Konzeption und Realisierung eines Systems zur Informationssuche in einem Dokumentenarchiv basierend auf Textinhalt und Metadaten

Bachelor-Vortrag Annika Kremer





### 1. Einleitung

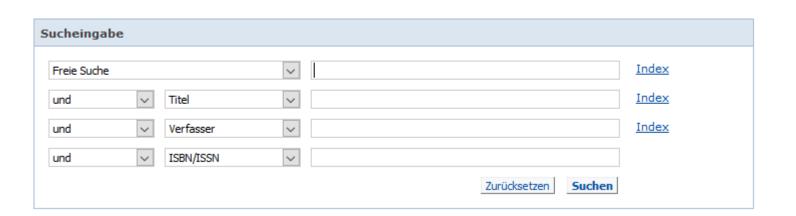




- Information Retrieval
- unstrukturierte/semistrukturierte Dokumente
- Beispiele:
  - Onlinekatalog Bibliothek
  - Websuche
  - E-Mail Posteingänge







## 2. Problemstellung



- Hauptanwendungsfall E-Mail Postfach
- Semistrukturierte Dokumente
  - Metadaten (z.B. Datum, Absender...)
  - Freitext
- Logische Verknüpfung der Suchkriterien mit AND, OR, NOT in beliebig tiefer Schachtelung



## 2. Problemstellung

```
Datei Bearbeiten Format Ansicht ?

(absender ("<maximilianSchuster@hochschule-trier.de>"))
(EmpfaengerNamen ("konradMueller@hochschule-trier.de"))
(Betreff ("Abgabe Hausarbeit"))
(datum ("Mon, 23 May 2017"))
(anzahlAnhaenge 1)
```

## 2.Problemstellung

```
(EMAIL-TYP "inbox")

;Freitext

Sehr geehrter Professor Müller,

im Anhang finden Sie meine Hausarbeit.

Mit freundlichen Grüßen

Maximilian Schuster
```

#### 3. Boolesches Retrieval



- Einsatz in Metadatensuche
- Charakterisieren von Mengen durch Attribute

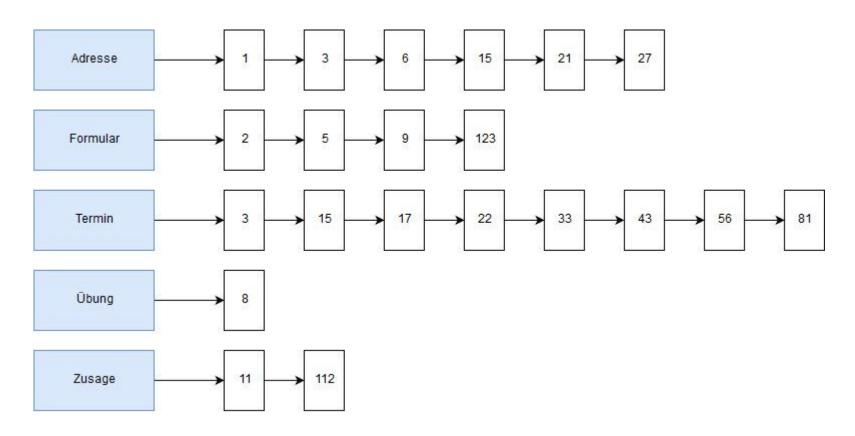
$$t: D \to T, t(d) = t_i$$
$$D_{t,t_i} = \{d \in D | t(d) = t_i\}$$

- Mengenoperationen
  - Durchschnitt AND  $D_{t,t_1} \cap D_{s,s_1}$
  - Vereinigung OR  $D_{t,t_1} \cup D_{s,s_1}$
  - Differenz NOT  $D \setminus D_{t,t_1}$

