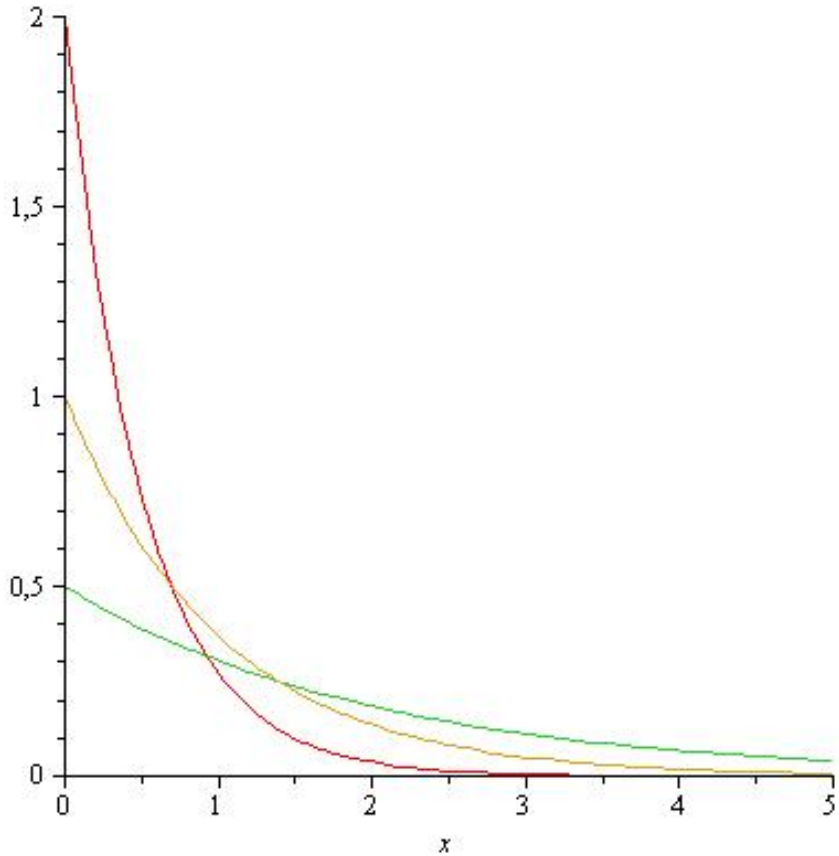
MCProb - Dynamisches Abbruchkriterium

Normalverteilung:



MCProb - Dynamisches Abbruchkriterium

Exponentialverteilung:



BalancedMCProb

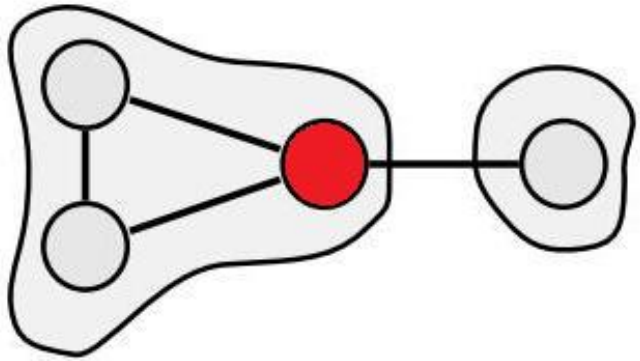
Idee: Wahrscheinlichkeit für Mehrheitsentscheidung anpassen.

Implementiert:

- ClusterCounter (CC)
- Dynamisch

StrongMajorClust

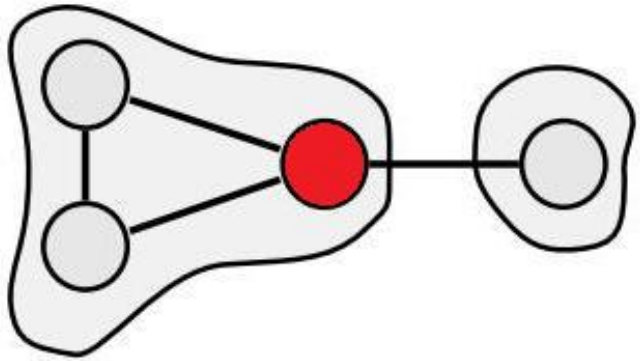
Führt ein Verstärker K zur Steuerung der Verkettungstendez ein.



$$\forall v \in V : c^* = \arg \max_{i:i\in\{1,\ldots,|V|\}} \begin{cases} \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i \neq c(v) \\ \frac{1}{K} \cdot \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i = c(v) \end{cases}$$

StrongMajorClust

Führt ein Verstärker K zur Steuerung der Verkettungstendez ein.

