

Nebenbedingungen in der Personensuche im Internet

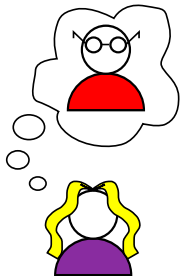
Constraints in Web People Search

Johannes Kiesel

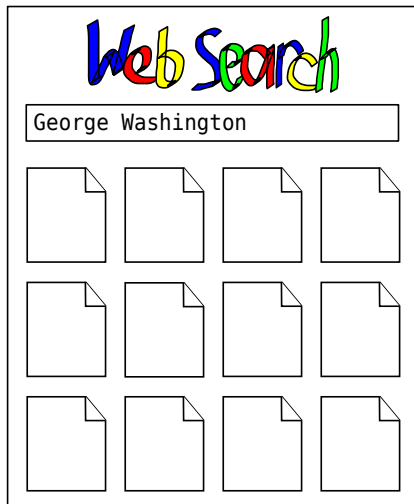
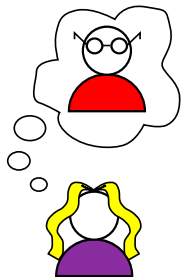
Bauhaus-Universität Weimar

3. Oktober 2012

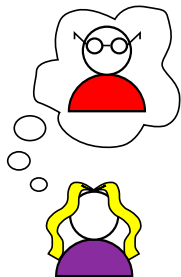
Was ist „Personensuche im Internet“?



Was ist „Personensuche im Internet“?



Was ist „Personensuche im Internet“?



Was ist „Personensuche im Internet“?



Web Search

George Washington Student

A screenshot of a web search interface. At the top, the text "Web Search" is written in a colorful, stylized font. Below it, a search bar contains the text "George Washington Student". Underneath the search bar, three search results are displayed as small icons in document-like frames. The first icon shows a stick figure with a green cap and a green top. The second icon shows a stick figure with glasses and a red top. The third icon shows a stick figure with a green cap and a green top.

