Verbesserte Kaskade zur Suchsitzungs- und Missionserkennung in Suchanfrage-Logs

Jakob Gomoll

Bauhaus-Universität Weimar
jakob.gomoll@uni-weimar.de

Bachelorverteidigung
Donnerstag, 26. April 2012

Motivation

Google

die zwölf apostel

Google-Suche | Auf gut Glück!

Motivation

sehenswürdigkeiten australien

opernhaus von sydney

ayers rock

die zwölf apostel

Google-Suche | Auf gut Glück!

Motivation

sehenswürdigkeiten australien

opernhaus von sydney

ayers rock

die zwölf apostel

Google-Suche | Auf gut Glück!

Zielstellung

ID	Anfrage	Zeit	Rang	Domain
786	london top 10 attractions	2012-04-20 19:11:58		
786	buckingham palace	2012-04-20 19:12:42	1	www.visitlondon.co.uk
786	tower bridge	2012-04-20 19:14:36	5	www.londononline.co.uk
786	presidential election	2012-04-22 21:29:53		
786	presidential election france	2012-04-22 21:30:06	1	topics.nytimes.com
786	houses of parlaiment	2012-04-23 17:49:32		
786	houses of parliament	2012-04-23 17:49:34	1	www.aviewoncities.com
786	big ben	2012-04-23 17:53:13	3	www.gothereguide.com

Zielstellung

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13

Zielstellung

ID	Anfrage	Zeit	Missions-ID
786	london top 10 attractions	2012-04-20 19:11:58	1
786	buckingham palace	2012-04-20 19:12:42	1
786	tower bridge	2012-04-20 19:14:36	1
786	presidential election	2012-04-22 21:29:53	2
786	presidential election france	2012-04-22 21:30:06	2
786	houses of parlaiment	2012-04-23 17:49:32	1
786	houses of parliament	2012-04-23 17:49:34	1
786	big ben	2012-04-23 17:53:13	1

Zielstellung

ID	Anfrage	Zeit	Missions-ID
786	london top 10 attractions	2012-04-20 19:11:58	1
786	buckingham palace	2012-04-20 19:12:42	1
786	tower bridge	2012-04-20 19:14:36	1
786	presidential election	2012-04-22 21:29:53	2
786	presidential election france	2012-04-22 21:30:06	2
786	houses of parlaiment	2012-04-23 17:49:32	1
786	houses of parliament	2012-04-23 17:49:34	1
786	big ben	2012-04-23 17:53:13	1

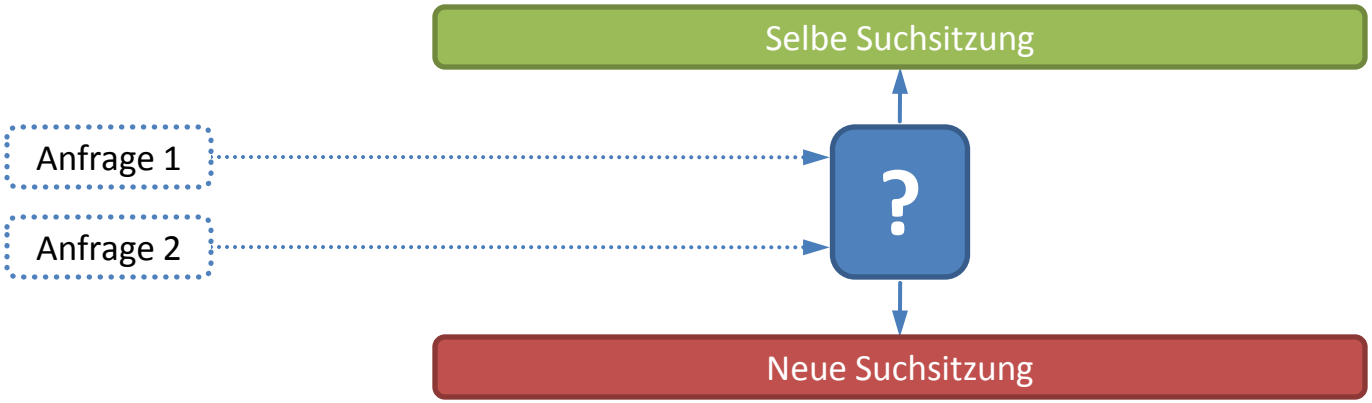
1. Erstellung eines neuen Gold-Standards
2. Automatische Einteilung des Anfrage-Logs in Sitzungen
3. Automatische Zuordnung der einzelnen Sitzungen zu Missionen

