# Insights into Cluster Labeling

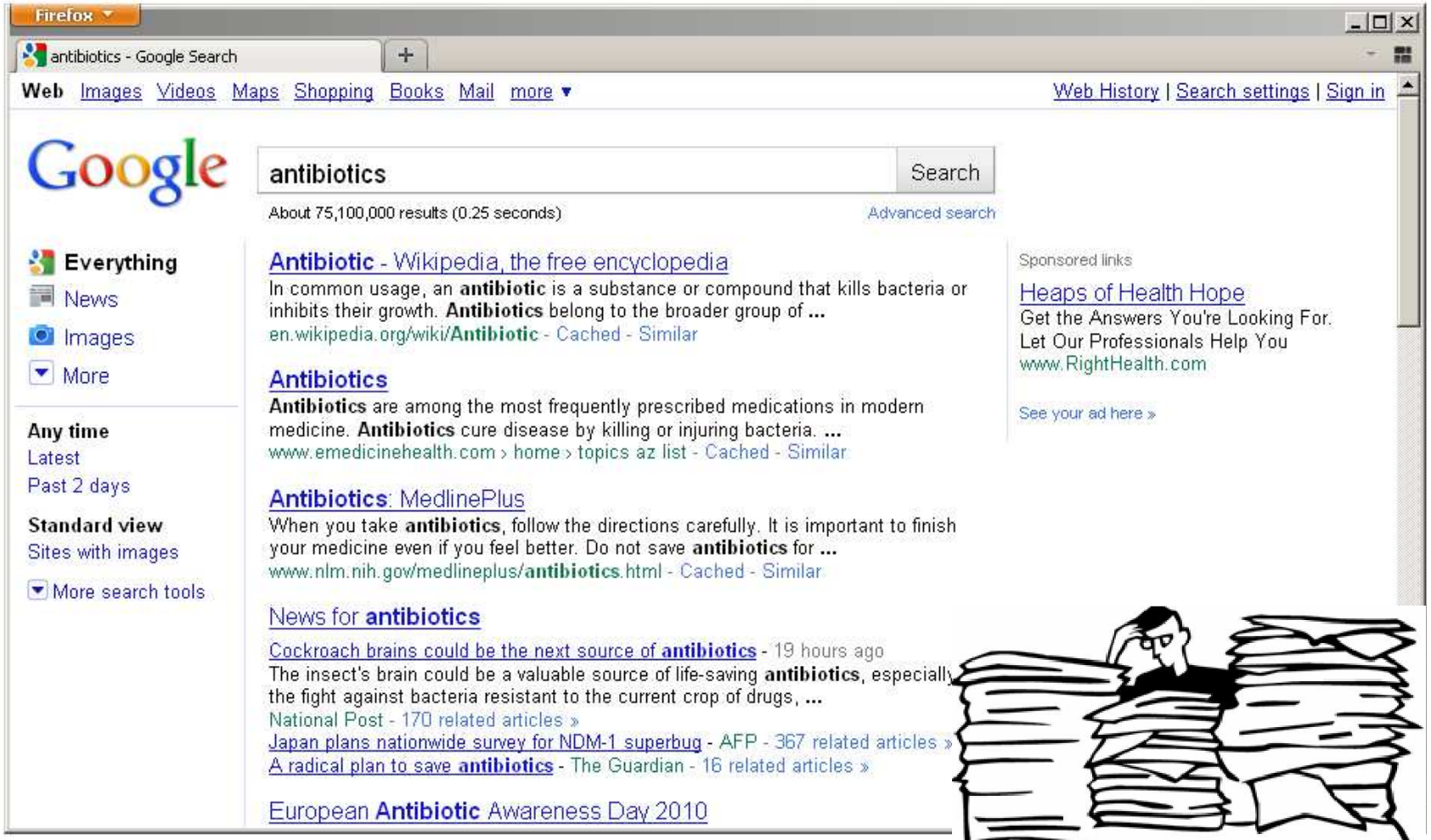
Dennis Hoppe

Web Technology and Information Systems

Bauhaus-Universität Weimar

# Cluster Labeling

## Application: Web Search Result Clustering






www.google.com

# Cluster Labeling

## Application: Web Search Result Clustering

Firefox

antibiotics - Lingo3G Document Clustering E...

About Lingo3G | Search feeds | Services | Open Source | Contact

Web | Bing | Yahoo | Wiki | Images | News | Jobs | PubMed

antibiotics — Search — More options

Tree | Visualization

- All Topics (100)
  - Infections (25)
    - Bacteria (6)
    - Antibiotics for Ear Infections (5)
    - Drugs (5)
    - Bacterial Infections (4)
    - Tract Infections (3)
    - Medical News and Health (2)
    - Infections Caused by Viruses (2)
    - Antibiotics are a Class (3)
    - Treatment (4)
    - Effects (5)
    - more | show all
  - Antibiotic Resistance (8)
  - Class of Antibiotics (9)
  - Antibiotics Working (8)
  - Buy Antibiotics (7)
  - Medical (9)
  - Penicillin (8)
  - Side Effects (5)
  - Children (5)
  - Viruses (5)
  - more | show all

Cluster **Tract Infections** with **3** documents

36 Appropriate Prescribing of Oral Beta-Lactam **Antibiotics** - August 1, 2000 - American Family Physician

Less Clear Indications The benefits of beta-lactam **antibiotics** in the treatment of bronchitis, skin and soft tissue infections, and urinary tract infections are less clear in the evidence-based literature. The marginal benefit of **antibiotics** ...

