

Nebenbedingungen in der Personensuche im Internet

Constraints in Web People Search

Johannes Kiesel

Bauhaus-Universität Weimar

23. April 2012

Nebenbedingungen in der Personensuche im Internet

Personensuche im Internet

- Einführung

- Aktuelle Systeme

- Vorgehensweise

Nebenbedingungen durch Personenattribute

- Kardinalität von Personenattributen

- Generierung von Nebenbedingungen

Experimente & Ergebnisse

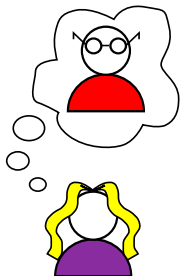
- Forschungsfragen

- Methodik

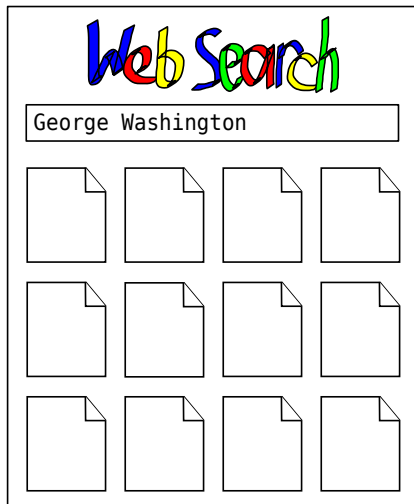
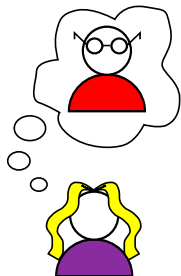
- Ergebnisse

- Ausblick

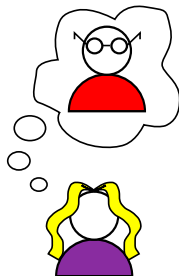
Was ist „Personensuche im Internet“?



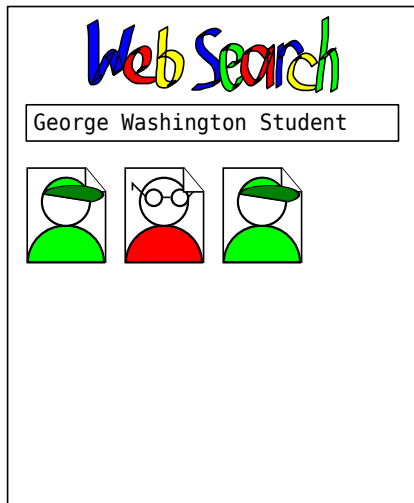
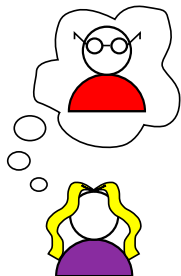
Was ist „Personensuche im Internet“?



Was ist „Personensuche im Internet“?



Was ist „Personensuche im Internet“?



My Saved Searches

[Clear Form](#)Full Name All name matchesJob Title
Company Name / URL / Ticker
Industry Keywords / SIC Codes
City / State / ZIP
 Person

Additional Filters

 x x Include People with Partial Profiles x Company Location[People](#)[Companies](#)[Home](#)

ZoomInfo™ Directory

1 2 3 4 5 6 7 8 9 Next ▶

[Add to List](#) [Export](#) [Set Alert](#) [Save Search](#) [Print](#)**787** People matching your criteria

