

Cluster-Labeling

Paradigmen und Validierung

Dennis Hoppe

Bauhaus-Universität Weimar

29. Juni 2010



Motivation



Google

antibiotics Suche

Ungefähr 21.100.000 Ergebnisse (0,52 Sekunden) Erweiterte Suche

 **Alles**
 **Bilder**
 **Bücher**
 **Mehr**

Das Web
[Seiten auf Deutsch](#)
[Seiten aus Deutschland](#)

Alle
[Letzte 2 Tage](#)

Antibiotic - [Wikipedia, the free encyclopedia](#) - [[Diese Seite übersetzen](#)]
In common usage, an **antibiotic** (from the Ancient Greek: ἀντί – anti, "against", and βίος – bios, "life") is a substance or compound that kills bacteria or ...
en.wikipedia.org/wiki/Antibiotic - [Im Cache](#) - [Ähnliche](#)

Antibiotics: MedlinePlus - [[Diese Seite übersetzen](#)]
Antibiotics are powerful medicines that fight bacterial infections. Used properly, **antibiotics** can save lives. They either kill bacteria or keep them from ...
www.nlm.nih.gov/medlineplus/antibiotics.html - [Im Cache](#) - [Ähnliche](#)

Antibiotics - [[Diese Seite übersetzen](#)]
Antibiotics are among the most frequently prescribed medications in modern medicine. **Antibiotics** cure disease by killing or injuring bacteria. ...
www.emedicinehealth.com › [home](#) › [topics az list](#) - [Im Cache](#) - [Ähnliche](#)

Abbildung: Suchmaschine Google (www.google.de)

Motivation

The screenshot shows the Carrot Search engine interface. At the top, there is a navigation bar with icons for Web, MSN, Yahoo, Wiki, Images, News, Jobs, and PubMed. Below this is a search bar containing the text 'antibiotics' and a 'Search' button. To the right of the search bar is a link for 'More options'.

On the left side, there is a 'Tree' view showing a hierarchical structure of search results. The tree is expanded to show 'Infections (21)', which includes sub-topics like 'Bacterial Infections (6)', 'Treat (8)', 'Effectiveness of Many Antibiotics (3)', 'Urinary Tract (2)', 'Ear (3)', 'Infections Caused by Viruses (2)', 'Kill Bacteria (2)', 'Strep Throat (2)', 'Treatment (2)', and 'Other topics (2)'. Other main categories include 'Bacteria (14)', 'Drugs (14)', 'Antibiotics Work (8)', 'Antibiotic Resistance (9)', 'Class of Antibiotics (9)', and 'Side Effects (7)'.

On the right side, there is a list of search results. The first result is titled 'Cluster Infections with 21 documents (search for more like this)'. Below this, there are several numbered results, each with a title, a brief description, and a URL. The results are:

- 3 | [Antibiotics: MedlinePlus](#) [Icons] [Antibiotics](#) are powerful medicines that fight bacterial infections. | <http://www.nlm.nih.gov/medlineplus/antibiotics.html> [Ask, Bluewi]
- 7 | [Using Antibiotics Wisely-Topic Overview](#) [Icons] Dec 17, 2009 ... What are **antibiotics**? **Antibiotics** are medicines | <http://www.webmd.com/a-to-z-guides/using-antibiotics-wisely-topi>
- 12 | [Antibiotics: Infections: Merck Manual Home Edition](#) [Icons] Although doctors try to use **antibiotics** for specific bacterial infec | <http://www.merck.com/mmhe/sec17/ch192/ch192a.html> [Ask, E]
- 27 | [Antibiotics and pandemic flu](#) [Icons] The risk of misusing **antibiotics** is increased by the threat posed | <http://www.ecdc.europa.eu/en/eaad/antibiotics/Pages/messagesA>

Abbildung: Suchmaschine Carrot Search (www.carrotsearch.com)

Agenda

1 Validierung von Cluster-Labeling-Verfahren

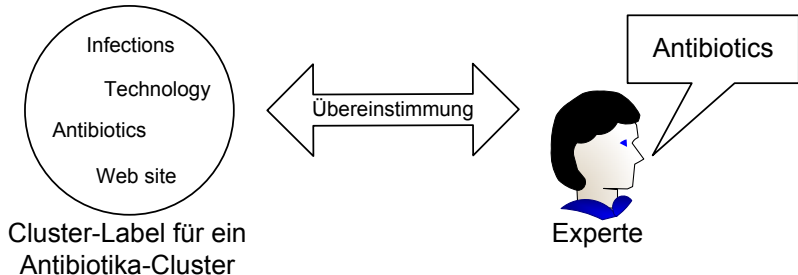
2 Paradigmen des Cluster-Labelings

Validierung von Cluster-Labeling-Verfahren

Ziel ist die Bestimmung der Güte eines Cluster-Labels

- a) Übereinstimmung mit einer externen Referenz prüfen
- b) Validierung anhand intrinsischer Qualitätsmaße
- c) Durchführung von Benutzerstudien

