

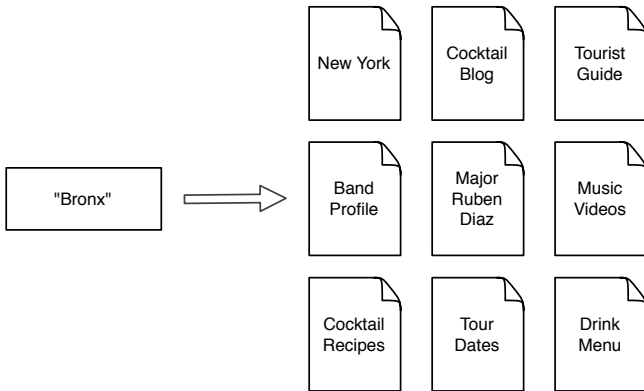
Ein neuer Ansatz für Clusterlabeling: Was war die Suchanfrage?

Maximilian Michel

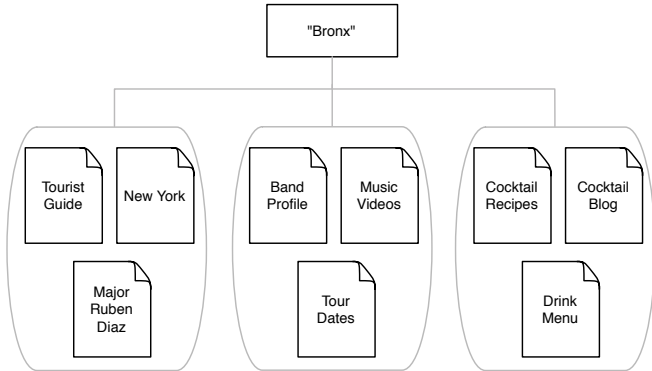
Bauhaus-Universität Weimar

23. April 2012

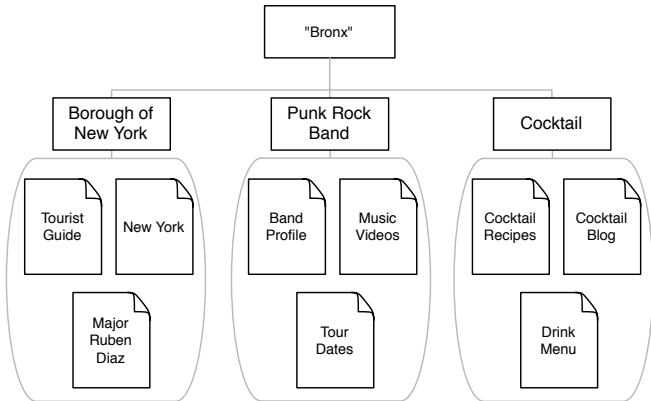
Clustering



Clustering



Clusterlabeling



Carrot Search

The screenshot shows the Carrot Search website interface. At the top left is the Carrot Search logo, a stylized orange carrot. The navigation bar includes links for Web, Wiki, Bing, News, Images, PubMed, and Jobs. A search bar contains the text "bronx" and a "Search" button with a "More options" link. Below the search bar are tabs for "Folders", "Circles", and "FoamTree". The "Folders" tab is active, showing a tree view of search results categorized by location and topic, such as "New York (32)", "Bronx County (7)", "Bronx Community (6)", "South Bronx (4)", "Photos (5)", "Bronx Zoo (3)", "Medical Center (3)", "Videos (3)", "Facebook (3)", and "Bronx Engineering (2)".

Top **95** results of about **7310000** for **bronx**

- [The Bronx - Wikipedia, the free encyclopedia](#)
The **Bronx** is the northernmost of the five boroughs of New York City. It is also known as **Bronx County**, the last of the 62 counties of New York State to be ...
http://en.wikipedia.org/wiki/The_Bronx [Bing, Google, Wikipedia, Yahoo]
- [I Love The Bronx!](#)
7 hours ago ... A listing of things to do and places to visit in the borough, presented by the **Bronx** Tourism Council. Includes links to famous places and ...
<http://www.ilovethebronx.com/> [Entireweb, Google, Teoma]
- [Bronx Zoo: Saving Wildlife and Wild Places](#)
General information, zoo history, map, education program summary, animal photos and descriptions, and calendar of