http://www.aafp.org/afp/20000801/611.html [Cull]

69 What **Antibiotics** Cure Pyelonephritis - HealthCentral

Everything you need to know about what **antibiotics** cure pyelonephritis. ... **Antibiotics**: Not Always the Answer for Upper Respiratory Tract Infections ...

http://www.healthcentral.com/incontinence/h/what-antibiotics-cure-pyelonephritis.html [Yahoo]

77 Urinary tract infection: antibiotic therapy recommendations

Urinary tract infection: antibiotic therapy recommendations The authors make no claims of the accuracy of the information contained herein; and these suggested doses are not a substitute for clinical judgement. Neither GlobalRPh Inc. BactrimDS ...

http://www.globalrph.com/antibiotic/uti.htm [Cull]

Query: **antibiotics** -- Source: **Web** (100 results, 22 ms) -- Clusterer: **Lingo3G** (48 ms)

search.carrotsearch.com

# Cluster Labeling

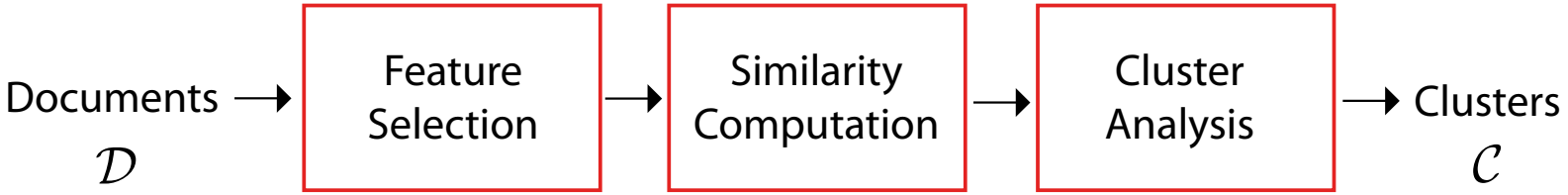
## Outline

- Formalization of Cluster Labels
- Evaluation of Cluster Labels
- Paradigms of Cluster Labeling

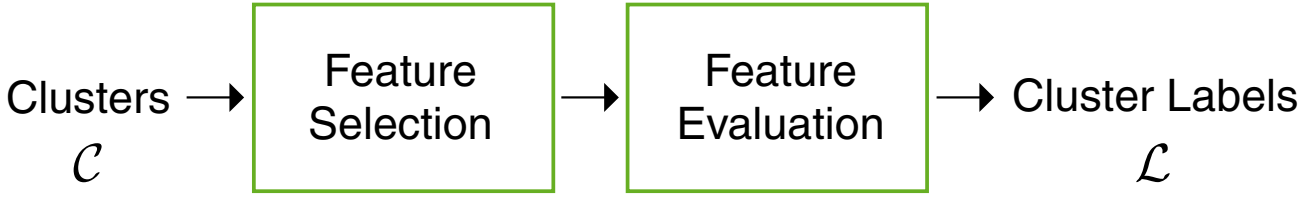
# Cluster Labeling

## Problem Statement

### Clustering $\mathcal{C}$




### Labling $\mathcal{L}$




Cluster Label $l \in \mathcal{L}$ : antibiotics, disease, infection, bacteria, drug

# Cluster Labeling

## Formalization of Cluster Labels

What accounts for “good” cluster labels?

- Comprehensibility
- Descriptiveness
- Discriminative power
- Uniqueness
- Non-redundancy
- Minimal Overlap
- Hierarchically consistency

The formalization is based on previous work done in [8].

# Cluster Labeling

## (a) Formalization of Cluster Labels: Comprehensibility ($f_1$)

Informal:
A reader should have a *clear imagination* of the contents of a cluster.

Formal:

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

where $l_c$ is the cluster label of cluster $c$, $p$ a phrase of $l_c$, and $L(G)$ determines a formal language identifying noun phrases.

# Cluster Labeling

## (a) Formalization of Cluster Labels: Comprehensibility ($f_1$)

Informal:
A reader should have a *clear imagination* of the contents of a cluster.

Formal:

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

where $l_c$ is the cluster label of cluster $c$, $p$ a phrase of $l_c$, and $L(G)$ determines a formal language identifying noun phrases.

Why select noun phrases as comprehensible cluster labels?

- Single terms [8] suffer from a loss of information.
- Named Entities [2, 9, 3] are too strict.
- Titles of web pages [5] are not always available.
- Frequent phrases [11] are often grammatically incorrect or meaningless.

# Cluster Labeling

## (a) Formalization of Cluster Labels: Comprehensibility ($f_1$)

Criterion:

$$f_1(p) = \text{NP}(p) \cdot \text{penalty}(p)$$

where

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

$$\text{penalty}(p) = \begin{cases} \exp \frac{-(|p| - |p|_{\text{opt}})^2}{2 \cdot d^2} & \text{, if } |p| > 1 \\ 0.5 & \text{, otherwise} \end{cases}$$

Note that the exponential expression was earlier used in [10] to penalize too short or too long phrases. [10] set $|p|_{opt} = 4$ and $d = 8$.

# Cluster Labeling

## (b) Formalization of Cluster Labels: Descriptiveness ($f_2$)

Informal:
*Every* document of a cluster should contain the associated cluster label.

Formal:

$$\forall c \in \mathcal{C}\ \exists p \in l_c\ \forall \underset{p' \notin l_c}{p' \in P_c} : \mathit{df}_c(p') \ll \mathit{df}_c(p)$$

where $P_c$ is the set of phrases in the cluster $c$.