Sort Order

Date Updated ▼

Contact Info

All ▼

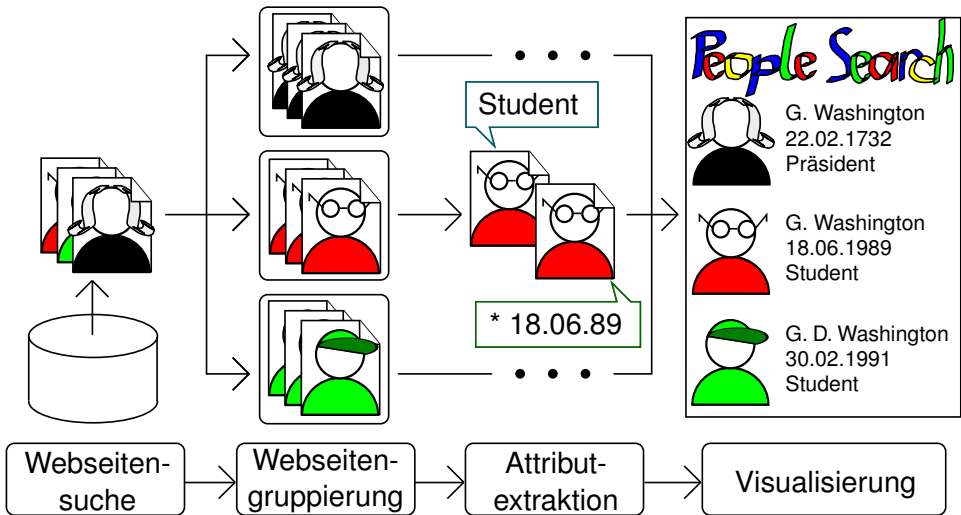
Last Update

No Limit ▼

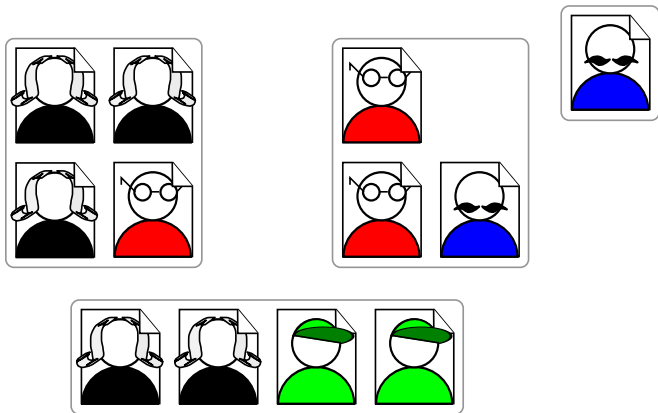
Results 1-25

<input type="checkbox"/>	George Washington	Washington Director DFW Elite Basketball	Email ✓ Phone ✓	3/3/12
<input type="checkbox"/>	George B. Washington	Attorney BAMN high school BAMN	Email ✓ Phone ✓	4/2/12
<input type="checkbox"/>	George Washington	Lab Technician for Southeastern Archeological Services Historic Kenmore		3/30/12
<input type="checkbox"/>	George E. Washington	President Invizion Inc	Email ✓ Phone ✓	3/29/12
<input type="checkbox"/>	George Washington	NIH-funded Researcher and Research Center Director University of Connecticut	Phone ✓	2/5/12

Vorgehensweise



Vorgehensweise



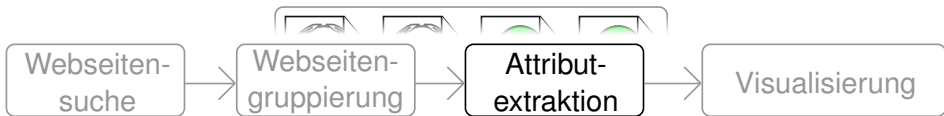
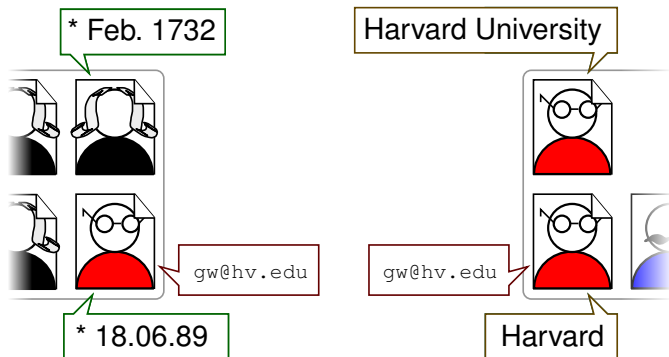
Webseiten-
suche