Clusterlabeling Verfahren

Cluster-Internal Labeling

Differential Labeling

Clusterlabeling Verfahren

Cluster-Internal Labeling

Weighted Centroid Covering*

[Stein, Meyer zu Eißel]

- Centroid-Dokument
- Top- k Terme

Differential Labeling

Clusterlabeling Verfahren

Cluster-Internal Labeling

Weighted Centroid Covering*

[Stein, Meyer zu Eißel]

- Centroid-Dokument
- Top- k Terme

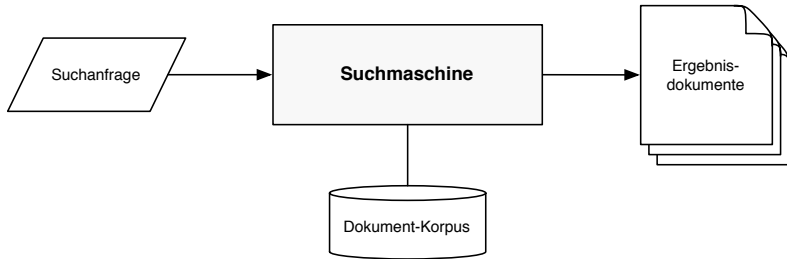
Differential Labeling

χ^2 -Labeling

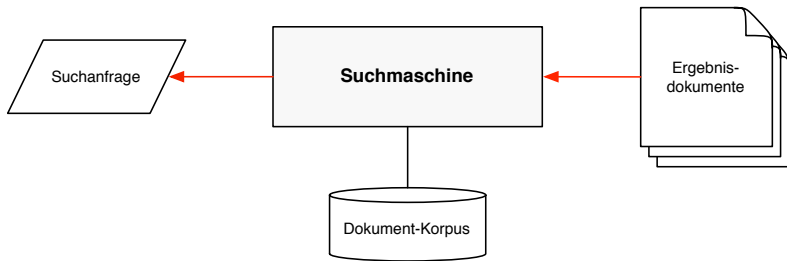
[Introduction to IR]