Criterion:

$$f_2(c,p) = 1 - \frac{1}{|P_c \setminus l_c|} \sum_{\substack{p' \in P_c \\ p' \notin l_c}} \frac{\mathit{df}_c(p')}{\mathit{df}_c(p)}$$

# Cluster Labeling

## (c) Formalization of Cluster Labels: Discriminative Power ($f_3$)

Informal:
A cluster label should *only* be present in documents of its own cluster.

Formal:

$$\forall c_i, c_j \in \mathcal{C} \; \exists p \in l_c \; : \; \frac{df_{c_i}(p)}{|c_i|} \ll \frac{df_{c_j}(p)}{|c_j|} \qquad c_i \neq c_j$$

Criterion:

$$f_3(c_j, p) = 1 - \frac{1}{k-1} \sum_{\substack{c_i \in \mathcal{C} \\ c_i \neq c_j}} \frac{|c_j| \cdot df_{c_i}(p)}{|c_i| \cdot df_{c_j}(p)}$$

# Cluster Labeling

## (e) Formalization of Cluster Labels: Uniqueness ($f_4$)

Informal:
Cluster labels should be unique.

Formal:

$$\forall \underset{c_i \neq c_j}{c_i, c_j \in \mathcal{C}} \; : \; l_{c_i} \cap l_{c_j} = \emptyset$$

Criterion:

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

# Cluster Labeling

## (f) Formalization of Cluster Labels: Non-redundancy ($f_5$)

Informal:
Cluster labels should not be synonymous.

Formal:

$$\forall c \in \mathcal{C}\ \forall \underset{p \neq p'}{p, p' \in l_c} : p \text{ and } p' \text{ are not synonymous}$$

Criterion:

$$f_5(c, p) = 1 - \frac{1}{|l_c| - 1} \sum_{\substack{p' \in l_c \\ p' \neq p}} \text{syn}(p, p')$$

where syn $: p \times p \mapsto \{0, 1\}$.

# Cluster Labeling

## Relevance of a phrase with respect to a cluster

All constraints can be combined into a single criterion:

$$\mathit{rel}(c,p) = \sum_{i=1}^{|\mathcal{F}|} \omega_i \cdot f_i(c,p)$$

where $\omega_i$ is a weighting factor and $\mathcal{F} = \{f|1 \ldots 5\}$, namely,

- $f_1$ Comprehensibility
- $f_2$ Descriptiveness
- $f_3$ Discriminative Power
- $f_4$ Uniqueness
- $f_5$ Non-redundancy

Note, that the effect of every constraint on the quality of a phrase is so far unevaluated.

# Cluster Labeling [^]

Do these constraints really select good phrases as cluster labels?

| Category | Top 5 Phrases | Worst 5 Phrases |
|---|---|---|
| Antibiotics | used Antibiotics | Technology |
| | other Antibiotics | queries |
| | Antibiotics Health | project |
| | Antibiotics Antibiotics | Print |
| | Antibiotics Work | time |
| Psycho (Movie) | Psycho | User |
| | Bates Motel Norman | TOPIC |
| | Marion Crane Janet Leigh | mail |
| | shower scene Hitchcock | list |
| | Martin Balsam | release |

