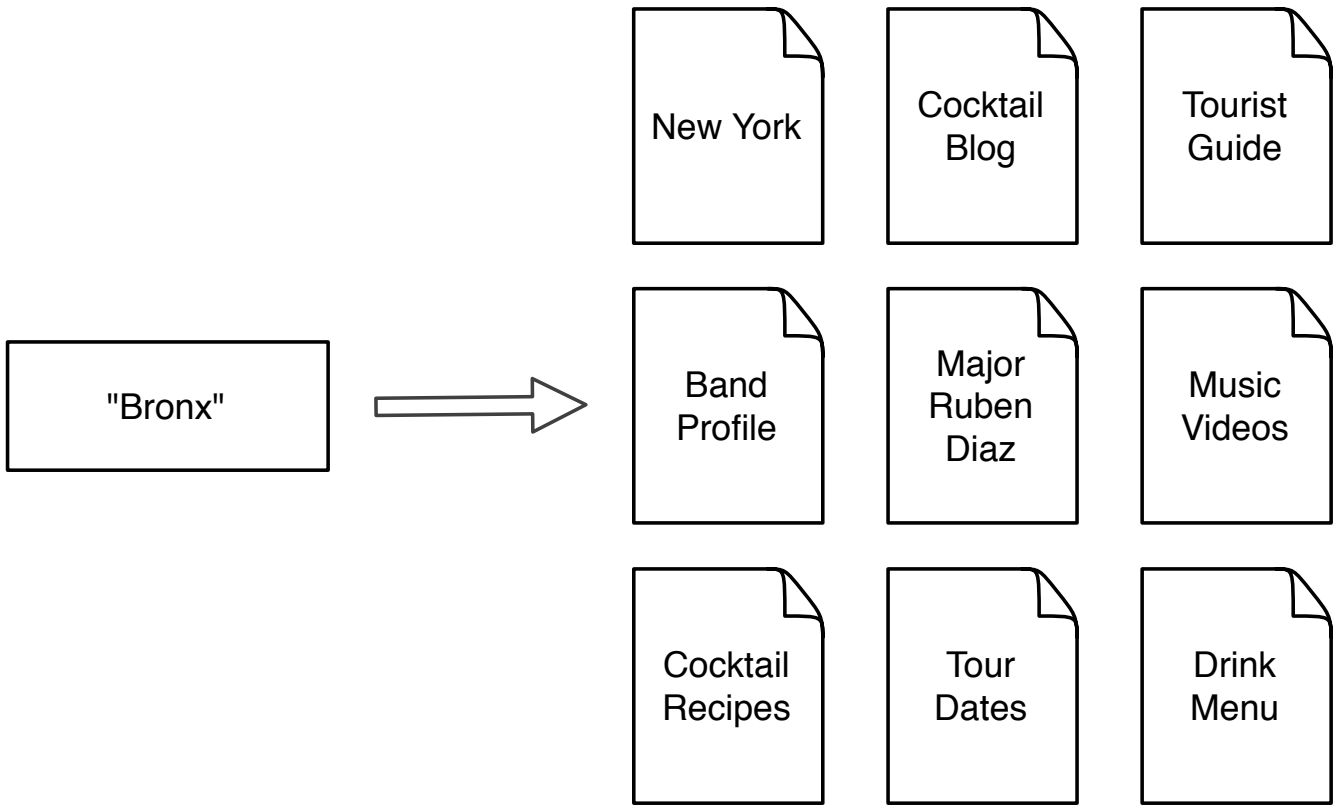
Ein neuer Ansatz für Clusterlabeling: Was war die Suchanfrage?