Webseiten-
gruppierung

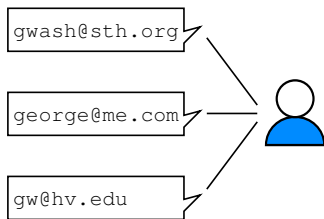
Attribut-
extraktion

Visualisierung

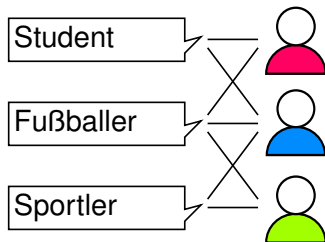
Vorgehensweise



Kardinalität von Personenattributen



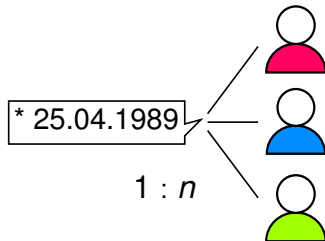
$n : 1$



$n : m$



$1 : 1$



$1 : n$

Die verwendeten Personenattribute

E-Mail

$n : 1$

$1 : 1$

Beruf

$n : m$

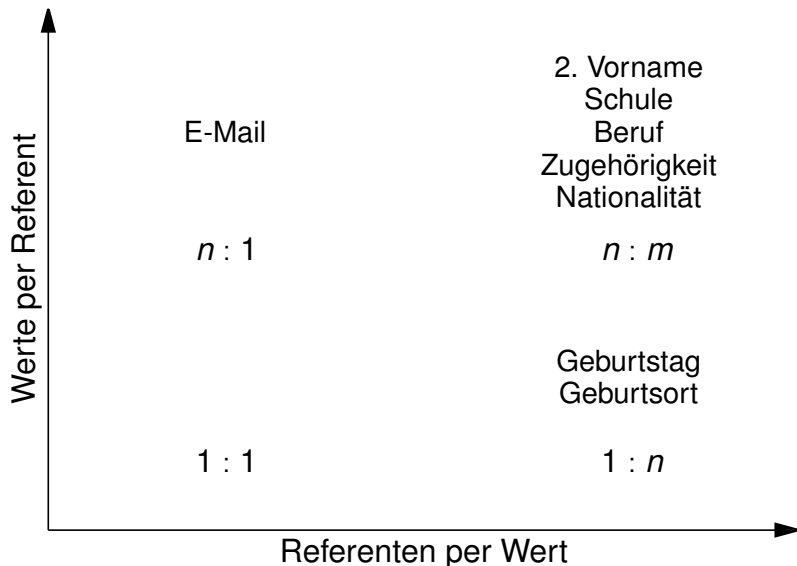
Geburtstag

$1 : n$

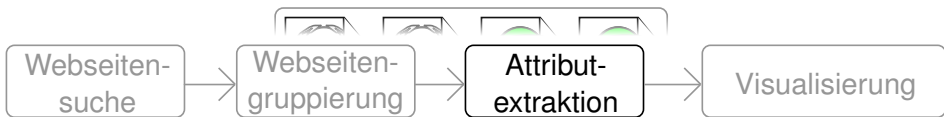
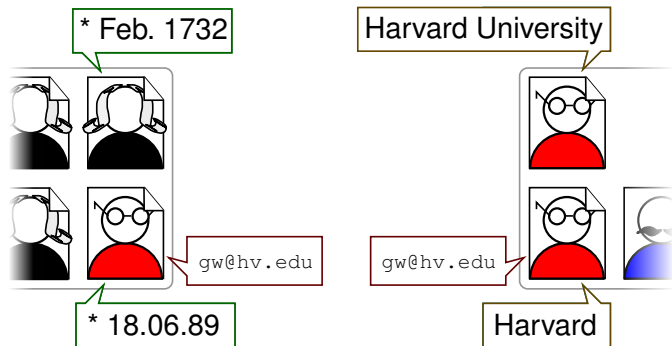
Die verwendeten Personenattribute

E-Mail	2. Vorname Schule Beruf Zugehörigkeit Nationalität
$n : 1$	$n : m$
	Geburtstag Geburtsort
$1 : 1$	$1 : n$

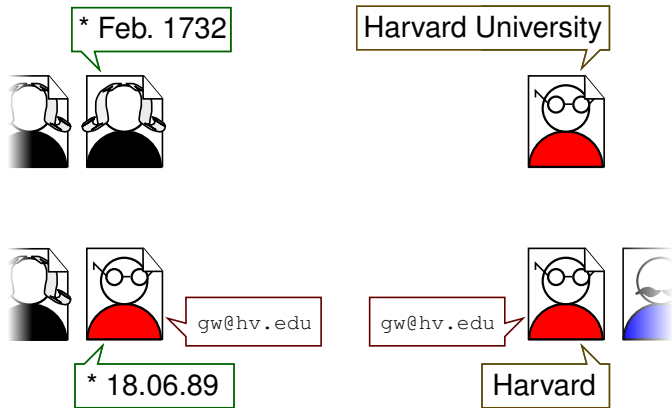
Die verwendeten Personenattribute



Generierung von Nebenbedingungen



Generierung von Nebenbedingungen



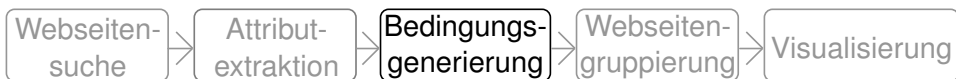
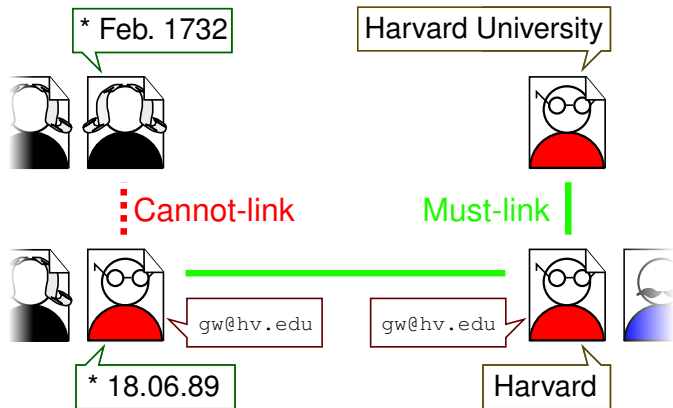
Webseiten-
suche