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 ▼

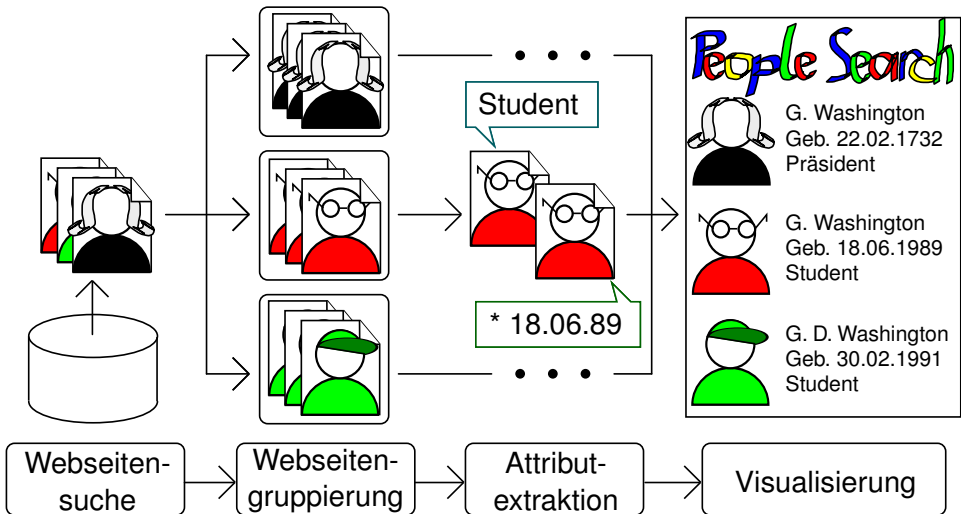
Last Update

No Limit ▼

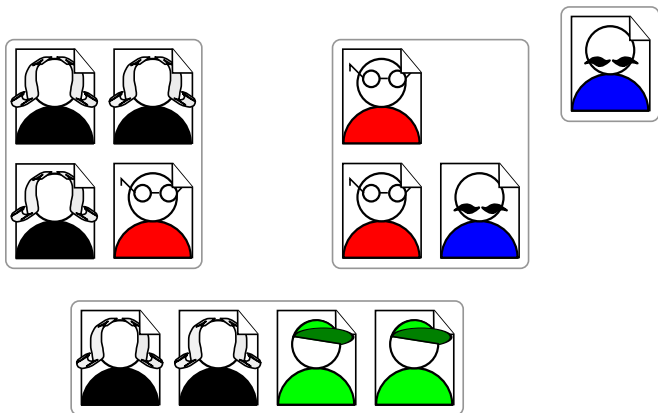
 Results 1-25

<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



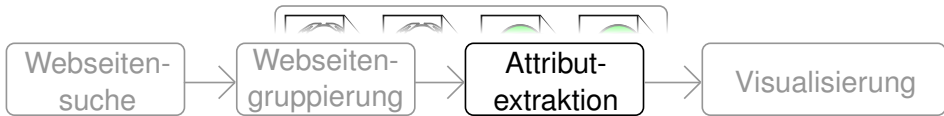
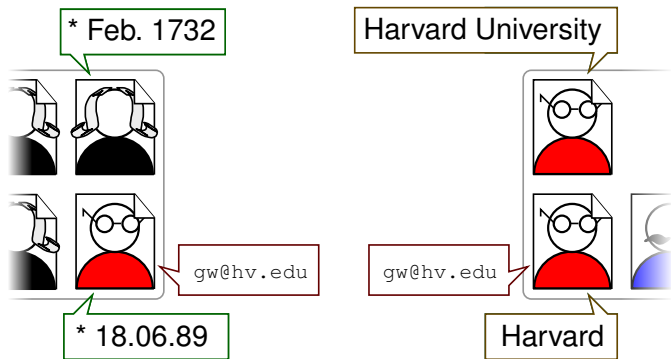
Webseiten-
suche

Webseiten-
gruppierung

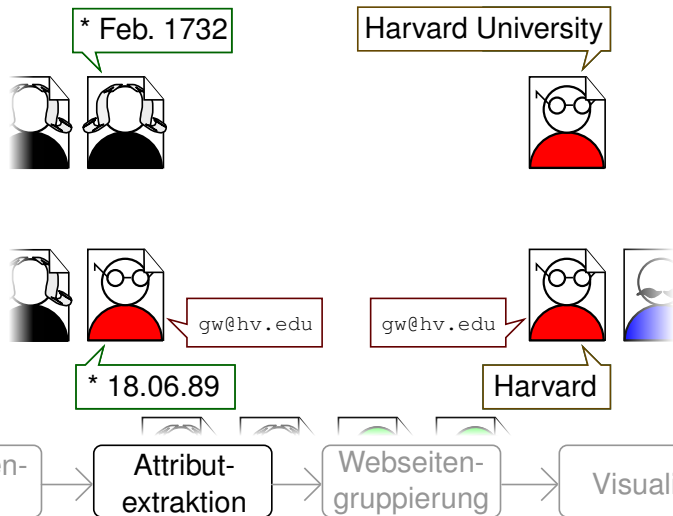
Attribut-
extraktion

Visualisierung

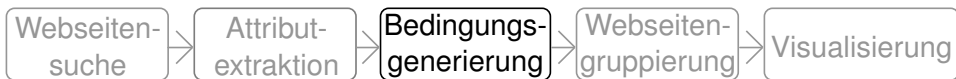
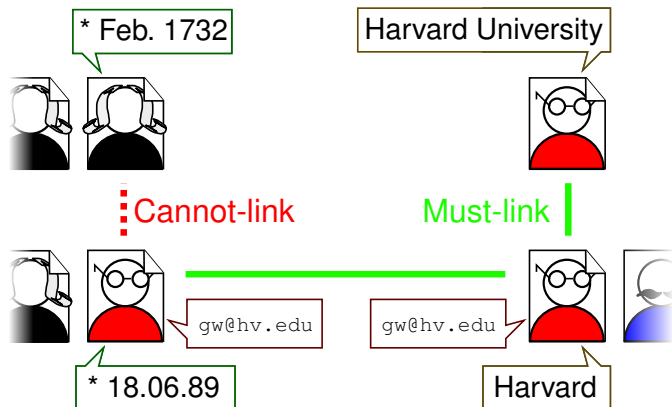
Vorgehensweise