- χ^2 -Test
- Top- k Terme

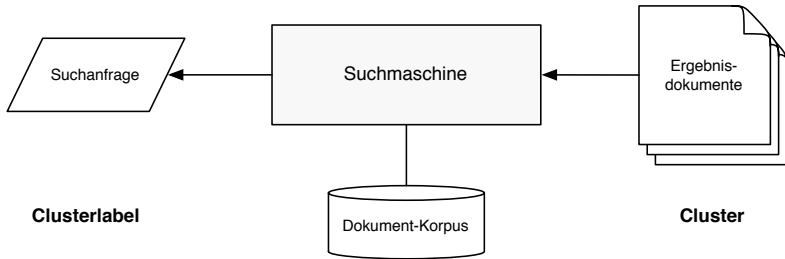
Suchmaschinenuche



Suchanfragenrekonstruktion

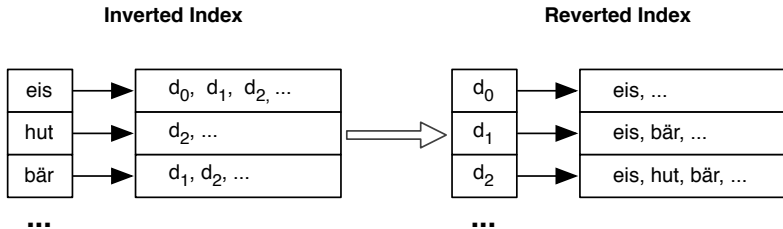


Suchanfragenrekonstruktion

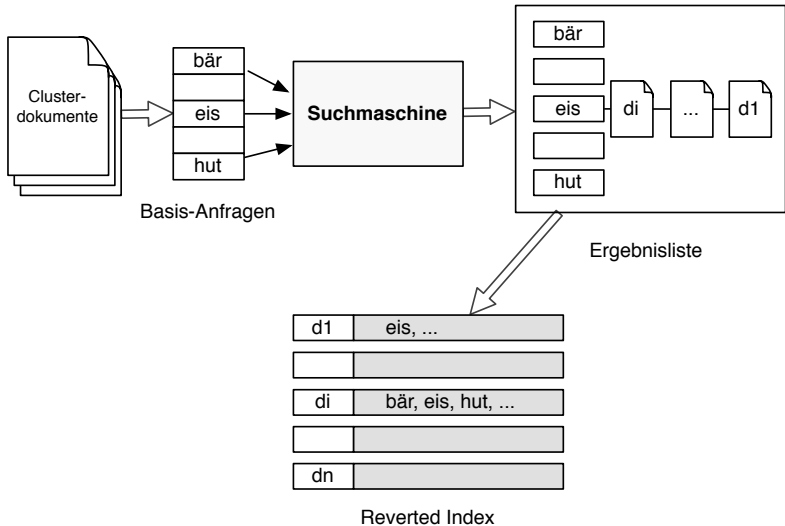


Reverted Index

[Pickens et. al]



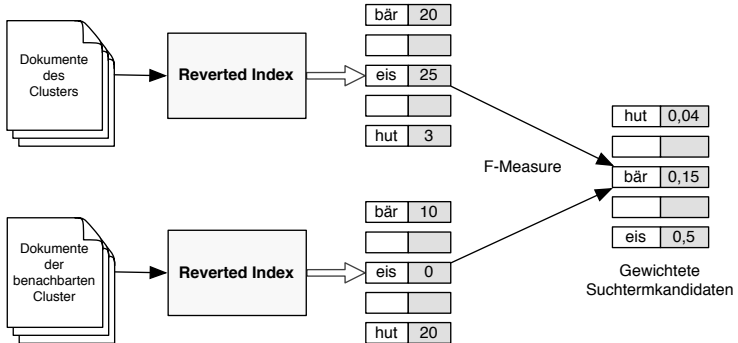
Reverted Index Konstruktion



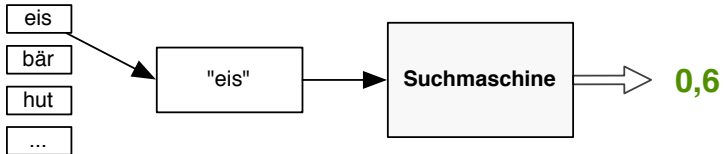
Rekonstruktion

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

Suchtermkandidaten

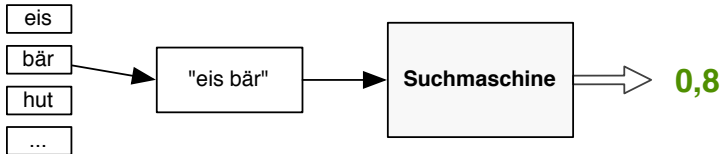


Compositing



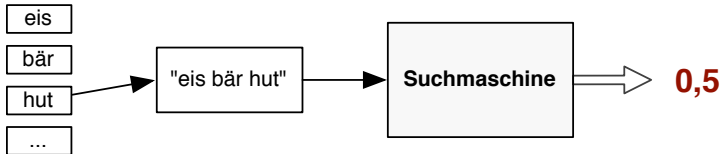
Gewichtete
Suchtermkandidaten

Compositing



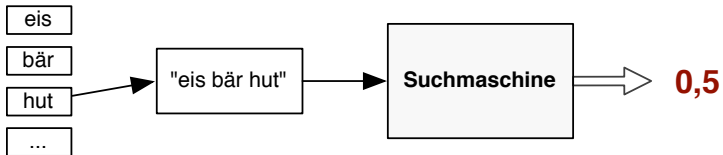
Gewichtete
Suchtermkandidaten

Compositing



Gewichtete
Suchtermkandidaten

Compositing



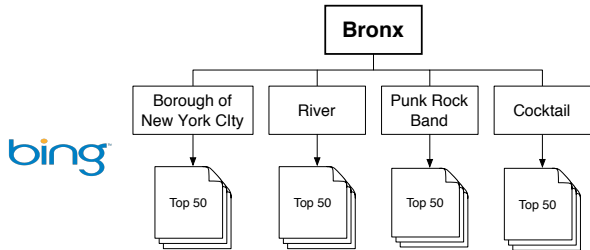
Gewichtete
Suchtermkandidaten

“eis bär”