#### 3. Boolesches Retrieval



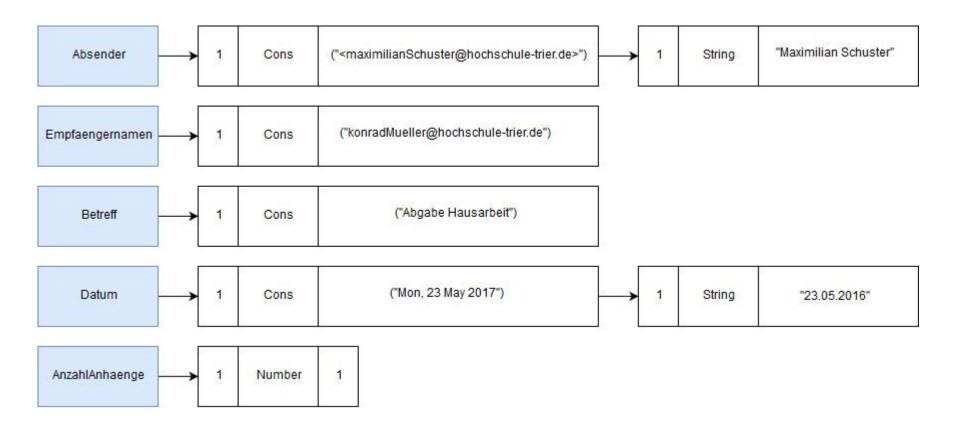
• Invertierte Liste → Speichern von *docIDs* 



#### 3. Boolesches Retrieval



#### Modifizierte invertiere Liste für Metadatensuche



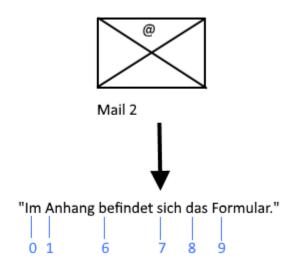


- Einsatz in Freitextsuche
  - Bestimmen der Übereinstimmung
  - Teiltreffer möglich
- Grundidee:
  - Repräsentation von Dokumenten und Anfragen als Vektoren
  - Ähnlichkeiten zwischen Vektoren bestimmen
    - → Ähnlichkeitsfunktion
    - → Ähnlichkeitswert
- Erlaubt Ranking



## • Indexierung der Terme

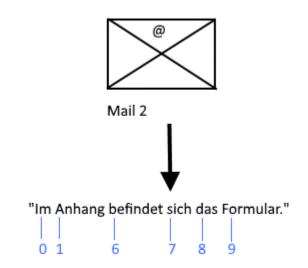






Indexierung der Terme

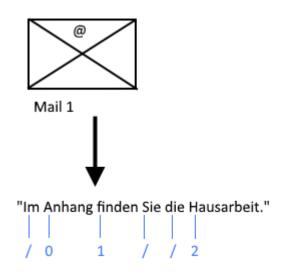


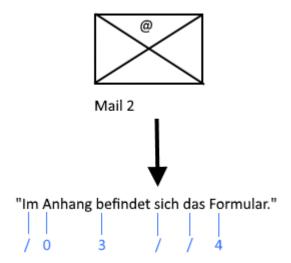


• Problem: Viele überflüssige Terme



- Lösung: Stoppwörter (engl. stop words)
- Indexierung mit Stoppwörtern:





## Dokumentvektoren mit Termhäufigkeiten



$$\overrightarrow{d1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$



$$\overrightarrow{d2} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

## Dokumentvektoren mit Termhäufigkeiten



$$\overrightarrow{d1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$



$$\overrightarrow{d2} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Problem: Wichtigkeit des Terms?



• Lösung: zusätzlich Dokumenthäufigkeit  $df_t$ 

• invertierte Dokumenthäufigkeit 
$$idf_t = \frac{1}{df_t}$$

• mit Dämpfung 
$$idf_t = \log\left(\frac{N}{df_t}\right)$$

• TF-IDF-Gewichtung 
$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

• Bei N = 2: 
$$\overrightarrow{df} = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$
  $\log\left(\frac{2}{df_t}\right)$   $\overrightarrow{idf} = \begin{pmatrix} 0 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{pmatrix}$ 

- Erster Term in jedem Dokument enthalten
  - → Gewicht 0 weil nicht aussagekräftig

Dokumentvektoren mit TF-IDF-Gewichtung

$$\vec{d1} = \begin{pmatrix} 0 \\ 0,3 \\ 0,3 \\ 0 \\ 0 \end{pmatrix} \qquad \vec{d2} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0,3 \\ 0,3 \end{pmatrix}$$