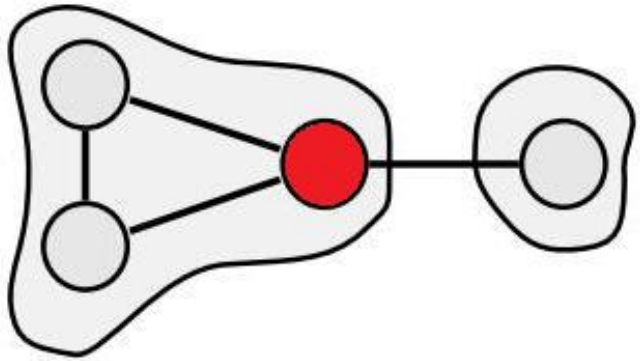


$$\forall v \in V : c^* = \arg \max_{i:i\in\{1,\ldots,|V|\}} \begin{cases} \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i \neq c(v) \\ \frac{1}{K} \cdot \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i = c(v) \end{cases}$$

Problem: Abbruchkriterium

StrongMajorClust

Führt ein Verstärker K zur Steuerung der Verkettungstendez ein.



$$\forall v \in V : c^* = \arg \max_{i:i\in\{1,\ldots,|V|\}} \begin{cases} \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i \neq c(v) \\ \frac{1}{K} \cdot \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v) & i = c(v) \end{cases}$$

Problem: Abbruchkriterium

Lösung:

- ClusterCounter (CC)

ExtendedMajorClust

$$\forall v \in V : \; c^* = \arg \max_{i:i\in\{1,\ldots,|V|\}} \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=i} \omega(u,v)$$

Führt zusätzlich eine Iteration auf Cluster-Ebene ein:

$\forall l, m \in c(V)$:

wenn

$$\sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=c(v)=l} \omega(u,v) < \sum_{\{u,v\}:\{u,v\}\in E \wedge c(u)=l \wedge c(v)=m} \omega(u,v)$$