Evaluation

Korpus

- Ambient-Dataset (AMBIGuous ENTRIES)
- 44 Themen, je 3-37 Unterthemen → 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**"

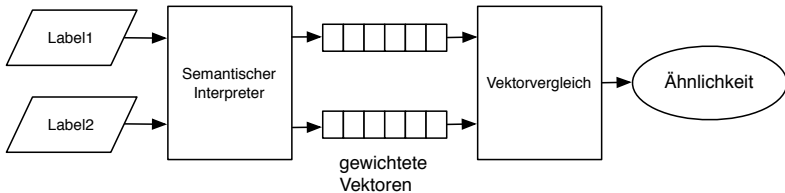
Label	F-Measure	Jaccard	Cos-Ähnl.
Gabriel Fahrenheit deutscher Wissenschaftler	0,66	0,5	0,66
Gabriel Fahrenheit Physiker	0,66	0,8	0,81
Gabriel Physiker Daniel Danzig	0,57	0,4	0,55

Explicit Semantic Analysis

[Gabilovich, Markovitch]



WIKIPEDIA
Die freie Enzyklopädie



Evaluation + ESA

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

Label	Cos-Ähnl.	ESA erweitert
Gabriel Fahrenheit deutscher Wissenschaftler	0,66	0,79
Gabriel <i>Fahrenheit</i> 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

The screenshot shows a software interface for user evaluation. At the top, there is a progress bar with a grey segment on the left and a white segment on the right. A circle containing the number '3' is positioned above the progress bar. Below the progress bar, there is a list of labels. The first label, '1 the bronx(band), an american punk rock band', is highlighted with a black background. Below it are two other labels: '2 Bronx - punk rock music band' and 'Bronx - metal fri punk rock band'. At the bottom of the list is the label 'Bronx - punk'.

3

1 **the bronx(band), an american punk rock band**

Bronx - punk rock music band

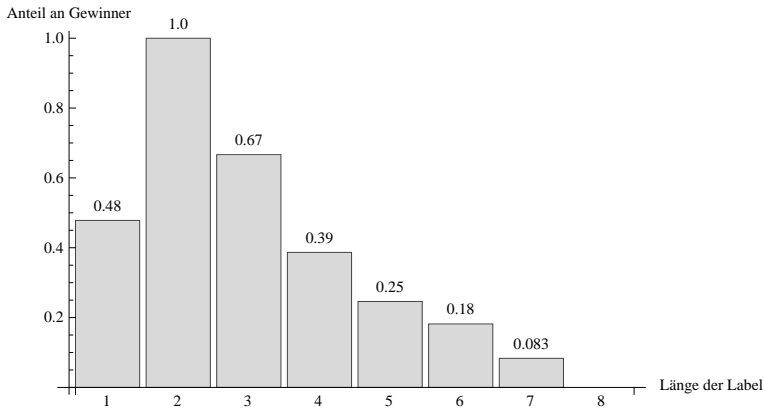
2 Bronx - metal fri punk rock band

Bronx - punk

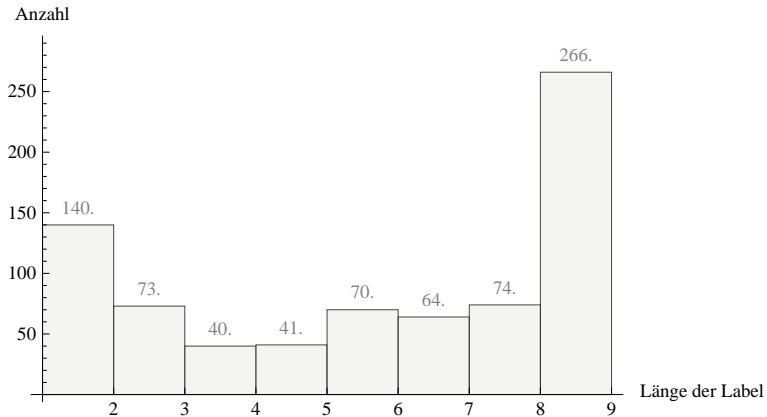
Ergebnis Nutzerevaluation

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



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!