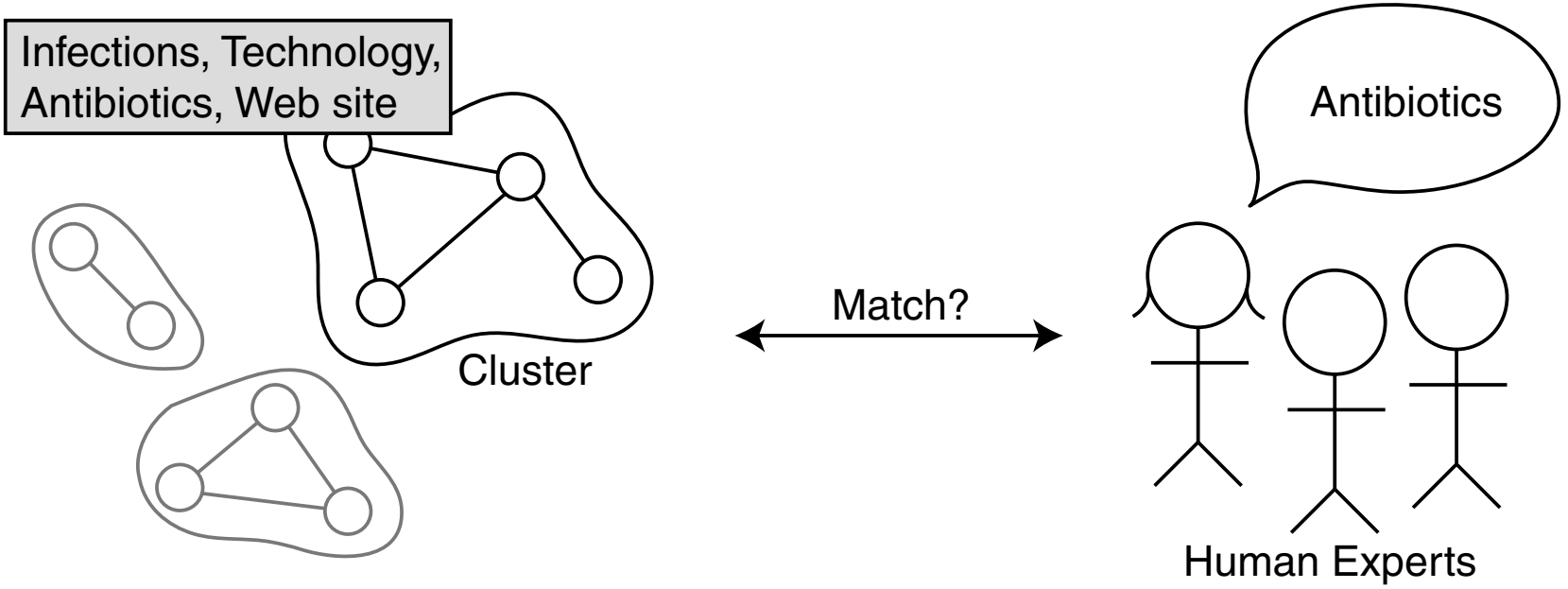
# Cluster Labeling

## Evaluation of Cluster Labels

- External Evaluation
- Internal Evaluation
- User Studies

# Cluster Labeling

## External Evaluation

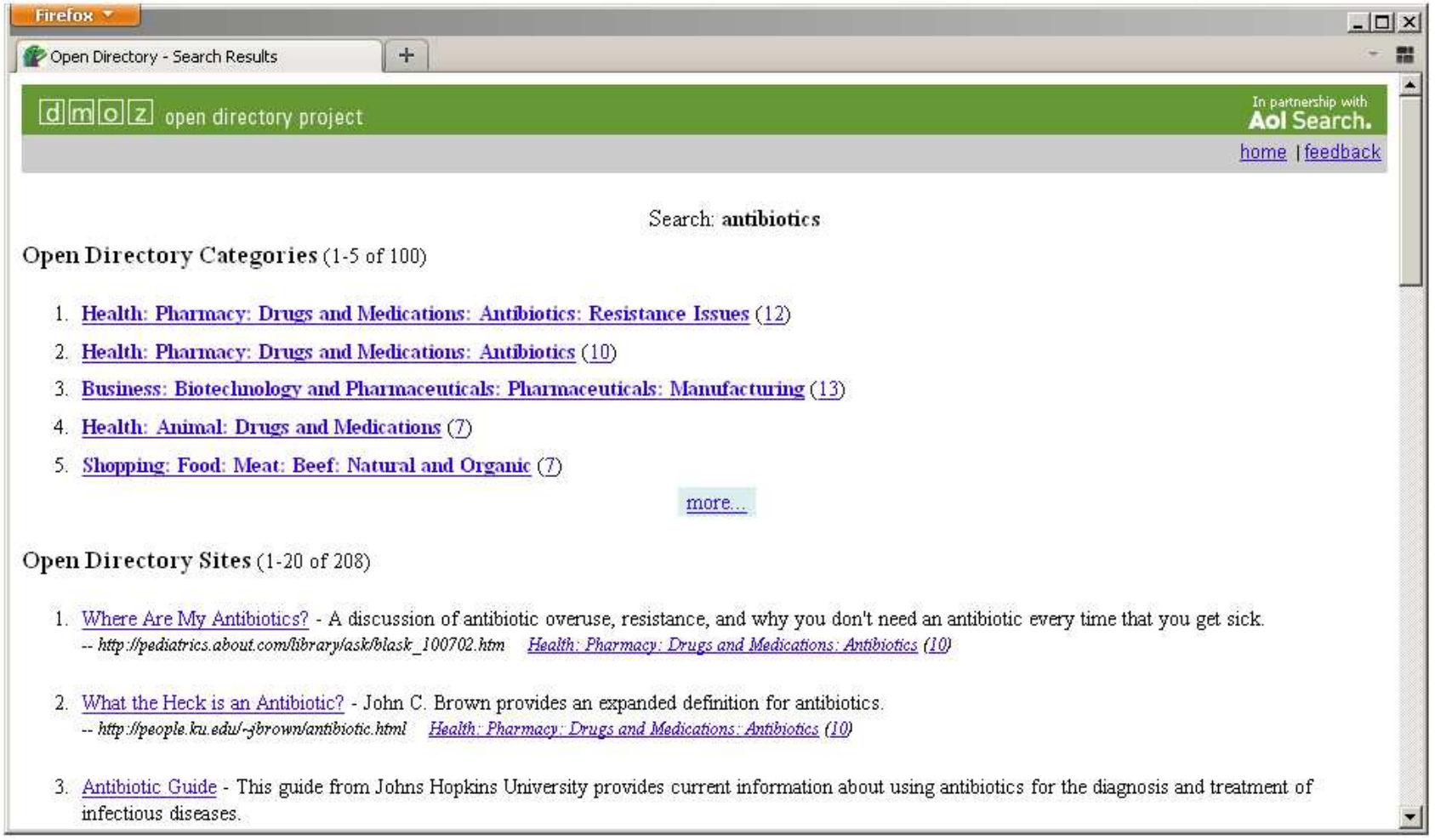



External Evaluation Measures

- Precision@N
- Match@N
- Mean Reciprocal Rank (MRR)

# Cluster Labeling

## External Evaluation: Human Experts



# Cluster Labeling

## External Evaluation

### Limitations

- Binary judgment about the relevance of a phrase is too strict.
- Used ranked-based measures are not sensitive regarding the order of phrases in a cluster label

# Cluster Labeling

## External Evaluation

Limitations

- Binary judgment about the relevance of a phrase is too strict.
- Used ranked-based measures are not sensitive regarding the order of phrases in a cluster label

Given a cluster about antibiotics; the reference label is “Antibiotics”, too.