Repräsentation der Anfrage durch Anfragevektor

• Beispiel: "Formular"

$$\vec{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0, 3 \end{bmatrix}$$

• Ähnlichkeitsfunktion: Cosinus-Maß

$$\frac{\overrightarrow{q} \cdot \overrightarrow{d2}}{|\overrightarrow{q}| \cdot |\overrightarrow{d2}|}$$

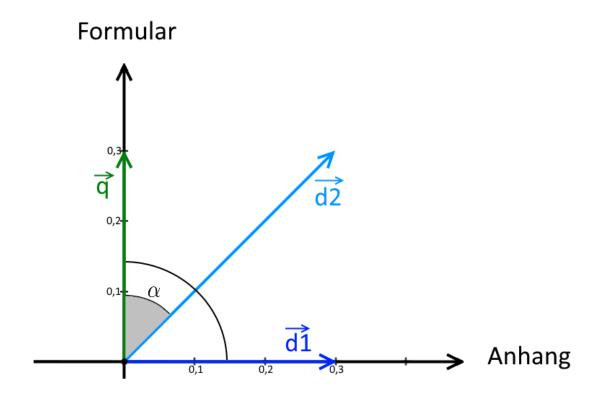
$$sim(\vec{q}, \vec{d1}) = \frac{\vec{q} \cdot \vec{d1}}{|\vec{q}| \cdot |\vec{d1}|} = \frac{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0, 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}}{\sqrt{0,3^2 + \sqrt{0,3^2 + 0,3^2}}} = \frac{0}{0,7243} = 0$$

$$sim(\vec{q}, \vec{d2}) = \frac{\vec{q} \cdot \vec{d2}}{|\vec{q}| \cdot |\vec{d2}|} = \frac{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0, 3 \\ 0, 3 \end{pmatrix} \begin{pmatrix} 0 \\ 0, 3 \\ 0, 3 \end{pmatrix}}{\sqrt{0,3^2 + \sqrt{0,3^2 + 0,3^2}}} = \frac{0,09}{0,7243} = 0,1243$$

$$sim(\vec{q}, \vec{d1}) = \frac{\vec{q} \cdot \vec{d1}}{|\vec{q}| \cdot |\vec{d1}|} = \frac{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0, 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}}{\sqrt{0,3^2 + \sqrt{0,3^2 + 0,3^2}}} = \frac{0}{0,7243} = 0$$

$$sim(\vec{q}, \vec{d2}) = \frac{\vec{q} \cdot \vec{d2}}{|\vec{q}| \cdot |\vec{d2}|} = \frac{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0, 3 \end{pmatrix} \begin{pmatrix} 0 \\ 0, 3 \\ 0, 3 \end{pmatrix}}{\sqrt{0,3^2 + \sqrt{0,3^2 + 0,3^2}}} = \frac{0,09}{0,7243} = 0,1243$$

- Längenunabhängige Ähnlichkeitsfunktion
- Cosinus-Maß



#### 5. Kombination



• AND: Durchschnitt der docIDs

• OR: Vereinigung der docIDs

• Problem: Score

```
    Bsp.:
```

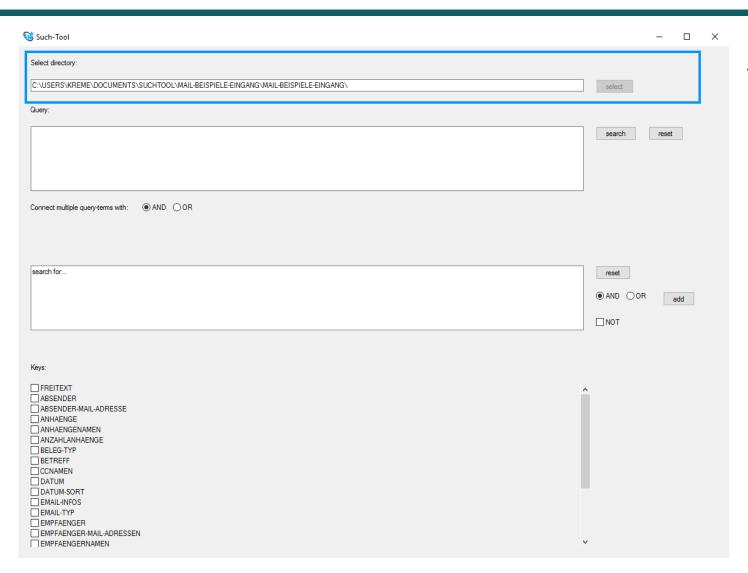
```
((Freitext = DFKI) OR
((Absender = Daniel) AND (Absender-Mail-Adresse = Daniel)))
```