Maximilian Michel

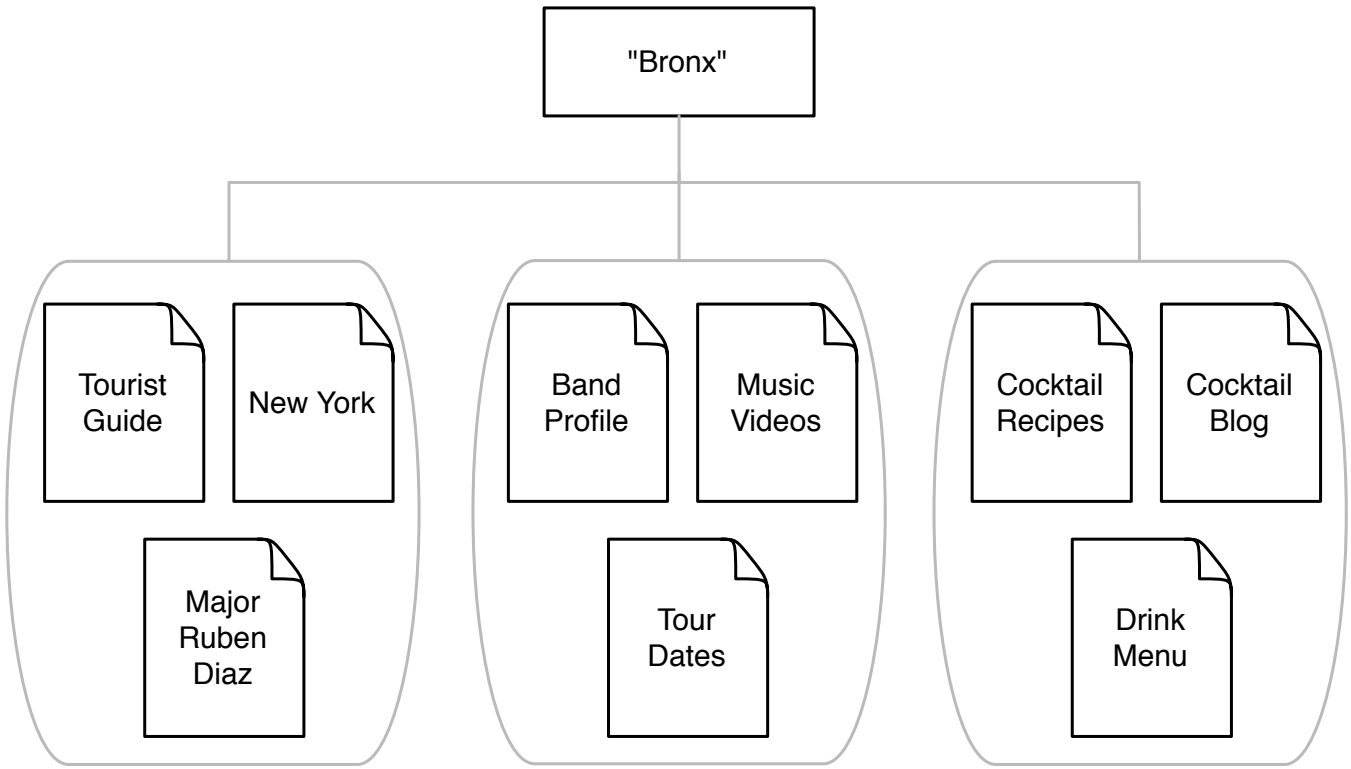
Bauhaus-Universität Weimar

23. April 2012

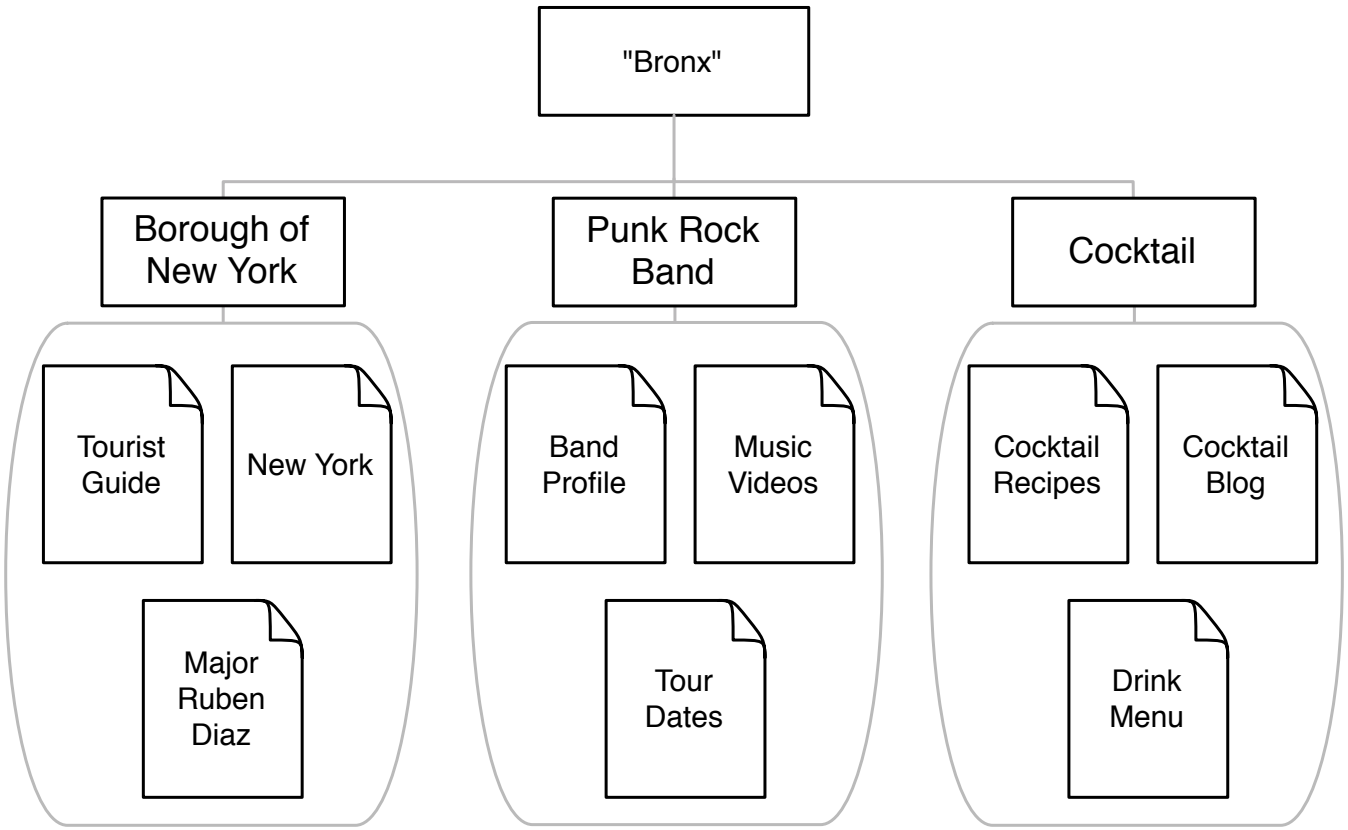
Clustering


"Bronx"
New York
Cocktail Blog
Tourist Guide
Band Profile
Major Ruben Diaz
Music Videos
Cocktail Recipes
Tour Dates
Drink Menu

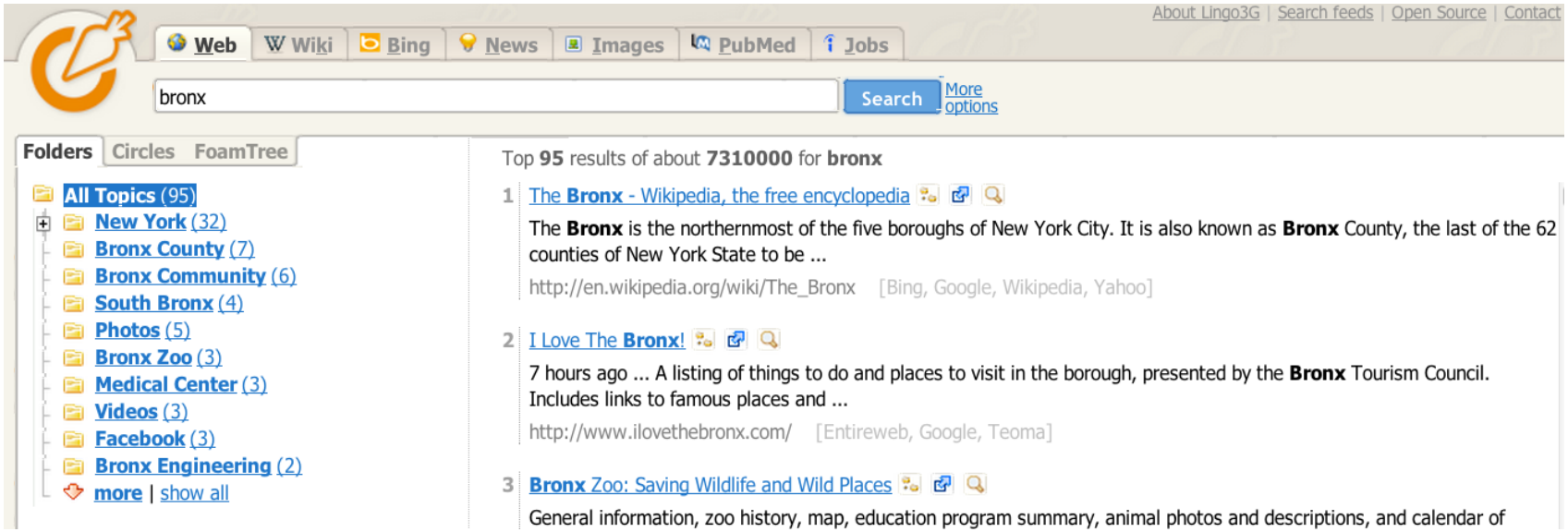
Clustering


"Bronx"
Tourist Guide
New York
Major Ruben Diaz
Band Profile
Music Videos
Tour Dates
Cocktail Recipes
Cocktail Blog
Drink Menu

Clusterlabeling



Carrot Search



Clusterlabeling Verfahren

Cluster-Internal Labeling

Differential Labeling

Clusterlabeling Verfahren

Cluster-Internal Labeling

Weighted Centroid Covering* [Stein, Meyer zu Eißen]

- Centroid-Dokument
- Top-k Terme

Differential Labeling

Clusterlabeling Verfahren

Cluster-Internal Labeling

Weighted Centroid Covering* [Stein, Meyer zu Eißen]

