

# Information Retrieval

## Clusteranalyse

Stefan Birkner

20. Oktober 2010

# Clusteranalyse



3 Cluster

# Clusteranalyse

## Anwendungen

- ▶ Gruppierung von Suchergebnissen
- ▶ Anreicherung von kleinen Ergebnislisten
- ▶ Automatische Thesauruserstellung













# Clusteranalyse

## Beispiel für Gruppierung von Suchergebnissen: yippy.com

The screenshot shows a search interface with a blue header containing 'clouds', 'sources', and 'sites'. Below the header, there are several filter categories:

- All Results (192) with a 'remix' button
- Manufacturer, Garments (17)
- Photos (20)
- Ancient Egypt (35)** (highlighted in red)
- Gallery, Original (2)
- Kids (6)
- Books (2)
- Egypt Travel (5)
- Pyramids (6)
- History (8)
- Modern (3)

Cluster **Ancient Egypt** contains **35** document:

1. [The Light of Egypt](#)     
Small gallery features facsimiles of origin  
[www.light-of-egypt.co.uk](http://www.light-of-egypt.co.uk) - [cache] - Open
2. [Art Gallery of Nile Arts](#)     
An art gallery with original watercolor pain  
[www.nilearts.com](http://www.nilearts.com) - [cache] - Open Direct
3. [Archaeologia](#)     
Sells rare and out-of-print archaeology bo  
[www.archaeologia.com](http://www.archaeologia.com) - [cache] - Open E
4. [Dakota Products](#)     
Aromatherapy is the practiced use of frag  
practiced widely.  
[www.dakotaproducts.com](http://www.dakotaproducts.com) - [cache] - Ope

# Clusteranalyse

## Eigenschaften einer guten Clusteranalyse

- ▶ Dokumente eines Clusters sind einander sehr ähnlich
- ▶ Dokumente unterschiedlicher Cluster sind einander kaum ähnlich

# Clusteranalyse

## Ähnlichkeitsmaß

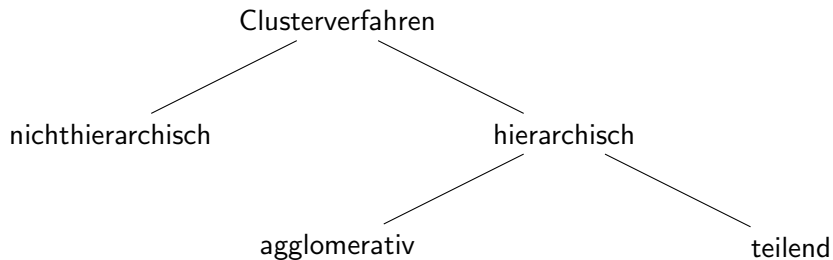
- ▶ Funktion, die für zwei Dokumente eine Zahl berechnet
- ▶ desto niedriger die Zahl, desto ähnlicher die Dokumente
- ▶ Beispiele: Dice-Koeffizient

# Clusteranalyse

## Ähnlichkeitsmaß

- ▶ Funktion, die für zwei Dokumente eine Zahl berechnet
- ▶ die Größe der Zahl gibt an, wie ähnlich sich die Dokumente sind
- ▶ Beispiele: Dice-Koeffizient

# Clusteranalyse





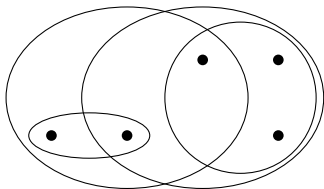
# Clusteranalyse

## Star-Technik

- ▶ nichthierarchisches Verfahren
- ▶ Für jedes Dokument wird ein Cluster erstellt, das alle Dokumente enthält, deren Ähnlichkeit unter einem frei zu wählenden Schwellwert liegt.

# Clusteranalyse

Beispiel für die Star-Technik



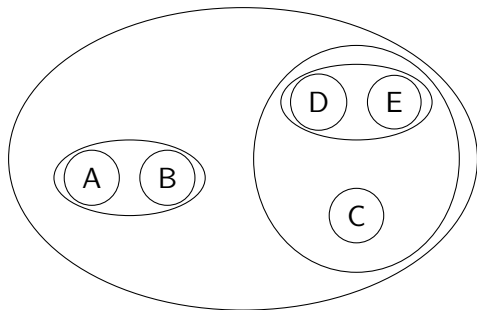
# Clusteranalyse

Hierarchische Verfahren: Agglomerative Verfahren

- ▶ Zu Beginn bildet jedes Dokument ein Cluster
- ▶ In jedem Folgeschritt werden jeweil zwei Cluster zu einem vereinigt.

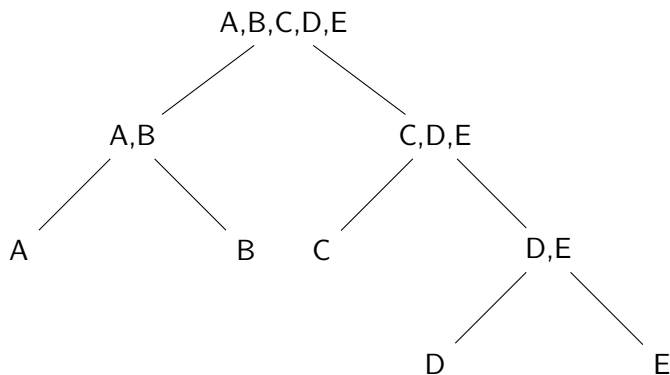
# Clusteranalyse

Beispiel für durch ein hierarchisches Verfahren berechnete Cluster



# Clusteranalyse

Dendrogramm zum vorhergehenden Beispiel



# Clusteranalyse

## Single-Link-Verfahren

- ▶ agglomeratives Verfahren
- ▶ Führe in jedem Schritt die zwei Cluster mit den ähnlichsten Elementen zusammen.

# Clusteranalyse

## Nachteile des Single-Link-Verfahrens

- ▶ Es können sich lange kettenförmige Cluster bilden.
- ▶ Uneinheitliche Clustergröße, da große Cluster dazu neigen, sich zu vergrößern.

# Clusteranalyse

## Complete-Link-Verfahren

- ▶ agglomeratives Verfahren
- ▶ Die Ähnlichkeit zweier Cluster entspricht der Ähnlichkeit der beiden Dokumente, die sich am unähnlichsten sind.
- ▶ Führe die beiden Cluster mit der größten Ähnlichkeit zusammen.



# Clusteranalyse

## Nachteile des Complete-Link-Verfahrens

- ▶ Ausreißer verschlechtern teilweise die Qualität der Cluster.

# Clusteranalyse

## Hierarchische Verfahren: Teilende Verfahren

- ▶ Zu Beginn gibt es ein Cluster, das alle Dokumente enthält.
- ▶ In jedem Folgeschritt wird ein Cluster in zwei Cluster aufgeteilt.

# Literatur

- ▶ Andreas Henrich: Information Retrieval 1
- ▶ [http://www.uni-bamberg.de/minf/ir1\\_buch/](http://www.uni-bamberg.de/minf/ir1_buch/)
- ▶ Kapitel 6.2