Cluster Label Examples:

a) Web site, Technology, Infections, *Antibiotics*

b) *Antibiotics*, Infections, Web site, Technology

# Cluster Labeling

## NDCG-Based External Measure

Normalized Discounted Cumulative Gain (NDCG)

| Relevance Level | Definition |
|---|---|
| 0 | No match |
| 1 | Partial match |
| 2 | Exact match |

$$DCG@N = \sum_{i=1}^{N} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

# Cluster Labeling

## External Evaluation: Vocabulary Problem

People have a “tremendous diversity in the words” they use “to describe the same object”, and therefore systems may fail to answer the user’s information needs [4].

Thus, one cannot expect that a selected reference label is the *only correct description for* a cluster.

Example
Given a cluster about antibiotics; the reference label is “Antibiotics”, too.

- Is “Penicillin” really a poor label? No match!
- Is “Antimicrobial compound” really a poor label? No match!
- Is “Bactericidal Agents” really a poor label? No match!
- Is “Substance that kills bacteria” really a poor label? No match!

# Cluster Labeling [^]

## Internal Evaluation

Based on the relevance of a phrase, $rel(c,p) = \sum_{i=1}^{|\mathcal{F}|} \omega_i \cdot f_i(c,p)$, we can associate a quality value to each phrase.

Normalized Discounted Cumulative Gain (NDCG)

$$DCG@N = \sum_{i=1}^{N} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

| N | Phrase | $rel_i$ |
|---|---|---|
| 1 | Infections | 4 |
| 2 | Web site | 1 |
| 3 | Technology | 0 |
| 4 | **Antibiotics** | 5 |
| | NDCG@4 | 0.27 |

| N | Phrase | $rel_i$ |
|---|---|---|
| 1 | **Antibiotics** | 5 |
| 2 | Infections | 4 |
| 3 | Technology | 0 |
| 4 | Web site | 1 |
| | NDCG@4 | 0.45 |

# Cluster Labeling

## Paradigms of Cluster Labeling [1]

- Data-Centric Algorithms
- Description-Centric Algorithms
- Description-Aware Algorithms

# Cluster Labeling

## Data-Centric Algorithms

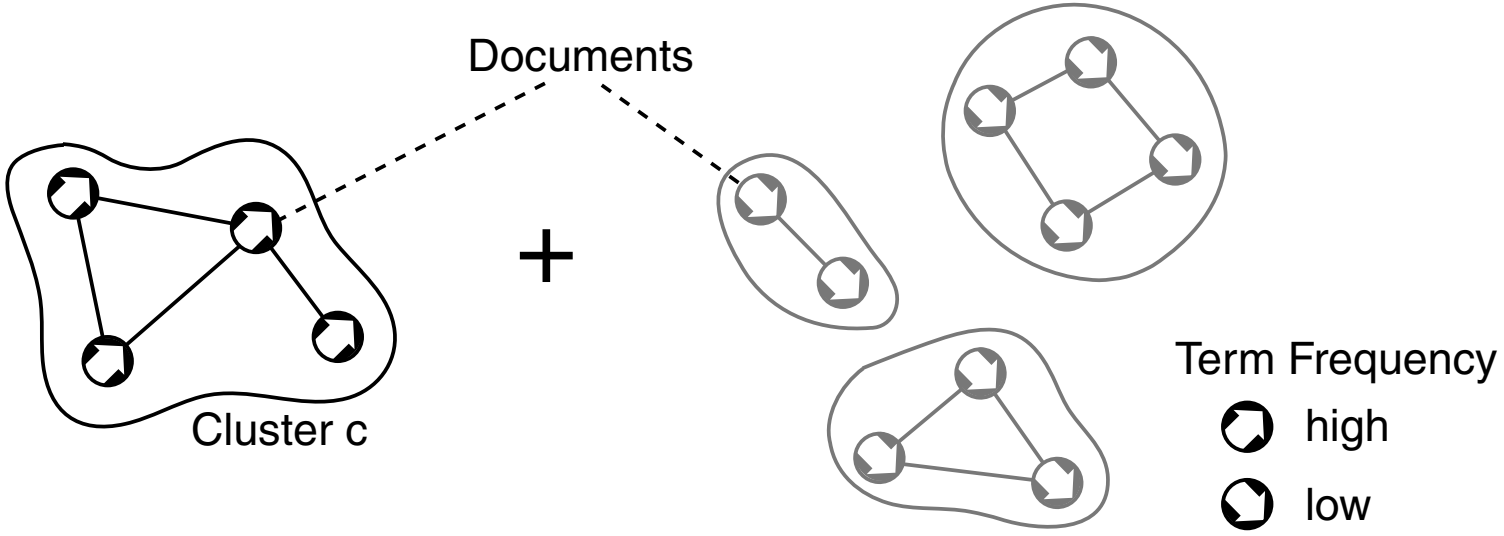

- Frequent Predictive Words (FPW) [7]
- Weighted Centroid Covering
- Scatter/Gather
- Tolerance Rough Set Clustering (TRSC)
- WebCAT
- Lassi

# Cluster Labeling

## Frequent Predictive Words

Terms $t$ are selected as cluster label from the cluster's centroid if they are

- very *frequent* within the cluster, and
- represent the cluster strongest (*predictive*).