- AND: Durchschnitt der docIDs
- OR: Vereinigung der docIDs
- Problem: Score
  - → Lösung: Metadatentreffer mit Score 1 versehen
  - → Addition der Score-Werte
  - → Kommazahl impliziert Freitexttreffer



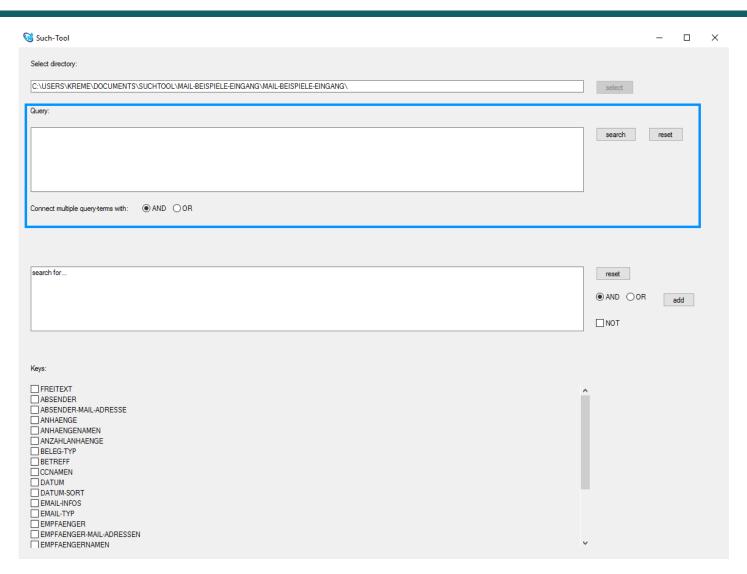
₩ Such-Tool	- 🗆 X
Select directory:	
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select
Query:	
	search reset
Connect multiple query-terms with:   AND OR	
search for	reset
	● AND ○ OR add
	□ NОТ
Keys:	
□ FREITEXT □ ABSENDER	^
□ ABSENDER-MAIL-ADRESSE □ ANHAENGE	
□ ANHAENGENAMEN □ ANZAHLANHAENGE	
BETREFF	
□ CCNAMEN □ DATUM	
DATUM-SORT	
□ EMAIL-INFOS □ EMAIL-TYP	
□ EMPFAENGER □ EMPFAENGERMALL-ADRESSEN □ EMPFAENGERMANEN	<b>~</b>
□ EMPFAENGERNAMEN	·





Verzeichnisauswahl





Gesamtanfrage



₩ Such-Tool	- 🗆 X	
Select directory:		
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select	
Query:		
	search reset	
Connect multiple query-terms with:   AND OR		
Conflict multiple query remis with.		
search for	reset	
	0.00	
	555	
	□NOT	
Keys:		
□ FREITEXT □ ABSENDER	^	Teila
□ ABSENDER-MAIL-ADRESSE □ ANHAENGE		Tella
ANHAENGENAMEN		
□ ANZAHLANHAENGE □ BELEG-TYP		
BETREFF		
□ CCNAMEN □ DATUM		
□ DATUM-SORT		
□ EMAIL-INFOS		
□ EMAIL-TYP □ EMPFAENGER		
☐ EMPFAENGER-MAIL-ADRESSEN		
□ EMPFAENGERNAMEN	V	