Attribut-
extraktion

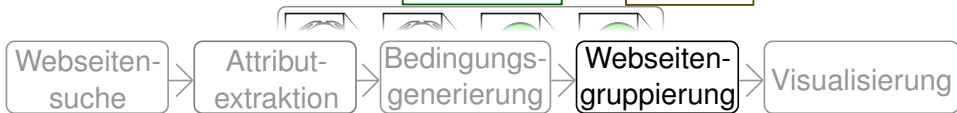
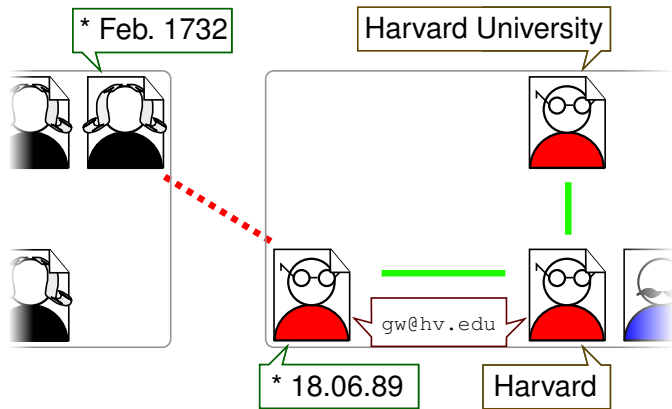
Webseiten-
gruppierung

Visualisierung

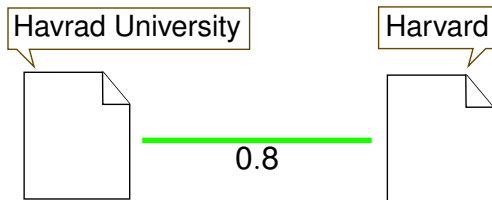
Generierung von Nebenbedingungen



Generierung von Nebenbedingungen



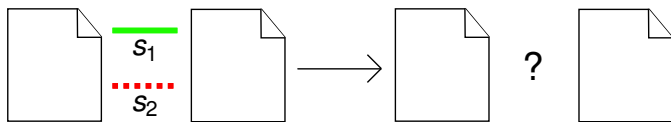
Abgleich von Attributwerten



Methoden:

- ▶ Exakter Abgleich
- ▶ Weicher Abgleich

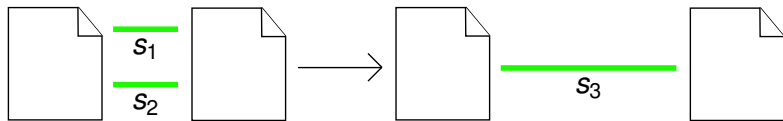
Addition der Nebenbedingungen



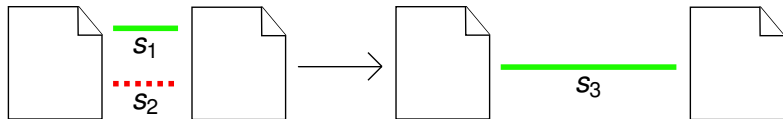
Wünschenswerte Eigenschaften:

- 1 Verstärkung
- 2 Abschwächung
- 3 Neutrale Bedingungen
- 4 Inverse Bedingungen
- 5 Assoziativität und Kommutativität
- 6 Typ-Gleichheit
- 7 Abgeschlossenheit
- 8 Asymptotische Grenzen
- 9 Widersprüchliche Bedingungen

Addition der Nebenbedingungen (forts.)

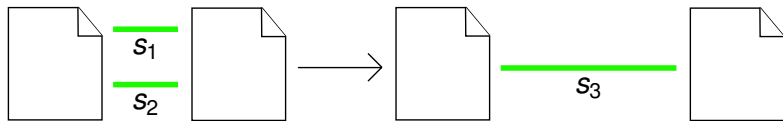


- ▶ Verstärkung: $s_3 \geq \max(s_1, s_2)$

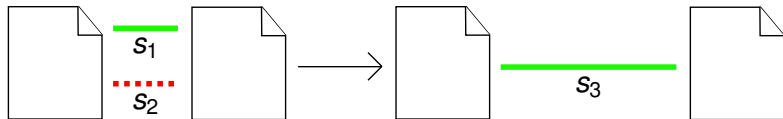


- ▶ Abschwächung: $s_3 \leq s_1$ wenn $s_1 > s_2$
- ▶ Inverse Bedingungen: $s_3 = 0$ wenn $s_1 = s_2$

Addition der Nebenbedingungen (forts.)