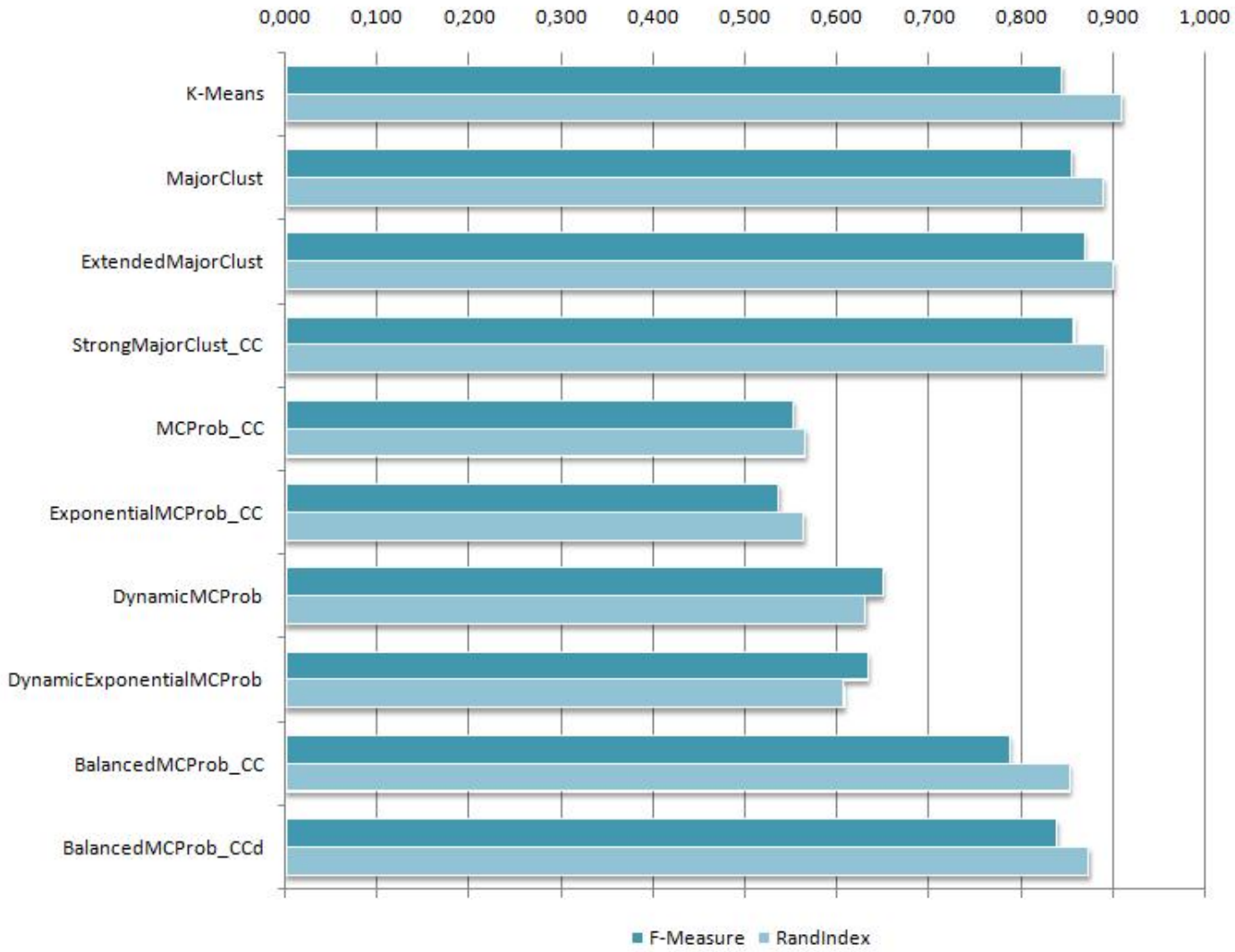
dann vereinige die Cluster l und m

Outline

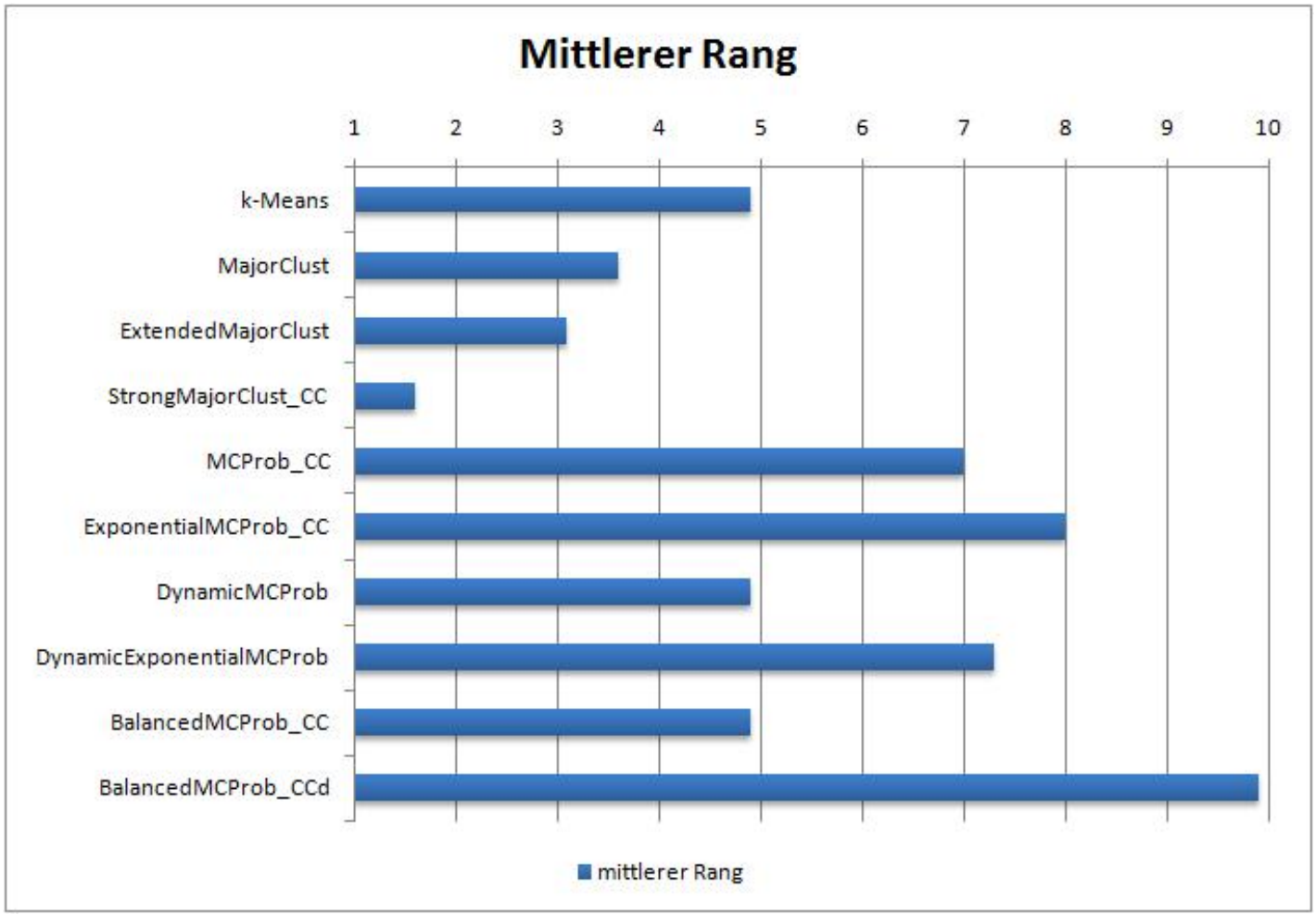
Versuchsumgebung

- Sieben Testkollektionen, 120 - 3000 Dokumente, basierend auf Reuters Collection (RCV1)
- BagOfWords- bzw. Vektorraummodell
- Kosinusähnlichkeit
- Graphausdünnung mit Harmonic Expected Similarity

Ergebnisse (gemittelt)



Laufzeit


Mittlerer Rang
1
2
3
4
5
6
7
8
9
10
k-Means
MajorClust
ExtendedMajorClust
StrongMajorClust_CC
MCProb_CC
ExponentialMCProb_CC
DynamicMCProb
DynamicExponentialMCProb
BalancedMCProb_CC
BalancedMCProb_CCd
mittlerer Rang

Outline

Zusammenfassung

- ExtendedMajorClust liefert die besten Ergebnisse!
- Probabilistische Varianten verbessern das Ergebnis nicht.
- Implementationen mit dem dynamischen Abbruchkriterium liefern bessere Ergebnisse als die mit CC.