Teilanfrage



₩ Such-Tool	- 0	×
Select directory:		
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select	
Query:		
	search reset	
Connect multiple query-terms with:		
		_
search for		
seatch to	reset	
	● AND ○ OR add	
	□ пот	
Keys:		
□ FREITEXT □ ABSENDER	^	
□ ABSENDER-MAIL-ADRESSE □ ANHAENGE		
ANHAENGENAMEN		
□ ANZAHLANHAENGE □ BELEG-TYP		
□ BETREFF □ CCNAMEN		
DATUM		
□ DATUM-SORT □ EMAIL-INFOS		
□ EMAIL-TYP □ EMPFAENGER		
□ EMPFAENGER-MAIL-ADRESSEN		
□ EMPFAENGERNAMEN	V	

Suchbereiche auswählen



Keys:
☑ FREITEXT □ ABSENDER □ ABSENDER-MAIL-ADRESSE □ ANHAENGE □ ANHAENGENAMEN □ ANZAHLANHAENGE □ BELEG-TYP □ BETREFF □ CCNAMEN □ DATUM □ DATUM □ DATUM-SORT □ EMAIL-INFOS □ EMAIL-TYP □ EMPFAENGER □ EMPFAENGER □ EMPFAENGERNAMEN

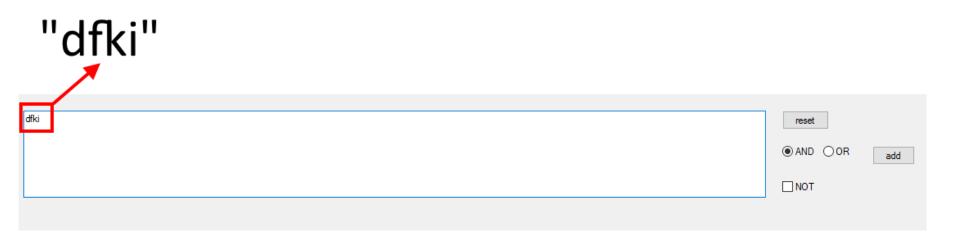
Freitextsuche auswählen



₩ Such-Tool	- 🗆 X
Select directory:  C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select
Query:	search reset
	searur reser
Connect multiple query-terms with:	
search for	reset
	● AND ○ OR add
Keys:	
☐ FREITEXT ☐ ABSENDER ☐ ABSENDER-MAIL-ADRESSE ☐ ANHAENGE ☐ ANHAENGEN ☐ ANHAENGEN ☐ BELEG-TYP	
□ BETREFF □ CCNAMEN □ DATUM □ DATUM-SORT	
☐ EMAIL-INFOS ☐ EMAIL-TYP ☐ EMPFAENGER	
☐ EMPFAENGER-MAIL-ADRESSEN ☐ EMPFAENGERNAMEN	V