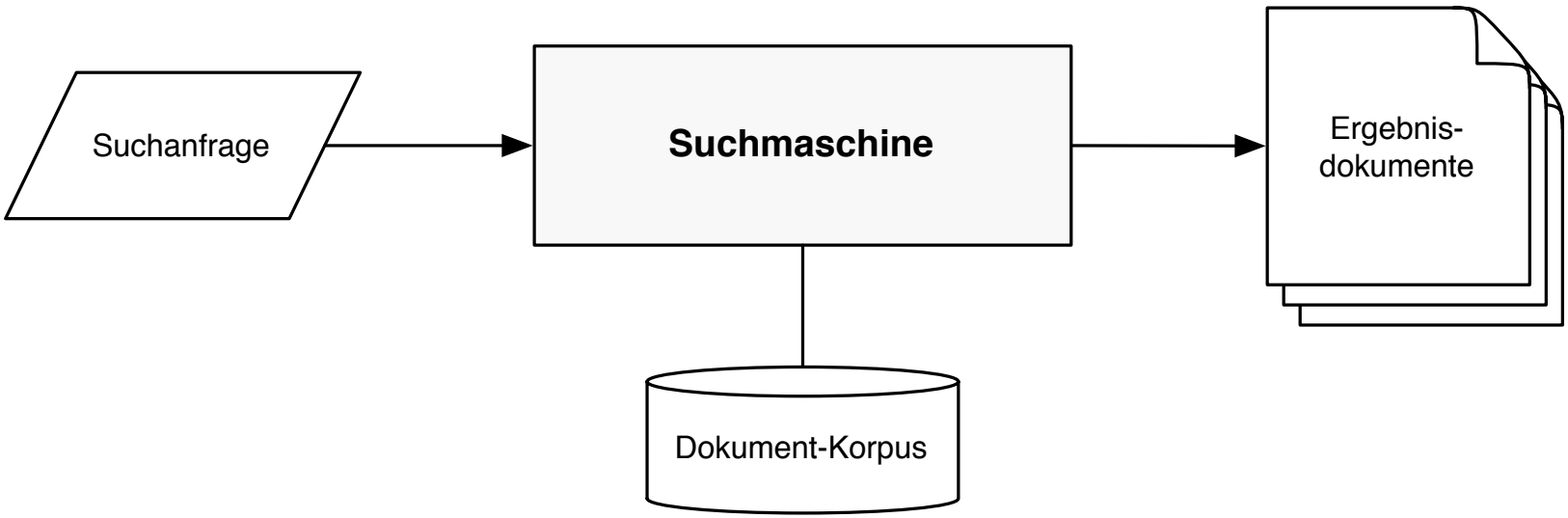
- Centroid-Dokument
- Top-k Terme

Differential Labeling

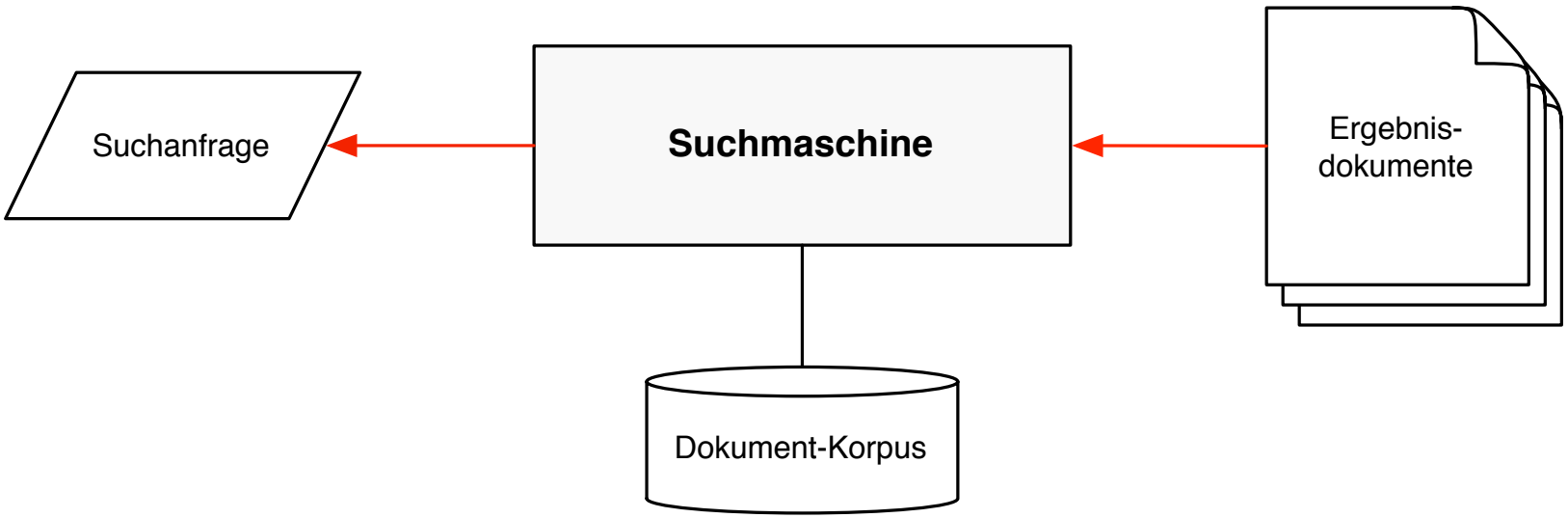
χ^2-Labeling [Introduction to IR]