Externe Validierung von Cluster-Labeln



Externe Validierung von Cluster-Labeln

Externe Validierungsmaße

- Precision@N,
- Match@N und
- Mean Reciprocal Rank

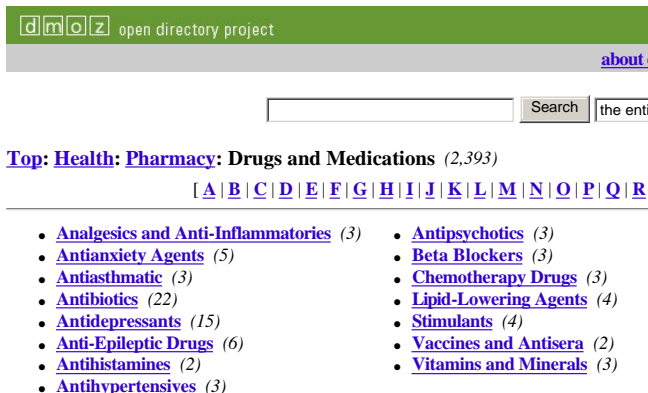


Referenz-Label: Antibiotics

Cluster-Label $\tau(c)$: Infections, Antibiotics, Technology, Web site

Maß	$N = 1$	$N = 2$	$N = 3$	$N = 4$
Precision@N	0	1/2	1/3	1/4

Externe Referenz: Open Directory Project (ODP)



The screenshot shows the Open Directory Project (ODP) website. At the top, there is a green header with the 'dmoz' logo and the text 'open directory project'. Below this is a grey navigation bar with a link to 'about'. A search bar is located in the center, with a 'Search' button and a placeholder text 'the enti'. Below the search bar, the text 'Top: [Health](#): [Pharmacy](#): [Drugs and Medications](#) (2,393)' is displayed. Underneath, there is a horizontal list of alphabetical links: [A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#). Below this list, there are two columns of category links, each followed by a number in parentheses:

- [Analgesics and Anti-Inflammatories](#) (3)
- [Antianxiety Agents](#) (5)
- [Antiasthmatic](#) (3)
- [Antibiotics](#) (22)
- [Antidepressants](#) (15)
- [Anti-Epileptic Drugs](#) (6)
- [Antihistamines](#) (2)
- [Antihypertensives](#) (3)
- [Antipsychotics](#) (3)
- [Beta Blockers](#) (3)
- [Chemotherapy Drugs](#) (3)
- [Lipid-Lowering Agents](#) (4)
- [Stimulants](#) (4)
- [Vaccines and Antisera](#) (2)
- [Vitamins and Minerals](#) (3)

Abbildung: Open Directory Project (www.dmoz.org)

Intrinsische Eigenschaften von Cluster-Labeln

Was wird von einem guten Cluster-Label erwartet?

- Verständlichkeit
- Überdeckung
- Trennschärfe
- Minimale Überlappung
- Eindeutigkeit
- Redundanzfreiheit

Intrinsische Eigenschaft: Verständlichkeit (f_1)

Informell: Ein Nutzer soll eine klare Vorstellung vom Inhalt eines Clusters bekommen.

Formal: $\forall c \in \mathcal{C} \quad \forall p \in \tau(c) : |p| > 1 \wedge p \in L(G)$

Nominalphrasen (NP)

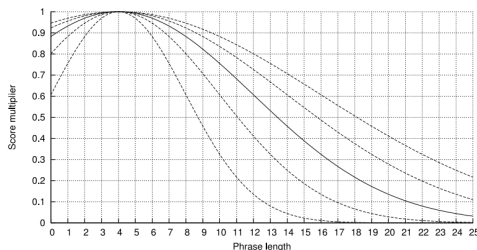
- „Antibiotika sind bakteriell wirkende Arzneistoffe.“
- „Antibiotika“ und „bakteriell wirkende Arzneistoffe“

Intrinsische Eigenschaft: Verständlichkeit (f_1)

Validierungsmaß: $f_1(p) = \text{penalty}(p) \cdot \text{NP}(p)$, mit

$$\text{penalty}(p) := \begin{cases} \exp \frac{-(|p| - |p|_{\text{opt}})^2}{2 \cdot d^2} & , \text{ wenn } |p| > 1 \\ 0,5 & , \text{ sonst} \end{cases}$$

$$\text{NP}(p) := \begin{cases} 1 & , \text{ wenn } p \in L(G) \\ 0 & , \text{ sonst} \end{cases}$$



Intrinsische Eigenschaft: Überdeckung (f_2)

Informell: Cluster-Label sollen in allen Dokumenten des Clusters vorkommen.