- ▶ Verstärkung: $s_3 \geq \max(s_1, s_2)$



- ▶ Abschwächung: $s_3 \leq s_1$ wenn $s_1 > s_2$
- ▶ Inverse Bedingungen: $s_3 = 0$ wenn $s_1 = s_2$

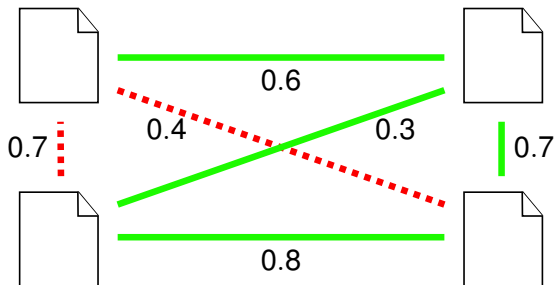
Methoden:

- ▶ Maximum
- ▶ Multiplikation

Zusätzlich:

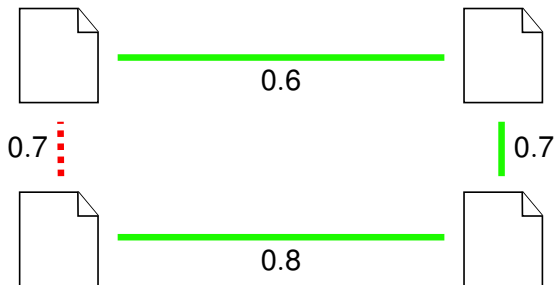
Gewichtung der Attribute

Transitivität und Konflikte



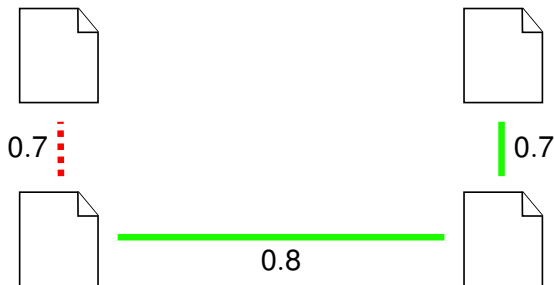
1. Anwendung eines Schwellwertes
2. Konfliktbehebung:
 - ▶ Übernahme der Must-links
 - ▶ Schwellwerterhöhung

Transitivität und Konflikte



1. Anwendung eines Schwellwertes
2. Konfliktbehebung:
 - ▶ Übernahme der Must-links
 - ▶ Schwellwerterhöhung

Transitivität und Konflikte



1. Anwendung eines Schwellwertes
2. Konfliktbehebung:
 - ▶ Übernahme der Must-links
 - ▶ Schwellwerterhöhung

Zusammenfassung:

Generierung und Anwendung von Nebenbedingungen

1. Abgleich von Attributwerten
 - ▶ Exakter Abgleich
 - ▶ Weicher Abgleich
2. Addition von Nebenbedingungen
 - ▶ Maximum
 - ▶ Multiplikation
3. Konfliktbehebung
 - ▶ Übernahme der Must-links
 - ▶ Schwellwerterhöhung
4. Gruppierung (Clustering)
 - ▶ Single Pass Clusterer
 - ▶ Hierarchischer Agglomerativer Clusterer (Single Link)

Forschungsfragen

- ▶ Wie präzise sind die generierten Nebenbedingungen?
- ▶ Wie hilfreich sind Nebenbedingungen durch einzelne Personenattribute bei der Personensuche?
- ▶ Ist es vorteilhaft mehrere Personenattribute zur Generierung von Nebenbedingungen zu nutzen?
- ▶ Kann die Qualität des Ergebnisses durch eine Gewichtung der Personenattribute erhöht werden?
- ▶ Wie verändert sich die Qualität durch die Nutzung von algorithmisch extrahierten Personenattributen?

Verwendeter Korpus: WePS-2

- ▶ 30 Anfragen („Vorname Nachname“) an Yahoo!
- ▶ Je 10 Namen von 3 verschiedenen Quellen
- ▶ Manuelle Annotation (Referent, Attributewerte) und Verwerfung der erhaltenen Dokumente

Teil des Korpus	D		R	
	TN	μ_{tn}	TN	μ_{tn}
Englische Wikipedia	940	94,0	107	10,7
ACL'08	816	81,6	142	14,2
1990 US Zensus	802	80,2	303	30,3
Gesamter Korpus	2558	85,3	552	18,4

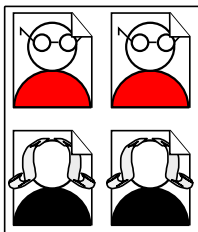
- ▶ TN: Wert in Bezug auf den Korpus („Korpusweit“)
- ▶ μ_{tn} : Durchschnittlicher Wert für die Dokumente einzelner Anfragen („Problemweit“)

Verwendeter Korpus: WePS-2 (Personenattribute)