- χ^2-Test
- Top-k Terme

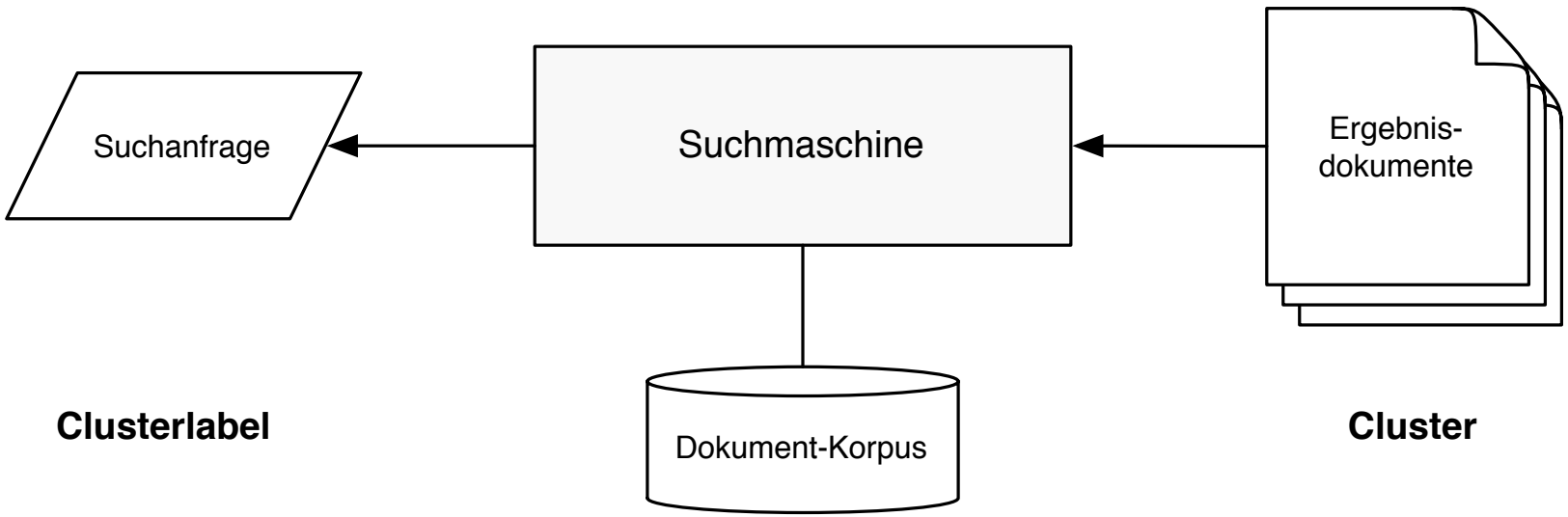
Suchmaschinensuche



Suchanfragenrekonstruktion

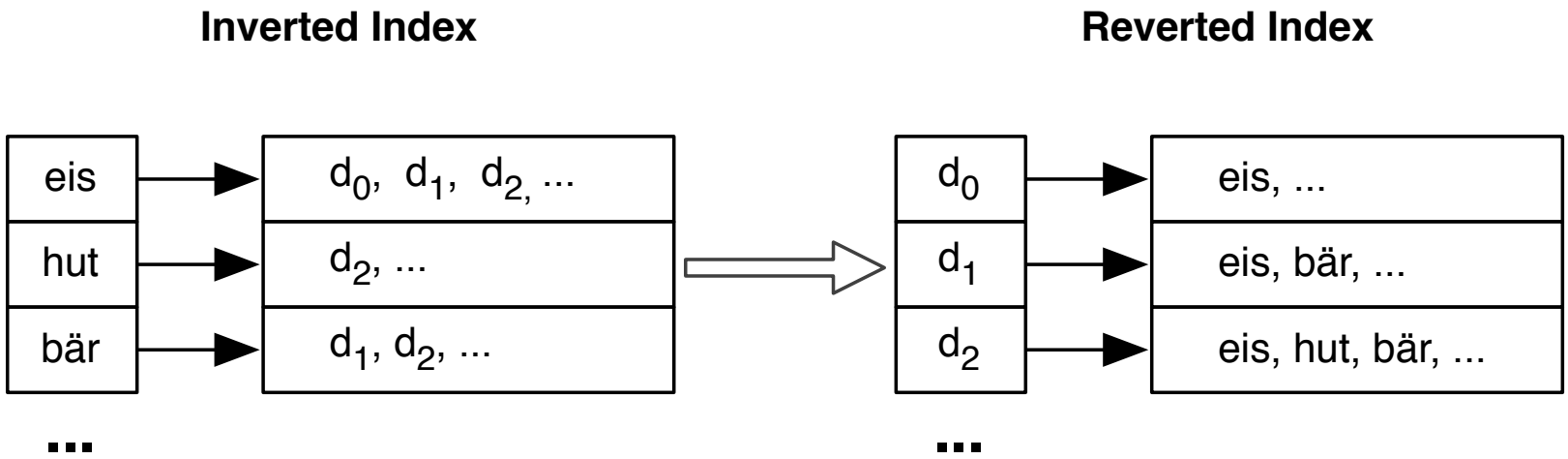


Suchanfragenrekonstruktion

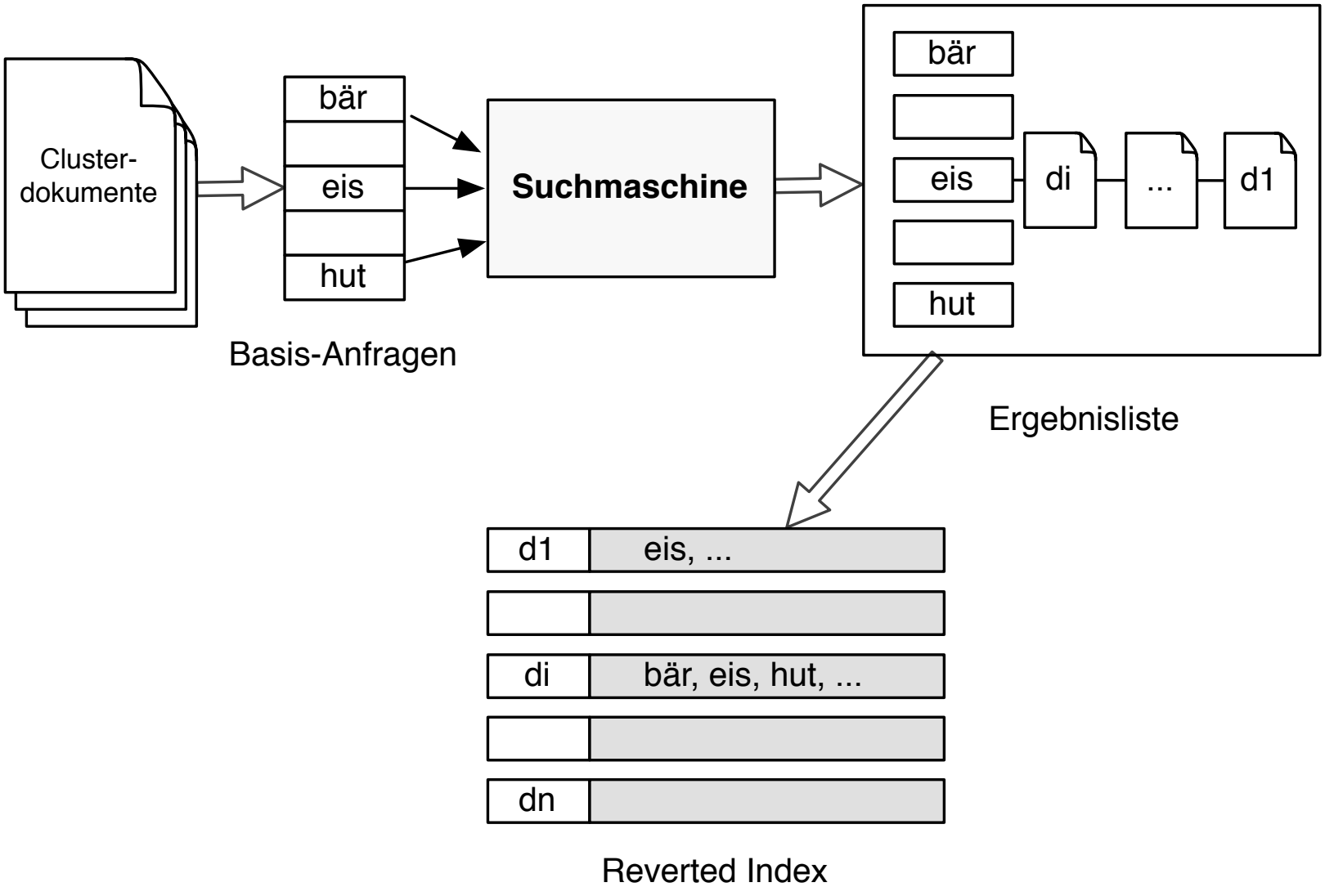


Reverted Index

[Pickens et. al]