Feature evaluation

$$f_c(t) = tf_c(t) \cdot \frac{tf_c(t)}{ctf}$$

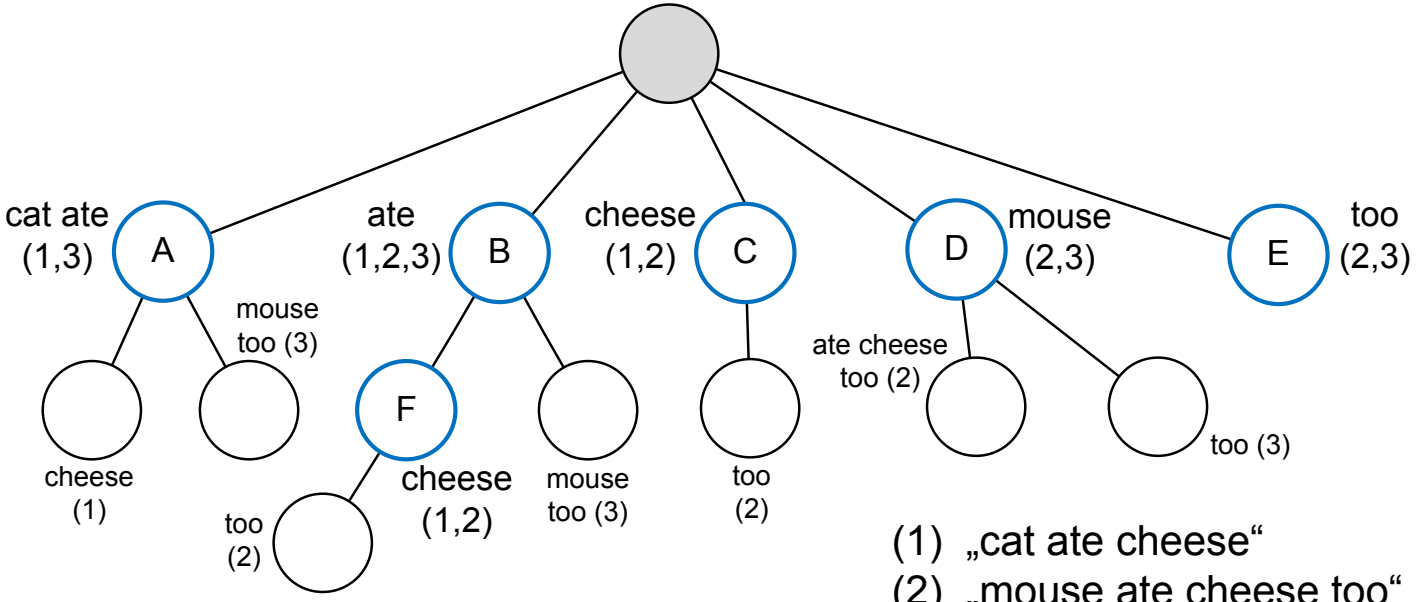
# Cluster Labeling

## Description-Aware Algorithms


- Suffix Tree Clustering (STC) [11]

# Cluster Labeling

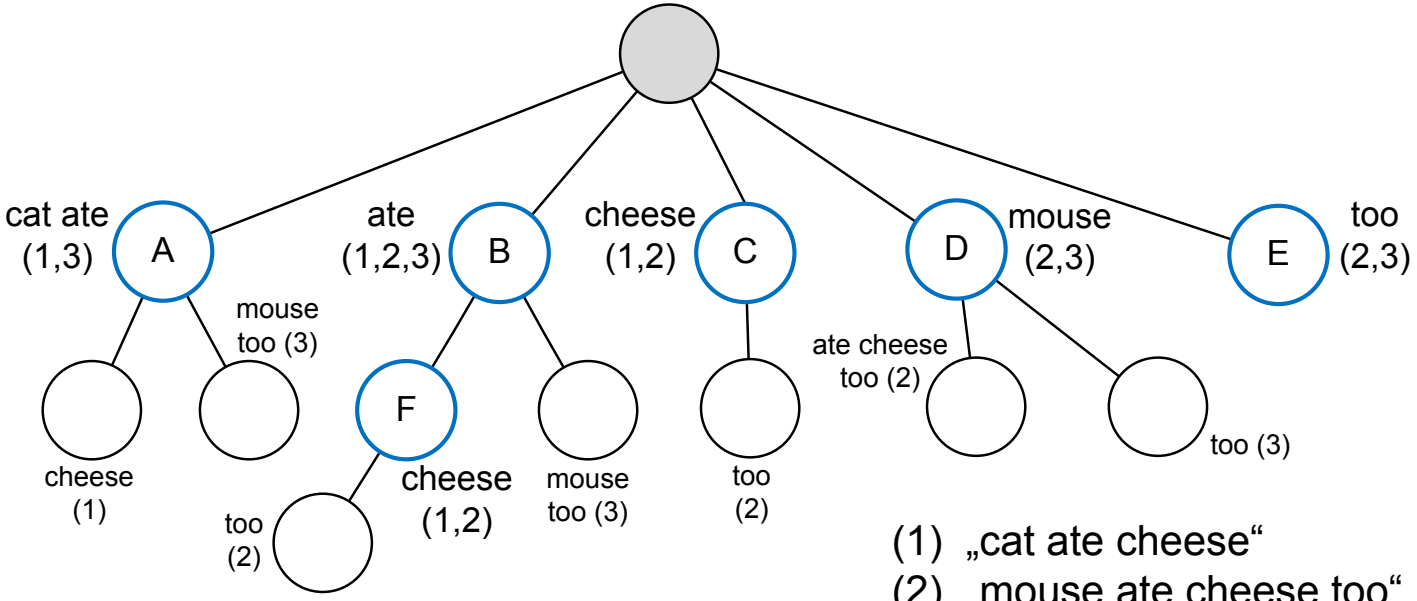
## Suffix Tree Clustering (STC)