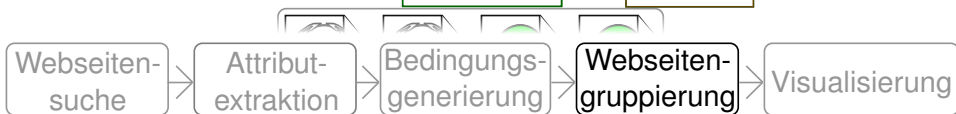
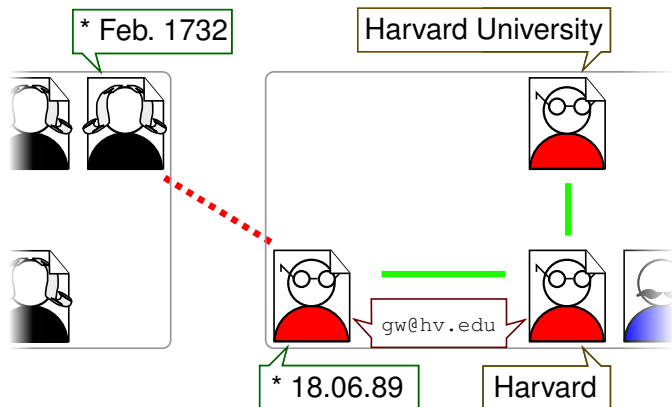
Generierung von Nebenbedingungen