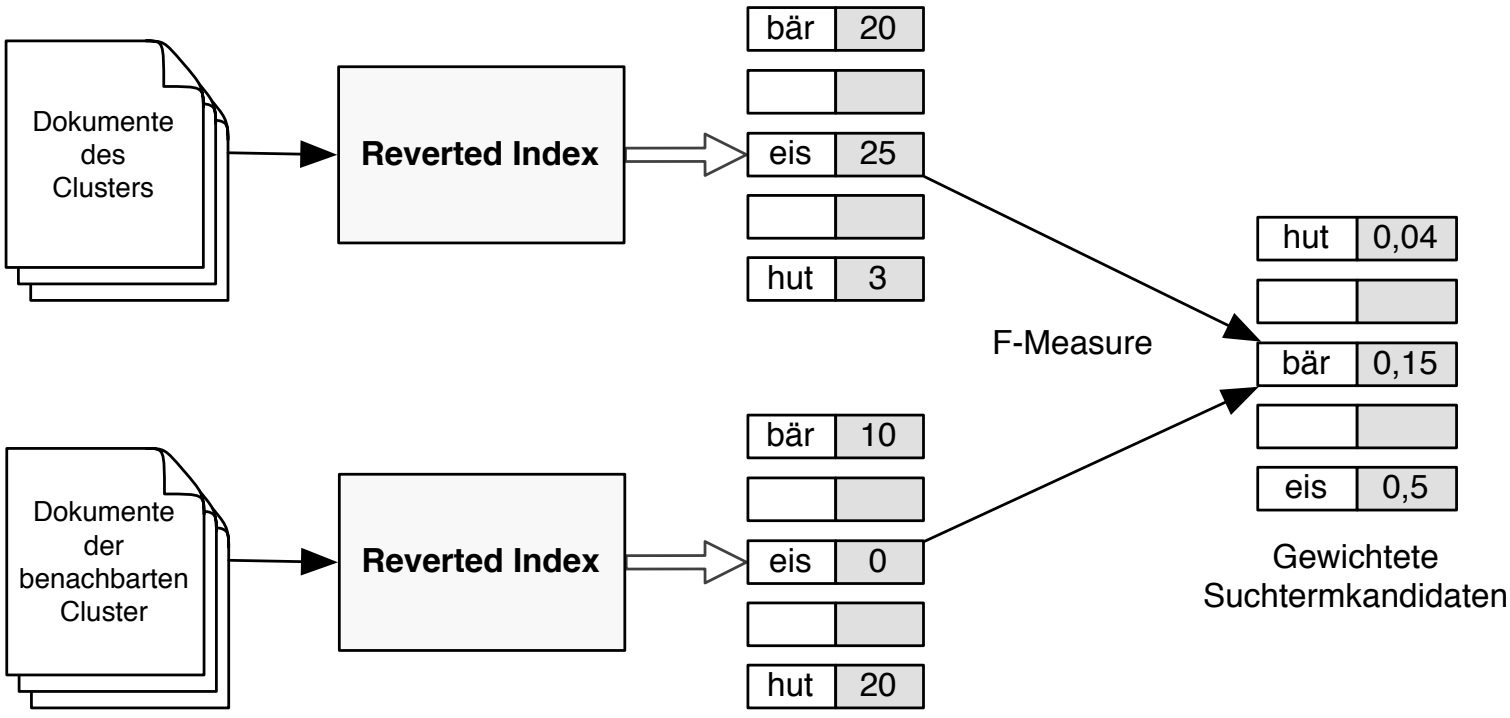
Reverted Index Konstruktion



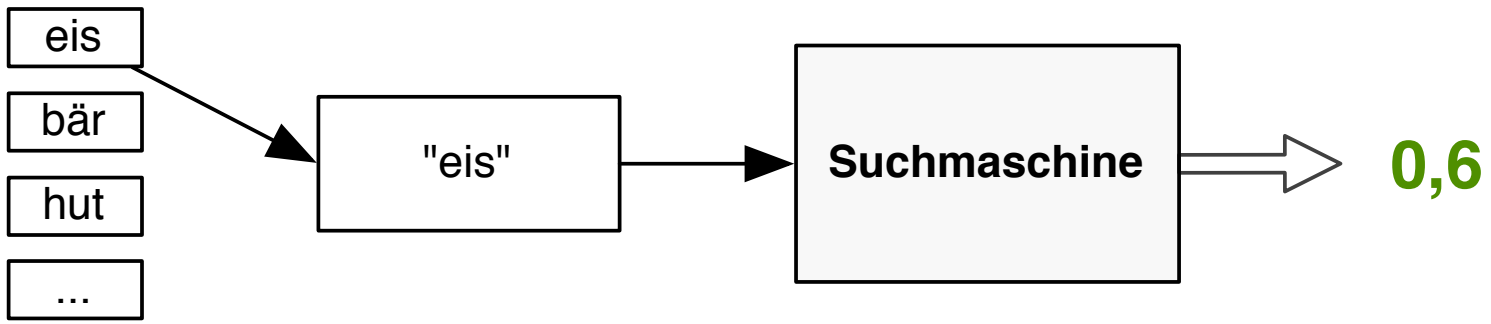
Rekonstruktion

1. Konstruktion Reverted Index
2. Finden und Gewichten der Suchtermkandidaten
3. Compositing der Suchtermkandidaten

Suchtermkandidaten

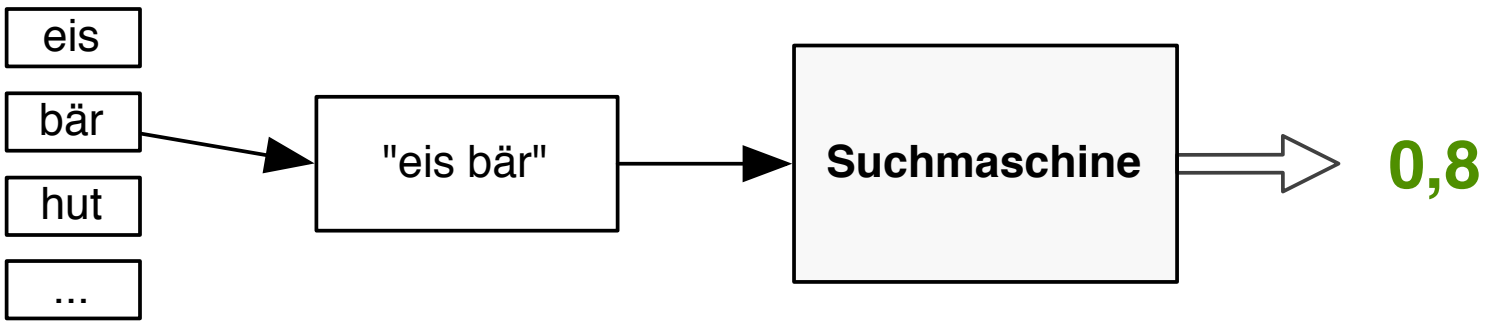


Compositing



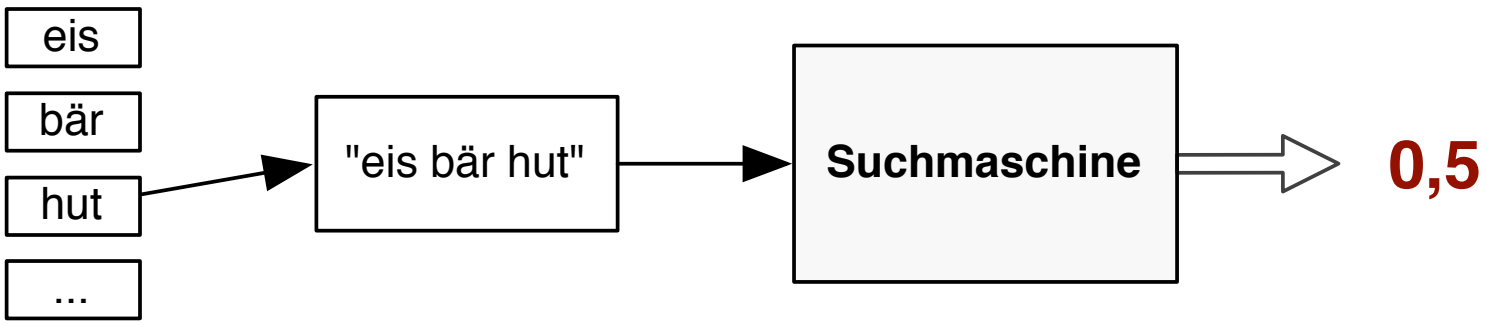
Gewichtete
Suchtermkandidaten

Compositing



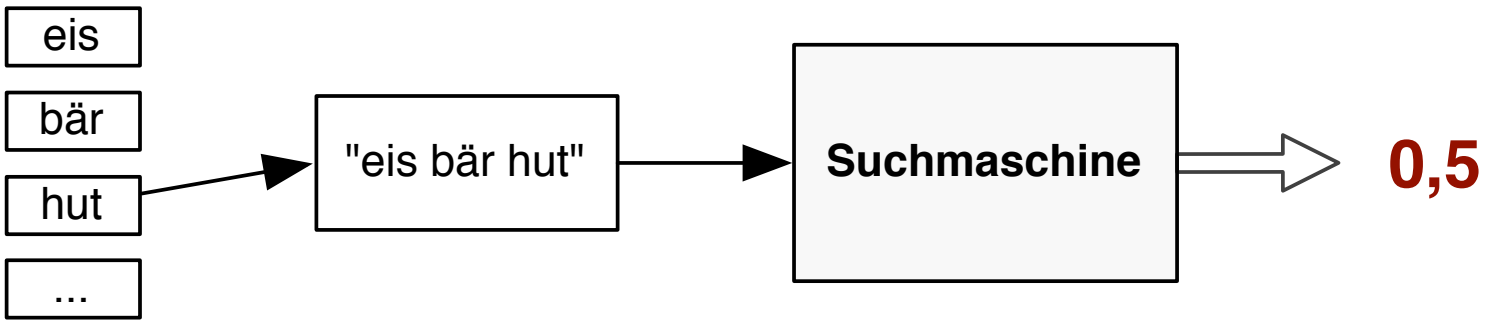
Gewichtete
Suchtermkandidaten

Compositing



Gewichtete