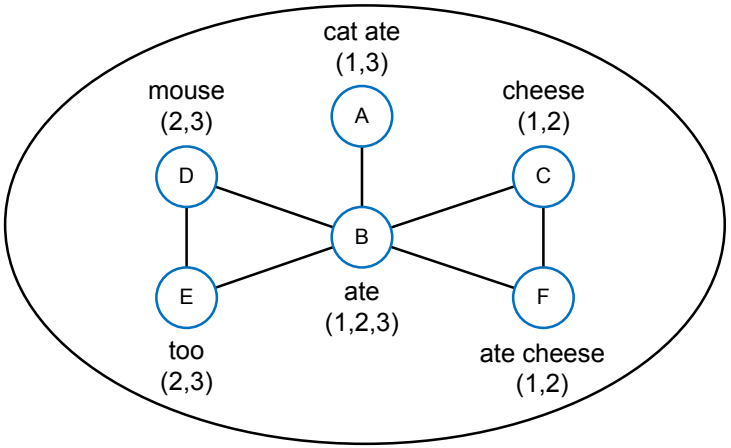

(1) „cat ate cheese“
(2) „mouse ate cheese too“
(3) „cat ate mouse too“

# Cluster Labeling

## Suffix Tree Clustering (STC)

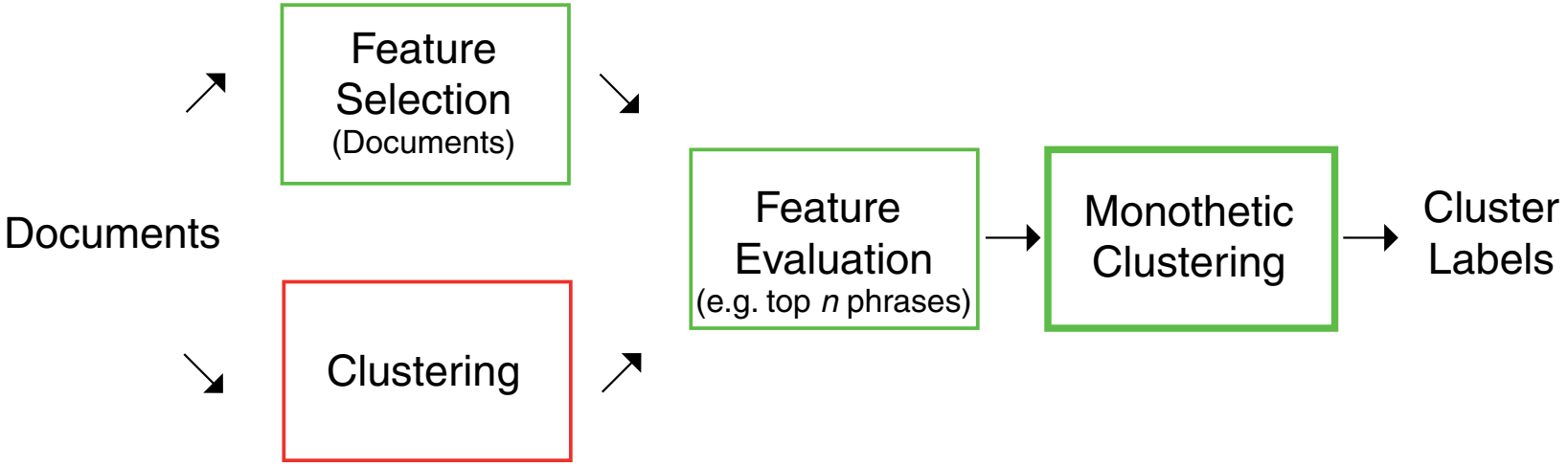



(1) „cat ate cheese“
(2) „mouse ate cheese too“
(3) „cat ate mouse too“



# Cluster Labeling

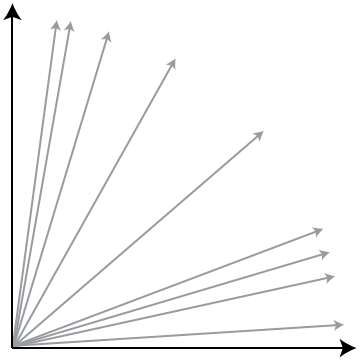
## Description-Centric Algorithms



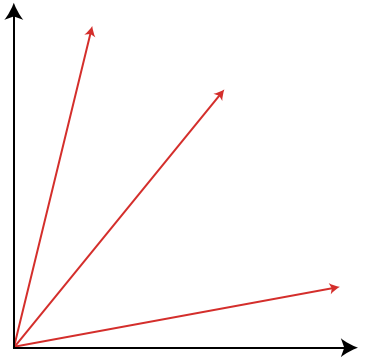

- Descriptive $k$-Means (DKM) [10]
- Lingo
- SRC
- DisCover

# Cluster Labeling

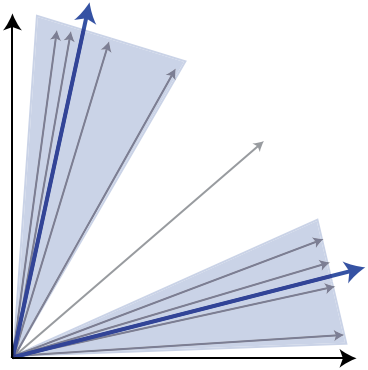
## Descriptive $k$-Means



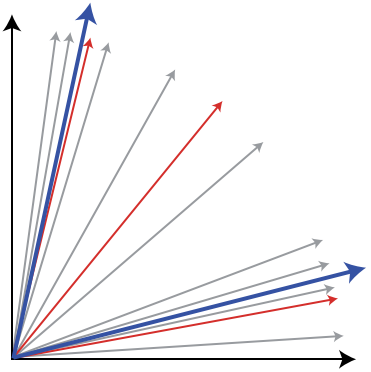
(1) Documents in vector space



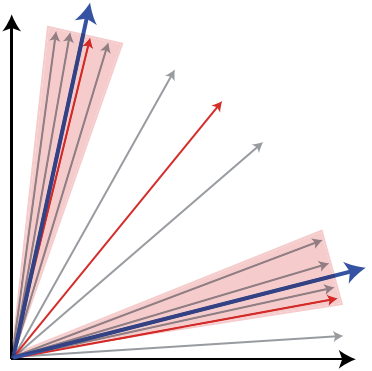
(2a) Feature selection: noun phrases



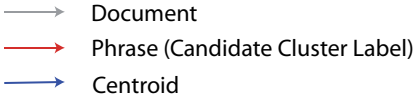
(2b) Clustering: centroids represent topics



(3) Feature evaluation. Phrases close to centroids become cluster label



(4) Monothetic clustering. Cluster label used as features



# Cluster Labeling [^]

## Paradigms of Cluster Labeling: Examples

| Category | Paradigm | Cluster Labels |
|---|---|---|
| MySQL | FPW | excel, jeremy, demo, authentic, forum |
| | STC | MySQL, Open Source Database, News, Search |
| | DKM | SQL Server, MySQL database server |
| PostgreSQL | FPW | hat, document, project, string, release |
| | STC | Support, Contact, Open Source, Search |
| | DKM | PostgreSQL database system, PostgreSQLServer |
| Antibiotics | FPW | antibiotics, disease, infection, bacteria, drug |
| | STC | Skip, Navigation, News, Search |
| | DKM | Antibiotic Resistant Bacteria |

# Cluster Labeling

## Experiments

Data set

- Open Directory Project (ODP)
- 5 selected categories ($\approx$ 250 documents)
- Example: Movies of Stanley Kubrick and Alfred Hitchcock

Evaluation

- Each criterion was evaluated separately
- NDCG-based internal measure
- Precision@N, Match@N

# Cluster Labeling

## Results

| Paradigm | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ |
|---|---|---|---|---|---|
| Keyphrase Extraction | 0.79 | 0.66 | 0.37 | 0.94 | 0.99 |
| Data-Centric Algorithms | 0.39 | 0.59 | 0.63 | 0.97 | **1.00** |
| Description-Aware Algorithms | 0.73 | **0.70** | 0.89 | **1.00** | 0.99 |
| Description-Centric Algorithms | **0.91** | 0.64 | **0.91** | **1.00** | **1.00** |

- $f_1$ Comprehensibility
- $f_2$ Descriptiveness
- $f_3$ Discriminative Power
- $f_5$ Uniqueness
- $f_6$ Non-redundancy

For example, comprehensibility:

$$f_{1|\text{all}}(\mathcal{L}) = \frac{1}{k} \sum_{c \in \mathcal{C}} \frac{1}{|l_c|} \sum_{p \in l_c} \text{NP}(p) \cdot \text{penalty}(p)$$

# Cluster Labeling [^]

## Results

- Using noun phrases yields to a better label quality.
- Using a reference clustering improves the label quality, too.
- Simple keyphrase-extraction techniques are competitive with data-centric algorithms.
- Description-centric algorithms achieve the best results.

# Cluster Labeling [^]

## Recap and Outlook

Recap

- Formalization of Cluster Label Properties
- Evaluation of Cluster Labels
- Paradigms of Cluster Labeling

Outlook

- Evaluate the effect of each cluster label contraint on the quality of a label.
- Considering new keyphrase extraction methods in addition to noun phrases and frequent phrases.

# Bibliography


[1] C. Carpineto, S. Osiński, G. Romano, D. Weiss. A Survey of Web Clustering Engines. *ACM Computing Surveys (CSUR)*, 41(3):Article 17, 2009.

[2] C. Clifton, R. Cooley, and J. Rennie. TopCat: Data Mining for Topic Identification in a Text Corpus. *IEEE Trans. Knowl. Data Eng.*, 16(8):949–964, 2004.

[3] W. de Winter and M. de Rijke. Identifying Facets in Query-Biased Sets of Blog Posts. In *Proceedings of ICWSM 2007*, pages 251–254.

[4] S.T. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, and R. Harshman. Using Latent Semantic Analysis to Improve Access to Textual Information. In *Proceedings of CHI 1988*, pages 281–285.

[5] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani. Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution. In *Proceedings of SPIRE 2006*, pages 25–36.

[6] S. Osiński, J. Stefanowski, and D. Weiss. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Proceedings of IIPWM 2004*, pages 359–368.

[7] A. Popescul and L.H. Ungar. “Automatic labeling of document clusters”. `http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf`, 2000.

[8] B. Stein and S. Meyer zu Eißen. Topic Identification: Framework and Application. In *Proceedings of I-Know 2004*, pages 353–360.

[9] H. Toda and R. Kataoka. A Clustering Method for News Articles Retrieval System. In *Proceedings of WWW 2005*, pages 988–989.

[10] D. Weiss. “Descriptive clustering as a method for exploring text collections”. Ph.D. dissertation. Poznań University of Technology, Poland, 2006.

[11] O. Zamir and O. Etzioni. Grouper: A dynamic Clustering Interface to Web Search Results. In *Proceedings of WWW 1999*, pages 1361–1374.