Formal: $\forall c \in \mathcal{C} \exists p \in \tau(c) \forall_{\substack{p' \in P_c \\ p' \notin \tau(c)}} : df_c(p') \ll df_c(p),$

mit P_c der Menge von Phrasen im Cluster c .

Validierungsmaß: $f_2(c, p) = 1 - \frac{1}{|P_c \setminus \tau(c)|} \sum_{\substack{p' \in P_c \\ p' \notin \tau(c)}} \frac{df(p')}{df(p)}$

Intrinsische Eigenschaft: Trennschärfe (f_3)

Informell: Cluster-Label sollen *nur* in Dokumenten des eigenen Clusters vorkommen.

$$\text{Formal: } \forall_{\substack{c_i, c_j \in \mathcal{C} \\ c_i \neq c_j}} \exists_{p \in \tau(c_j)} : \frac{\text{df}_{c_i}(p)}{|c_i|} \ll \frac{\text{df}_{c_j}(p)}{|c_j|}$$

$$\text{Validierungsmaß: } f_3(c_j, p) = 1 - \frac{1}{k-1} \sum_{\substack{c_i \in \mathcal{C} \\ c_i \neq c_j}} \frac{|c_j|}{|c_i|} \frac{\text{df}_{c_i}(p)}{\text{df}_{c_j}(p)}$$

Intrinsische Eigenschaft: Minimale Überlappung (f_4)

Informell: Jedes Dokument soll nur in einem Cluster vorkommen.

Formal: $\forall_{\substack{c_i, c_j \in \mathcal{C} \\ c_i \neq c_j}} \exists_{p \in \tau(c_j)} : \frac{|c_i(p) \cap c_j(p)|}{|c_i(p) \cup c_j(p)|} \ll 1$

Validierungsmaß: $f_4(c_j, p) = 1 - \frac{1}{k-1} \sum_{\substack{c_i \in \mathcal{C} \\ c_i \neq c_j}} \frac{|c_i(p) \cap c_j(p)|}{|c_i(p) \cup c_j(p)|}$

Beispiel: Trennschärfe und Minimale Überlappung

$p = \text{Antibiotics}$

$$|c_j| = 20$$

$$df_{c_j}(p) = 20$$

$$|c_i| = 10$$

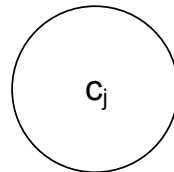
$$df_{c_i}(p) = 10$$

$$c_j \cap c_i = \emptyset$$

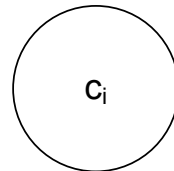
$$f_3(c_j, p) = 0 \quad (\text{Trennschärfe})$$

$$f_4(c_j, p) = 1 \quad (\text{Minimale Überlappung})$$

Antibiotics



Health



Intrinsische Eigenschaft: Eindeutigkeit (f_5)

Informell: Cluster-Label eines Clusterings sollen unterschiedlich sein.

Formal: $\forall_{\substack{c_i, c_j \in \mathcal{C} \\ c_i \neq c_j}} : \tau(c_i) \cap \tau(c_j) = \emptyset$

Validierungsmaß: $f_5(c_j, p) = 1 - \frac{1}{k-1} \sum_{\substack{c_i \in \mathcal{C} \\ c_i \neq c_j}} \frac{|p \cap \tau(c_i)|}{|p \cup \tau(c_j)|}$

Intrinsische Eigenschaft: Redundanzfreiheit (f_6)

Informell: Cluster-Label sollen keine Synonyme enthalten.

Formal: $\forall_{c \in \mathcal{C}} \forall_{\substack{p, p' \in \tau(c) \\ p \neq p'}} : p \text{ und } p' \text{ sind nicht synonym}$

Validierungsmaß: $f_6(c, p) = 1 - \frac{1}{|\tau(c)|-1} \sum_{\substack{p' \in \tau(c) \\ p' \neq p}} \text{syn}(p, p'),$

mit $\text{syn} : p \times p \mapsto \{0, 1\}$

Bewertung der Relevanz einer Phrase für ein Cluster

- Relevanz rel einer Phrase p für ein Cluster c ist definiert mit

$$rel = \sum_{i=1}^6 f_i(c, p)$$

f_1 Verständlichkeit

f_2 Überdeckung

f_3 Trennschärfe

f_4 Minimale Überlappung

f_5 Eindeutigkeit

f_6 Redundanzfreiheit

Validierung quantifizierter intrinsischer Label-Eigenschaften

Wählen Qualitätsmaße gute Phrasen aus?