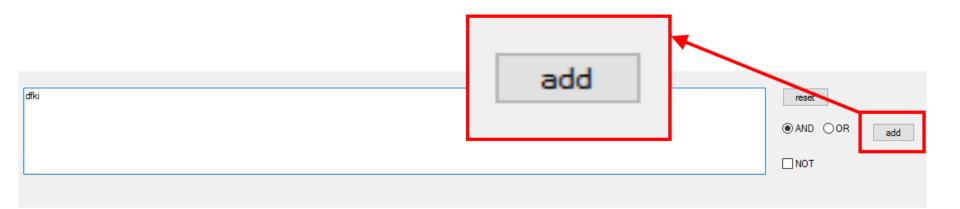


• Suchtext eingeben





• Teilanfrage hinzufügen

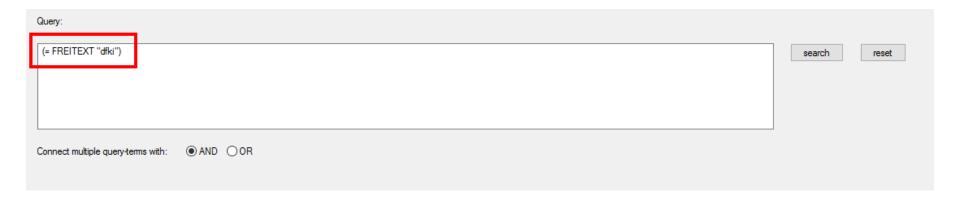




₩ Such-Tool	- D X
Select directory:	
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select
Query:	
(= FREITEXT "dfki")	search reset
Connect multiple query-terms with:   AND OR	
dfki	reset
	● AND ○ OR add
	□NOT
Keys:	
☐ FREITEXT	٨
□ ABSENDER □ ABSENDER-MAIL-ADRESSE	
□ ANHAENGE □ ANHAENGENAMEN	
□ ANZAHLANHAENGE □ BELEG-TYP	
□ BETREFF □ CCNAMEN □ CNAMEN	
□ DATUM □ DATUM-SORT □ EMAIL-INFOS	
□ EMAIL-TYP □ EMPFAENGER	
☐ EMPFAENGER-MAIL-ADRESSEN ☐ EMPFAENGERNAMEN	<b>~</b>



Anzeige





Anzeige

# (= FREITEXT "dfki")



## 6. Benutzeroberfläche und Bedienung



Query:	
(= FREITEXT "dfiki")	search reset
Connect multiple query-terms with:	

# 6. Benutzeroberfläche und Bedienung







<b>⋘</b> Such-Tool		_	_ ×	×
Select directory:				
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select			
Query:				
	search	reset		
Connect multiple query-terms with: ● AND ○ OR				
search for	reset			
	● AND ○ OR	ado	d	
	NOT			
Keys:  ☐FREITEXT				
☐ ABSENDER ☐ ABSENDER-MAIL-ADRESSE				
ANHAENGE ANHAENGENAMEN				
☐ ANZAHLANHAENGE ☐ BELEG-TYP				
□ BETREFF □ CCNAMEN				
□ DATUM □ DATUM-SORT				
□ EMAIL-INFOS □ EMAIL-TYP □ EMAIL-TYP				
□EMPFAENGER □EMPFAENGER-MAIL-ADRESSEN □EMPFAENGERNAMEN	<b>~</b>			
Learner and an				



FREITEXT   ABSENDER   ABSENDER-MAIL-ADRESSE   ANHAENGE   ANHAENGENAMEN   ANZAHLANHAENGE   BELEG-TYP   BETREFF   CCNAMEN   DATUM   DATUM   DATUM-SORT   EMAIL-INFOS   EMAIL-TYP   EMPFAENGER	Keys:
	☐ FREITEXT ☐ ABSENDER ☐ ABSENDER-MAIL-ADRESSE ☐ ANHAENGE ☐ ANHAENGENAMEN ☐ ANZAHLANHAENGE ☐ BELEG-TYP ☐ BETREFF ☐ CCNAMEN ☐ DATUM ☐ DATUM-SORT ☐ EMAIL-INFOS ☐ EMAIL-TYP
EMPFAENGER-MAIL-ADRESSEN EMPFAENGERNAMEN	EMPFAENGER EMPFAENGER-MAIL-ADRESSEN

Metadaten auswählen

Informatik
Hauptcampus

H O C H
S C H U L E
R

Select directory:  C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\  Query:  [= FREITEXT "dfk:")  Search reset  Connect multiple query-terms with: ○ AND ● OR	
Query:  [(= FREITEXT "dfki")  search reset  Connect multiple query-terms with: ○ AND ⑥ OR	
(= FREITEXT "dfki")  Connect multiple query-terms with: ○ AND	
Connect multiple query terms with: ○ AND	
	1
daniel	ı
● AND ○ OR add	ı
□ пот	ı
	L
Keys:	
☐ FREITEXT ☐ ABSENDER	
☑ ABSENDER-MAIL-ADRESSE □ ANHAENGE □ ANHAENGENAMEN	
□ ANZAHLANHAENGE □ BELEG-TYP	
□ BETREFF □ CCNAMEN	
□ DATUM □ DATUM-SORT	
□ EMAIL-INFOS □ EMAIL-TYP ■ CHINESTRUSTS	
□ EMPFAENGER □ EMPFAENGER-MAIL-ADRESSEN □ EMPFAENGERNAMEN  ✓	