Suchtermkandidaten

Compositing



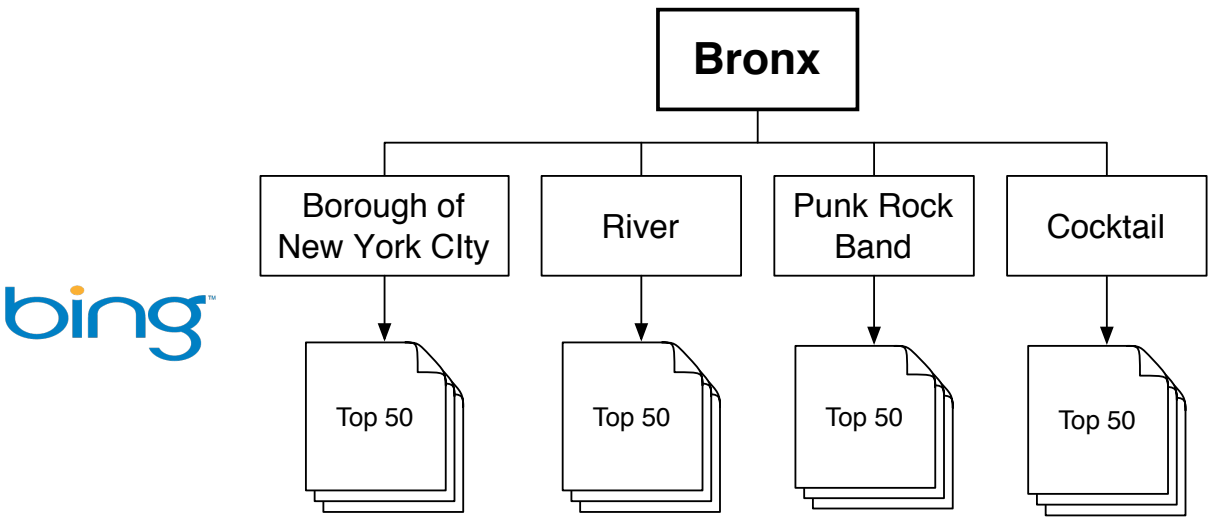
Gewichtete
Suchtermkandidaten

“eis bär”

Evaluation

Korpus

- Ambient-Dataset (AMBIgous ENTries)
- 44 Themen, je 3-37 Unterthemen $\rightarrow$ 791 Cluster
- 39 550 Dokumente



Evaluation

Referenzlabel: "**Gabriel** Fahrenheit, ein **deutscher Physiker**"

Label	F-Measure	Jaccard	Cos-Ähnl.
Gabriel Fahrenheit **deutscher** Wissenschaftler	0,66	0,5	0,66
Gabriel Fahrenheit **Physiker**	0,66	0,8	0,81
Gabriel Physiker Daniel Danzig	0,57	0,4	0,55