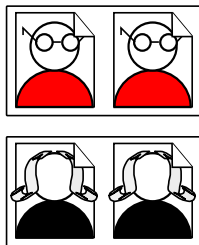
Attribut	$\frac{ V }{ D }$	Werte per Referent		Referenten per Wert	
		TN	μ_{tn}	TN	μ_{tn}
Beruf	1,07	4,42	9,60	1,44	1,04
Zugehörigkeit	1,04	4,92	8,36	1,03	1,00
Schule	0,17	2,24	2,81	1,10	1,01
Geburtsort	0,10	1,69	2,44	1,05	1,00
Geburtstag	0,10	1,12	1,11	2,22	1,05
2. Vorname	0,09	1,06	1,02	1,87	1,07
Nationalität	0,08	1,30	1,18	1,30	1,00
E-Mail	0,07	1,29	1,03	1,01	1,01

Evaluation des Clusterings: BCubed $F_{\alpha=0,5}$ -Measure

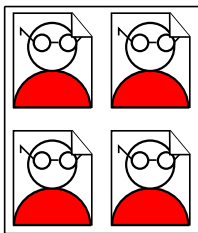
Homogenität



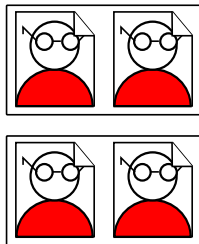
→
Höhere
Qualität



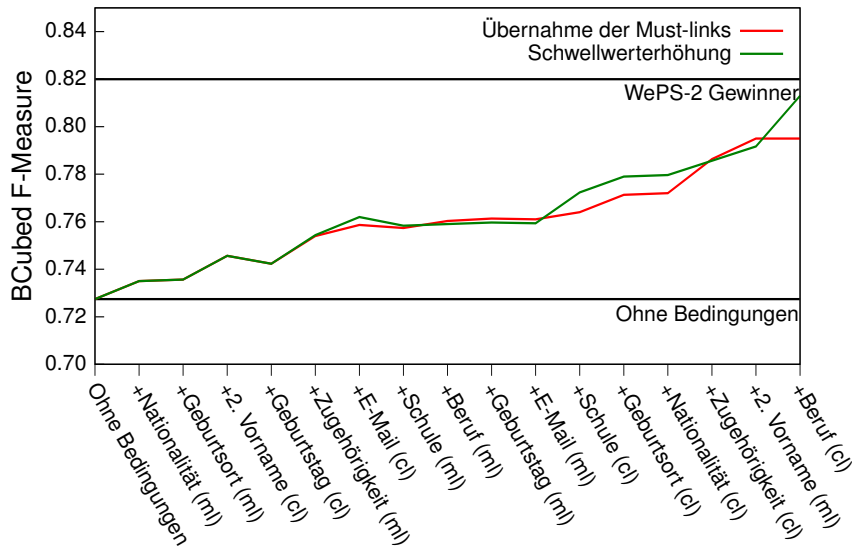
Vollständigkeit



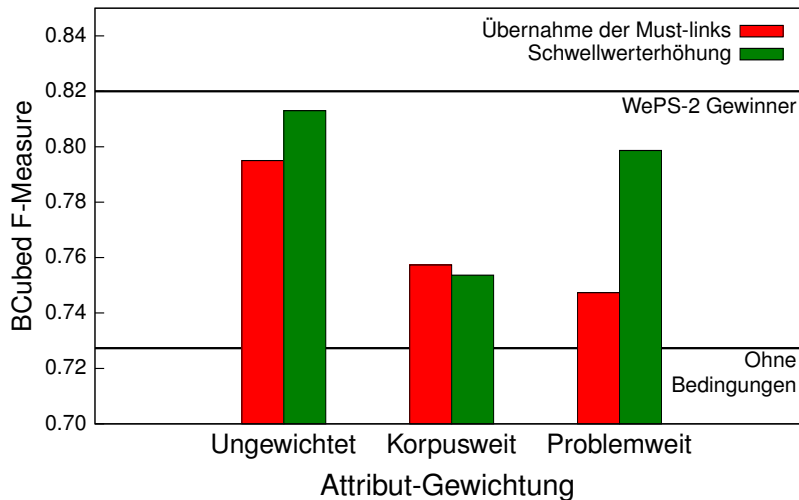
←
Höhere
Qualität



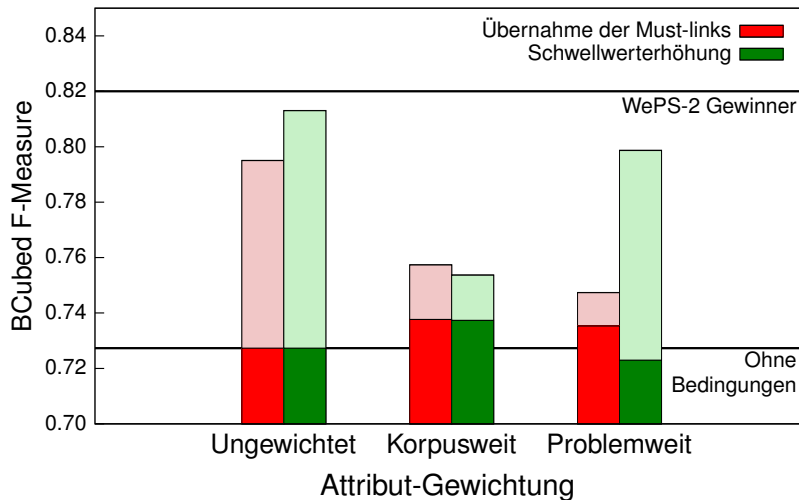
Auswirkungen der Nebenbedingungen



Gewichtung der Personenattribute



Algorithmisch extrahierte Personenattribute



Zusammenfassung