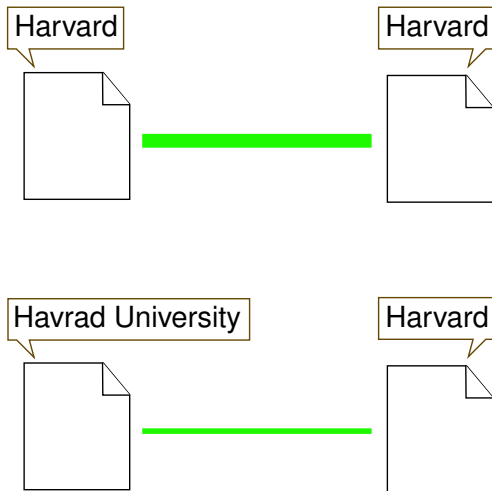
Generierung von Nebenbedingungen



Generierung von Nebenbedingungen

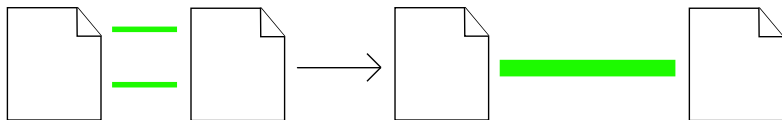


Abgleich von Attributwerten

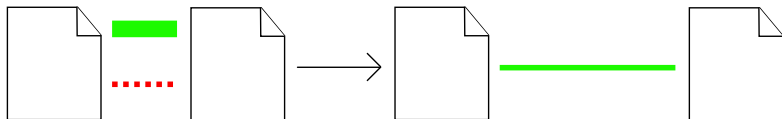


Addition der Nebenbedingungen

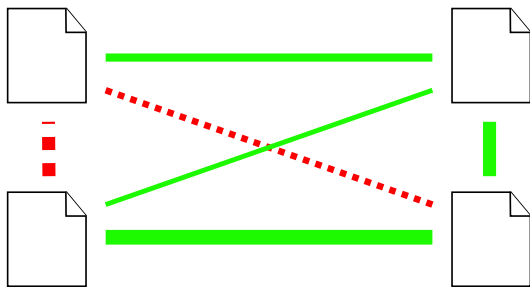
▶ Verstärkung



▶ Abschwächung



Transitivität und Konflikte



1. Anwendung eines Schwellwertes
2. Konfliktbehebung

Transitivität und Konflikte



1. Anwendung eines Schwellwertes
2. Konfliktbehebung

Transitivität und Konflikte



1. Anwendung eines Schwellwertes
2. Konfliktbehebung

Zusammenfassung: Generierung und Anwendung von Nebenbedingungen

- ▶ Abgleich von Attributwerten
- ▶ Addition von Nebenbedingungen
- ▶ Konfliktbehebung
- ▶ Gruppierung (Constrained Clustering)

Verwendeter Korpus: WePS-2

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

Teil des Korpus	Dokumente	Referenten
Englische Wikipedia	94,0	10,7
ACL'08	81,6	14,2
1990 US Zensus	80,2	30,3
Gesamter Korpus	85,3	18,4

Durchschnittswerte für eine Anfrage

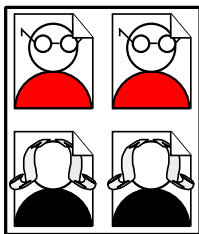
Verwendeter Korpus: WePS-2 (Personenattribute)

Attribut	Werte per Dokument	Werte per Referent	Referenten per Wert
Beruf	1,07	9,60	1,04
Zugehörigkeit	1,04	8,36	1,00
Schule	0,17	2,81	1,01
Geburtsort	0,10	2,44	1,00
Geburtstag	0,10	1,11	1,05
2. Vorname	0,09	1,02	1,07
Nationalität	0,08	1,18	1,00
E-Mail	0,07	1,03	1,01

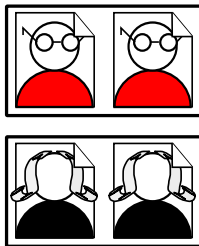
Durchschnittswerte für eine Anfrage

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

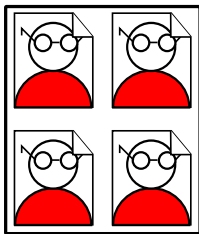
Homogenität



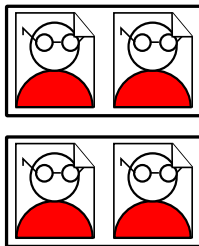
→
Höhere
Qualität



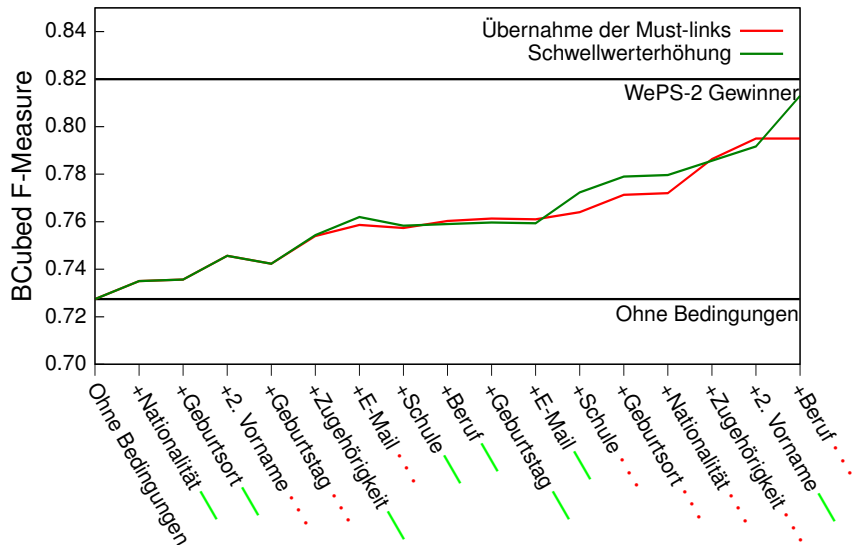
Vollständigkeit



←
Höhere
Qualität



Auswirkungen der Nebenbedingungen



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
BUW	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