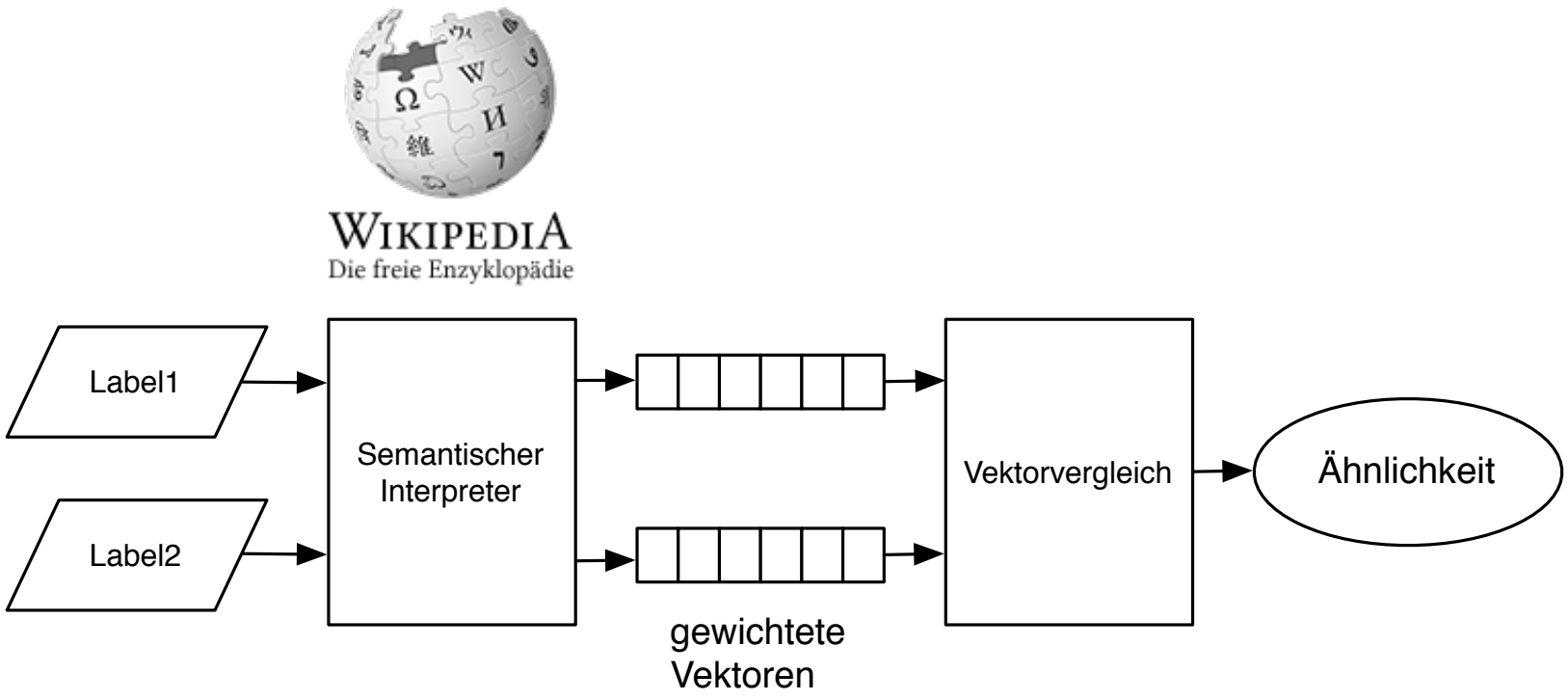
Evaluation

Referenzlabel: "**Gabriel** Fahrenheit, ein **deutscher Physiker**"

Label	F-Measure	Jaccard	Cos-Ähnl.
Gabriel Fahrenheit **deutscher** Wissenschaftler	0,66	0,5	0,66
Gabriel Fahrenheit **Physiker**	**0,66**	**0,8**	**0,81**
Gabriel Physiker Daniel Danzig	0,57	0,4	0,55

Explicit Semantic Analysis

[Gabrilovich, Markovitch]



Evaluation + ESA

Referenzlabel: "**Gabriel** Fahrenheit, ein **deutscher Physiker**";

Label	Cos-Ähnl.	ESA erweitert
Gabriel Fahrenheit **deutscher** Wissenschaftler	0,66	**0,79**
Gabriel *Fahrenheit* **Physiker**	**0,81**	0,66
Gabriel Physiker Daniel Danzig	0,55	0,42

Ergebnis

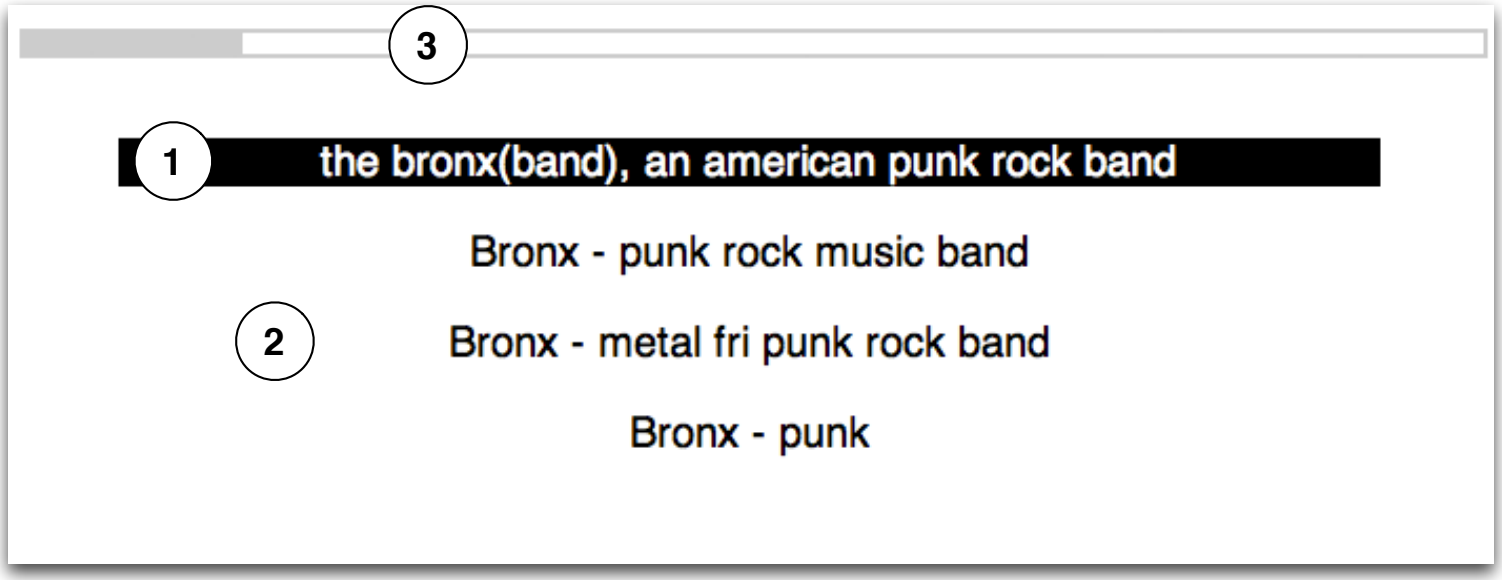
	Rekonst.	χ^2	WCC
F-Measure	0,103	**0,137**	0,056
Jaccard-Index	0,051	**0,068**	0,028
Cos-Ähnlichkeit	**0,367**	0,352	0,188
ESA erweitert	**0,443**	0,434	0,311