- ▶ Anwendung von *Constrained Clustering* in der Personensuche im Internet
- ▶ Verwendung von unsicheren Informationen von verschiedenen Personenattributen
- ▶ Vorschlag eines generischen Systems zur Generierung von Nebenbedingungen

Verfahren	$F_{\alpha=0,5}$
PolyUHK	0,82
Thesis	0,81
UVA	0,81
ITC-UT	0,81
XMEDIA	0,72
UCI	0,71

Ausblick

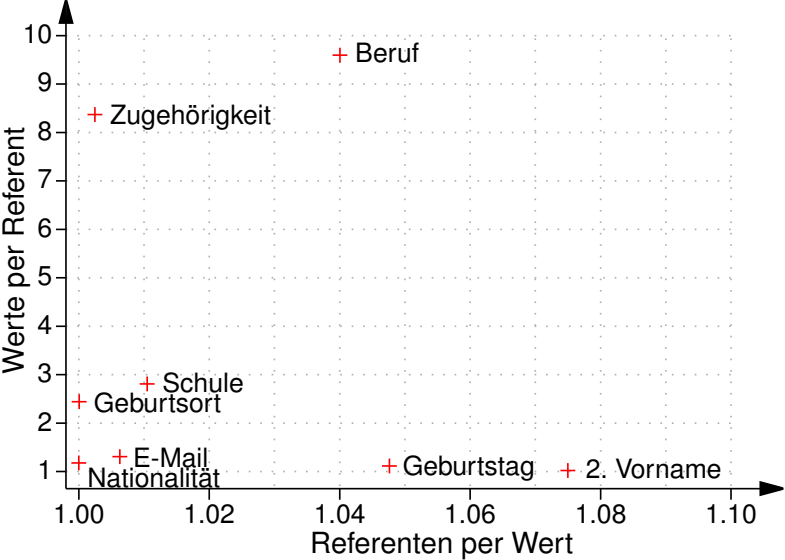
- ▶ Anwendung der gewichteten Nebenbedingungen zur Veränderung der Dokumentähnlichkeit
- ▶ Testen verschiedener Methoden zur automatischen Bestimmung von Schwellwerten
- ▶ Genauere Analyse der verschiedenen Kombinationen von Personenattributen

Ausblick

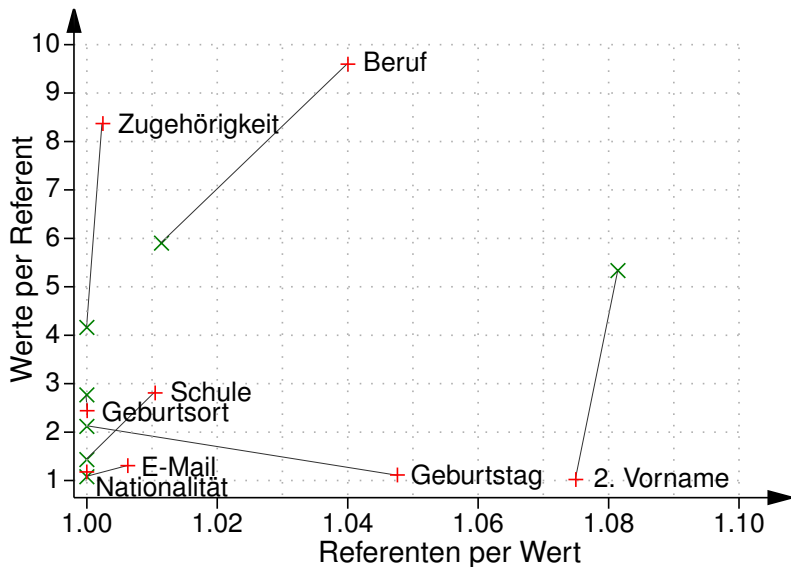
- ▶ Anwendung der gewichteten Nebenbedingungen zur Veränderung der Dokumentähnlichkeit
- ▶ Testen verschiedener Methoden zur automatischen Bestimmung von Schwellwerten
- ▶ Genauere Analyse der verschiedenen Kombinationen von Personenattributen