ODP-Kategorie	Besten 5 Phrasen	Schlechteste 5 Phrasen
Antibiotics	used Antibiotics other Antibiotics Antibiotics Health Antibiotics Antibiotics Antibiotics Work	Technology queries project Print time
Psycho	Psycho Bates Motel Norman Marion Crane Janet Leigh shower scene Hitchcock Martin Balsam	User TOPIC mail list release

Neues internes Validierungsmaß

Normalized Discounted Cumulative Gain (NDCG) (Järvelin & Kekäläinen, 2002)

- $DCG@N = \sum_{i=1}^N (2^{rel_i} - 1) / (\log_2(1 + i))$
- Normierung anhand der idealen Relevanzliste \rightarrow NDCG

N	Phrase	Relevanz
1	Infections	4
2	Web site	1
3	Technology	0
4	Antibiotics	6
NDCG@4		0,27

N	Phrase	Relevanz
1	Antibiotics	6
2	Infections	4
3	Technology	0
4	Web site	1
NDCG@4		0,45

Korrelation von NDCG mit externem Validierungsmaß

Linearen Zusammenhang zwischen P@N & NDCG@N zeigen

- Korrelationskoeffizient (Kor) nach Pearson
- t-Test belegt statistisch signifikanten Zusammenhang beider Merkmale, wenn für die Prüfgröße (PG) gilt: $PG > t$

N	Kor(P@N,NDCG@N)	PG	$t_{(0,99;10)}$
1	0,72	3,28	3,169
2	0,90	6,70	3,169
3	0,93	8,46	3,169
4	0,96	11,71	3,169
5	0,97	13,99	3,169

Agenda

1 Validierung von Cluster-Labeling-Verfahren

2 Paradigmen des Cluster-Labelings

Paradigmen des Cluster-Labelings

- 1 Datenzentrierte Ansätze
 - Frequent Predictive Words
 - Weighted Centroid Covering
- 2 Beschreibungsbeachtende Ansätze
 - Suffixbaum-Clustering
- 3 Beschreibungszentrierte Ansätze
 - Topical k -Means
 - Descriptive k -Means
 - Lingo

Beispiel: Frequent Predictive Words

Um welche Kategorien handelt es sich?

ODP-Kategorie

Cluster-Label

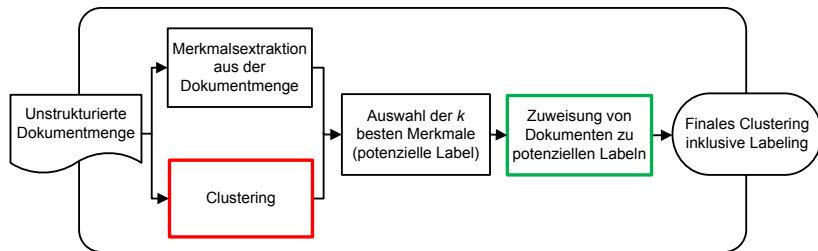
- | | |
|---|---|
| ? | spice, slicer, told, fred, baker |
| ? | excel, jeremy, demo, authentic, forum |
| ? | hat, document, project, string, release |
| ? | data, address, match, bow, custom |
| ? | antibiotics, disease, infection, bacteria, drug |

Beispiel: Frequent Predictive Words

Um welche Kategorien handelt es sich?

ODP-Kategorie	Cluster-Label
IBM DB2	spice, slicer, told, fred, baker
MySQL	excel, jeremy, demo, authentic, forum
PostgreSQL	hat, document, project, string, release
Data Warehousing	data, address, match, bow, custom
Antibiotics	antibiotics, disease, infection, bacteria, drug