Nutzerevaluation

- Labels von 100 Clustern
- 23 Probanden
- 15-30 Minuten

Nutzerevaluation

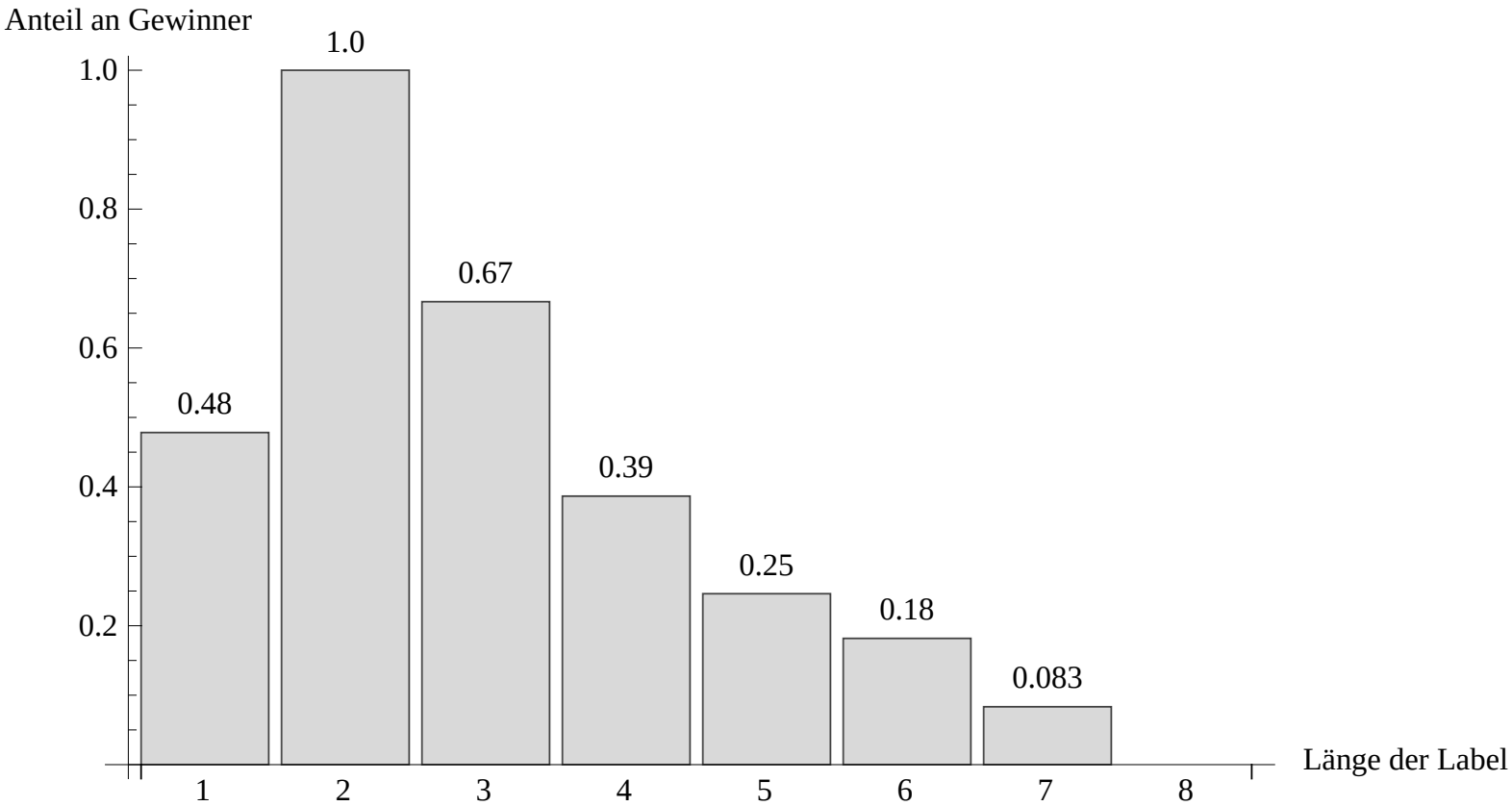
- Labels von 100 Clustern
- 23 Probanden
- 15-30 Minuten



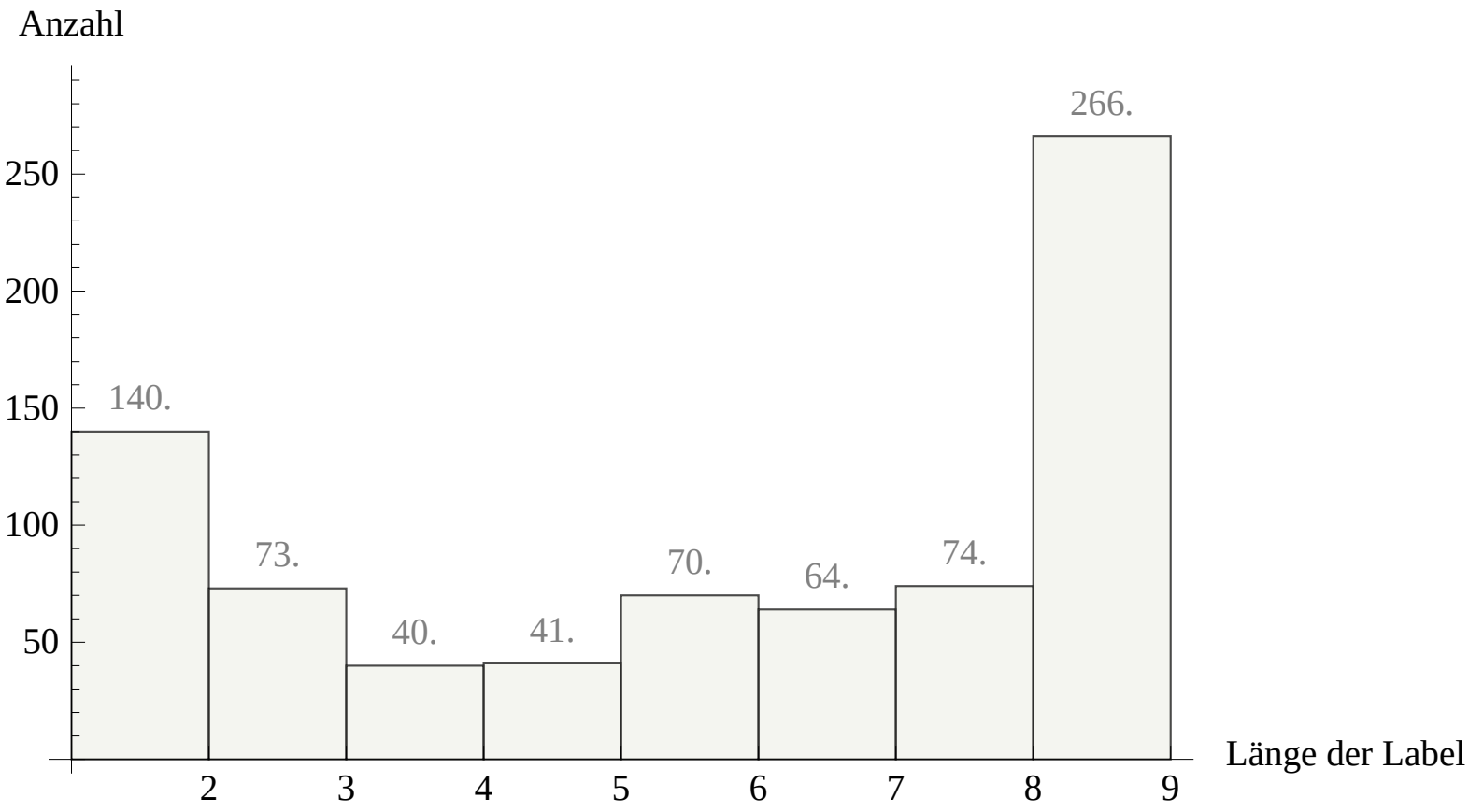
Ergebnis Nutzerevaluation

Verfahren	Klicks (relativ)	Gewinner
χ^2-Labeling	**1084 (0,45)**	**53**
Rekonst.	936 (0,39)	36
WCC	380 (0,16)	11
Gesamt	2400	100

Nutzerevaluation



Histogramm


Anzahl
250
200
150
100
50
140.
73.
40.
41.
70.
64.
74.
266.
2
3
4
5
6
7
8
9
Länge der Label

Zum Schluss

Zusammenfassung

- Suchanfragen als Clusterlabels
- Clusterkorpus aus Ambient Dataset
- Neue semantische Evaluations-Methode
- Nutzerstudie

Zum Schluss

Zusammenfassung

- Suchanfragen als Clusterlabels
- Clusterkorpus aus Ambient Dataset
- Neue semantische Evaluations-Methode
- Nutzerstudie

Ausblick

- Label-Länge
- Bessere Nutzerstudie
- Andere Suchmaschinen
- Weitere Anwendungsgebiete

Zum Schluss

Zusammenfassung

- Suchanfragen als Clusterlabels
- Clusterkorpus aus Ambient Dataset
- Neue semantische Evaluations-Methode
- Nutzerstudie

Ausblick

- Label-Länge
- Bessere Nutzerstudie
- Andere Suchmaschinen
- Weitere Anwendungsgebiete

Vielen Dank!