• Erneut Suchtext eingeben



26.08.2017



• Hinzufügen



26.08.2017 42

Informatik
Hauptcampus

H O C H
S C H U L E
R

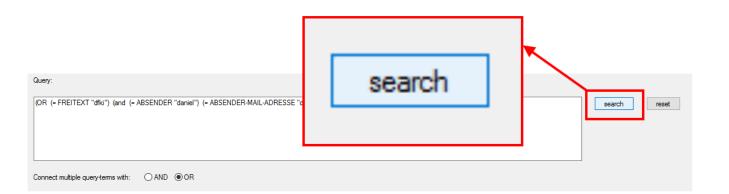
₩ Such-Tool	- 🗆 X
Select directory:	
C:\USERS\KREME\DOCUMENTS\SUCHTOOL\MAIL-BEISPIELE-EINGANG\MAIL-BEISPIELE-EINGANG\	select
Query:	
(OR (= FREITEXT "dfki") (and (= ABSENDER "daniel") (= ABSENDER-MAIL-ADRESSE "daniel")))	search reset
Connect multiple query-terms with: OAND OOR	
daniel	reset  ● AND ○ OR add  □ NOT
Keys:	
☐ FREITEXT  ☑ ABSENDER  ☑ ABSENDER-MAIL-ADRESSE  ☐ ANHAENGE  ☐ ANHAENGENAMEN  ☐ ANZAHLANHAENGE  ☐ BELEG-TYP  ☐ BETREFF  ☐ CCNAMEN  ☐ DATUM  ☐ DATUM-SORT  ☐ EMAIL-INFOS	
□ EMAIL-TYP □ EMPFAENGER □ EMPFAENGER-MAIL-ADRESSEN □ EMPFAENGERNAMEN	<b>v</b>



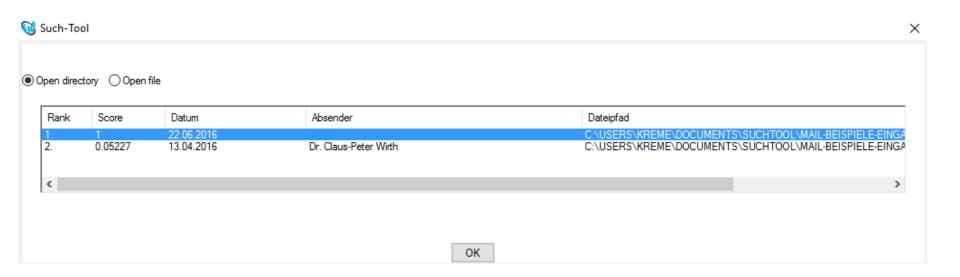
(OR (= FREITEXT "dfki") (and (= ABSENDER "daniel") (= ABSENDER-MAIL-ADRESSE "daniel")))



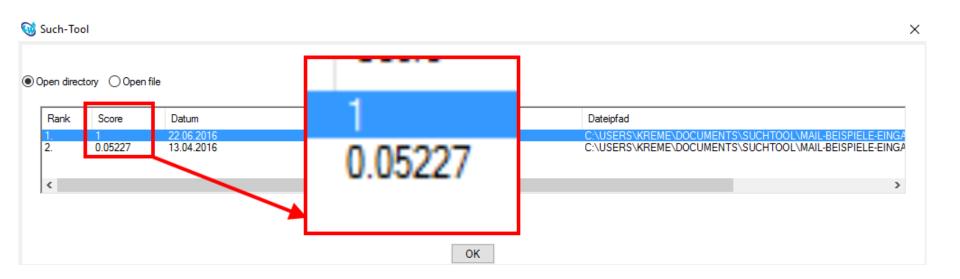












• Noch Fragen?



- https://de.wikipedia.org/wiki/Google
- http://www.underconsideration.com/brandnew/archives/new log o for bing.php
- •http://bibserv.fh-
- trier.de:8080/webOPACClient/search.do?methodToCall=start
- •http://www.arbeitstipps.de/email-postfach-aufraeumen-3-tipps.html
- •http://de.freepik.com/freie-ikonen/fragezeichen 731610.htm