Beschreibungszentrierte Ansätze



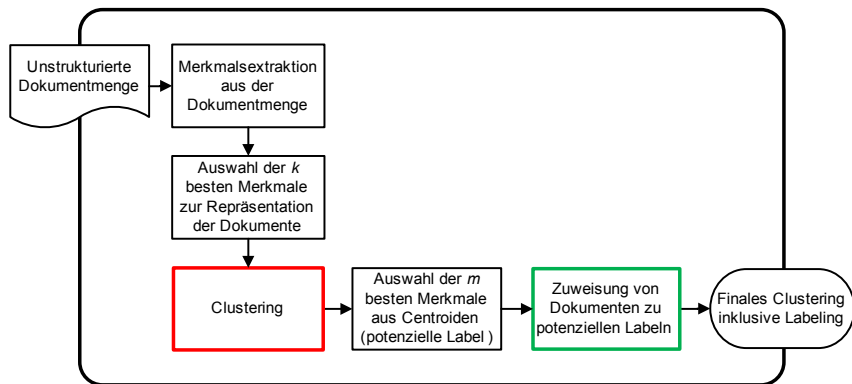
Untersuchte Verfahren

- Descriptive k -Means (Weiss, 2006)
- Lingo (Osinski u. a., 2004)

Beispiel: Descriptive k -Means

ODP-Kategorie	Cluster-Label
IBM DB2	IBM Data Management, IBM SQL Partners
MySQL	SQL Server, MySQL database server
PostgreSQL	PostgreSQL database system
Data Warehousing	data quality management solutions
Antibiotics	Antibiotic Resistant Bacteria

Neues Verfahren: Topical k -Means



Topical k -Means

Informativeness einer Phrase (Tomokiyo & Hurst, 2003)

- Welche Phrase sagt am meisten über ein Cluster aus?
- Abgrenzung einer Phrase von den restlichen Clustern

Sinnvolle Phrase für Cluster c_j ?

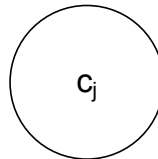
$p_1 = \text{MySQL}$

$p_2 = \text{SQL}$

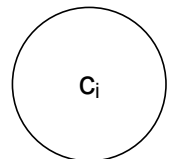
$p_3 = \text{Database}$

$p_4 = \text{PostgreSQL}$

MySQL



PostgreSQL



Auswertung

Verfahren	Precision@1	NDCG@1
Nominalphrasen	0,40	0,74
Frequent and Predictive Words	0,20	0,64
Descriptive k -Means	0,61	0,82
Topical k -Means	0,55	0,84

Qualität von Cluster-Labeln steigt bei Verwendung von

- Nominalphrasen
- einer Referenzkategorisierung

Auswertung

Verfahren	f_1	f_2	f_3	f_4	f_5	f_6
Schlüsselwortverfahren	0.79	0.66	0.37	1.00	0.94	0.99
Datenzentrierte Ansätze	0.39	0.59	0.63	1.00	0.97	1.00
Beschreibungsbeachtende Ansätze	0.73	0.70	0.89	0.68	1.00	0.99
Beschreibungszentrierte Ansätze	0.91	0.64	0.91	0.44	1.00	1.00

f_1 Verständlichkeit

f_2 Überdeckung

f_3 Trennschärfe

f_4 Minimale Überlappung

f_5 Eindeutigkeit

f_6 Redundanzfreiheit

Zusammenfassung

- 1 Validierung von Cluster-Labeling-Verfahren
- 2 Paradigmen des Cluster-Labelings

Ausblick

- Hierarchisches Cluster-Labeling
- Gewichtung intrinsischer Validierungsmaße
- Einsatz neuer Schlüsselwortverfahren
- Verwendung von externem Wissen

Erkennung von Themen in Dokumentmengen

automatic-taxonomy-generation
text-summarization
automatic-
cluster-
labeling faceted-search **keyword-**
extraction labeling-of-topic-models
search-result-clustering self-organizing-maps **tag-clouds** text-
classification **topic-detection-and-**
tracking

Referenzen

- [Järvelin & Kekäläinen 2002] Järvelin, Kalervo ; Kekäläinen, Jaana: Cumulated gain-based evaluation of IR techniques. In: *ACM Transactions on Information Systems* 20 (2002), Oktober, Nr. 4, 446.
<http://dx.doi.org/10.1145/582415.582418>. – DOI 10.1145/582415.582418. – ISSN 10468188
- [Osinski u. a. 2004] Osinski, S. ; Stefanowski, J. ; Weiss, D.: Lingo: Search results clustering algorithm based on singular value decomposition. In: *Intelligent information processing and web mining: proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, Mai 17-20, 2004*, Springer Verlag, 2004, 359
- [Tomokiyo & Hurst 2003] Tomokiyo, T. ; Hurst, M.: A language model approach to keyphrase extraction. In: *Proceedings of the ACL Workshop on Multiword Expressions*, 2003, 3440
- [Weiss 2006] Weiss, D.: *Descriptive clustering as a method for exploring text collections*. 2006