Vielen Dank für Ihre Aufmerksamkeit

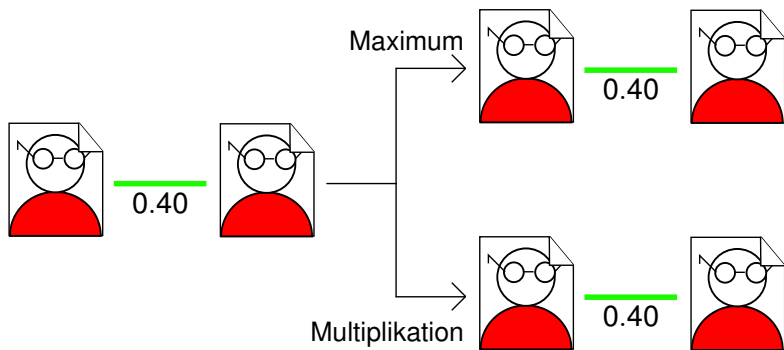
Referenten per Wert/Werte per Referent



Referenten per Wert/Werte per Referent

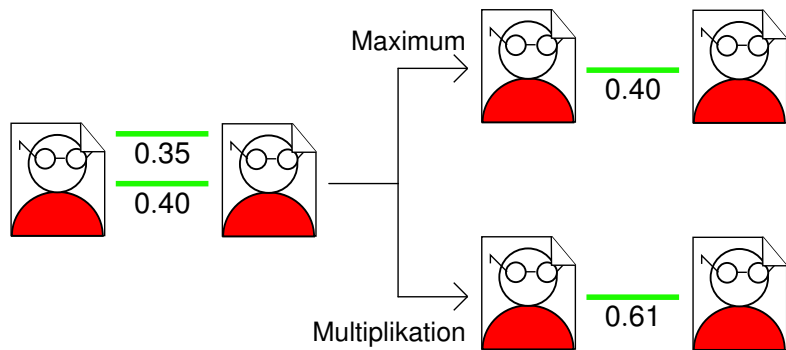


Addition von Nebenbedingungen



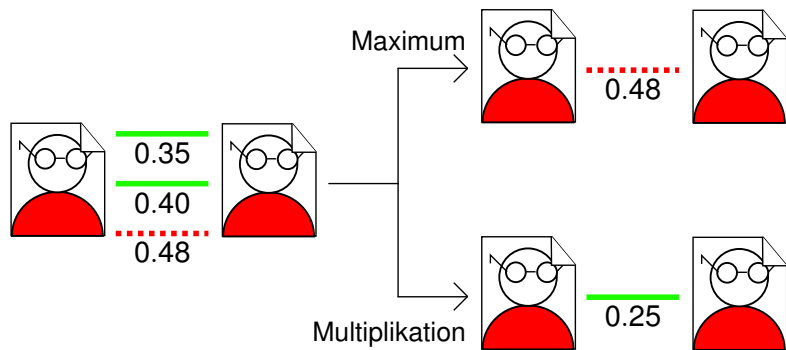
$$1 - (1 - 0.40)$$

Addition von Nebenbedingungen



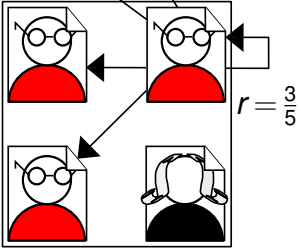
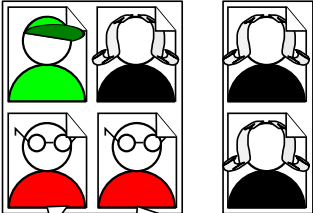
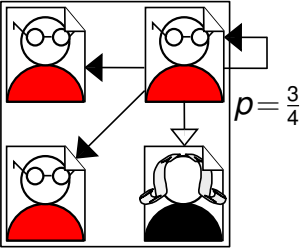
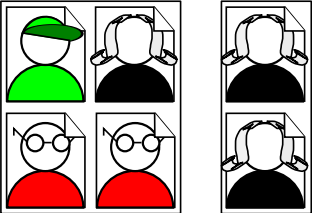
$$1 - (1 - 0.40) \cdot (1 - 0.35)$$

Addition von Nebenbedingungen



$$1 - \frac{(1 - 0.40) \cdot (1 - 0.35)}{(1 - 0.48)}$$

BCubed Precision und Recall



Einordnung in den WePS-2 Workshop

Verfahren	F-Measure			
	$\alpha=0,5$	$\alpha=0,2$	P	R
PolyUHK	0,82	0,80	0,87	0,79
<i>Thesis</i>	0,81	0,84	0,79	0,85
UVA	0,81	0,80	0,85	0,80
ITC-UT	0,81	0,76	0,93	0,73
XMEDIA	0,72	0,68	0,82	0,66
UCI	0,71	0,77	0,66	0,84