Überarbeiteter Gold-Standard

- Gayo-Avello, 2009
  - 11 500 Suchanfragen
  - 215 Nutzer
  - Viele Mängel
  - Keine Missionen
- Lucchese et al., 2011
  - 1 500 Suchanfragen
  - 13 Nutzer
  - Zu klein
  - 97 % aller Anfragen verworfen
- Neuer Gold-Standard
  - 8 840 Suchanfragen
  - 127 Nutzer
  - Angemessene Größe
  - Bessere Qualität

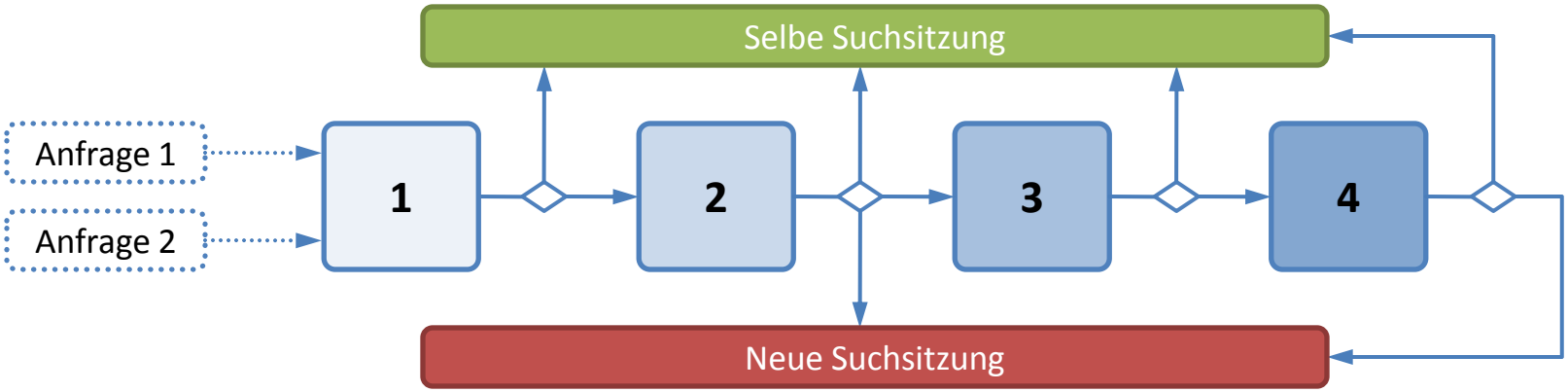
Anordnung in einer Kaskade

- Matthias Hagen, Benno Stein und Tino Rüb. *Query Session Detection as a Cascade*. CIKM, 2011.



Anordnung in einer Kaskade

- Matthias Hagen, Benno Stein und Tino Rüb. *Query Session Detection as a Cascade*. CIKM, 2011.
  1. Termüberlappung
  2. Geometrische Methode
  3. Explizite Semantische Analyse
  4. Anfrage-Erweiterung



Schritt 1

- Wiederholung

```
presidential election
presidential election
```

Generalisierung

```
presidential election france
presidential election
```

Spezialisierung

```
presidential election
presidential election france
```

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13

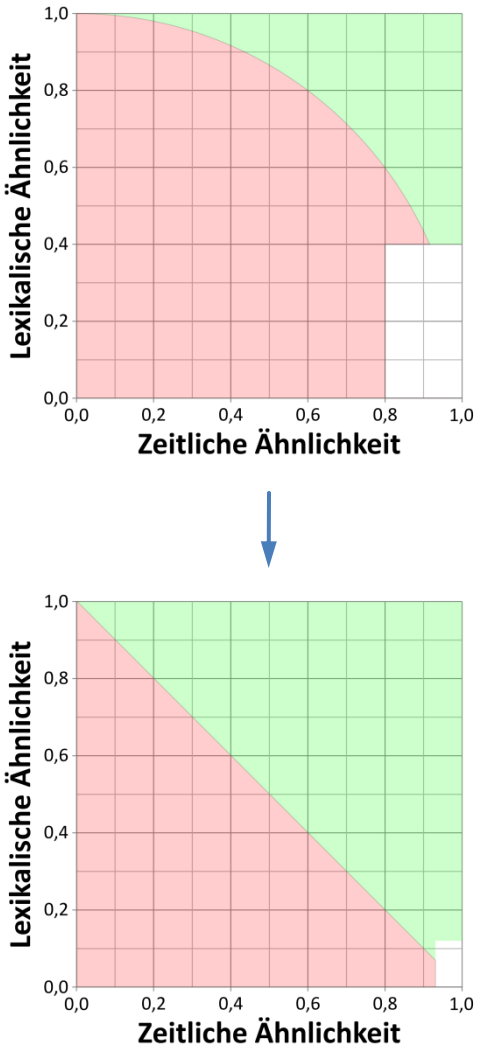
- Termüberlappung → Stringvergleich

Schritt 2

- Geometrische Methode

  [Gayo-Avello, 2009]

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13



Schritt 3

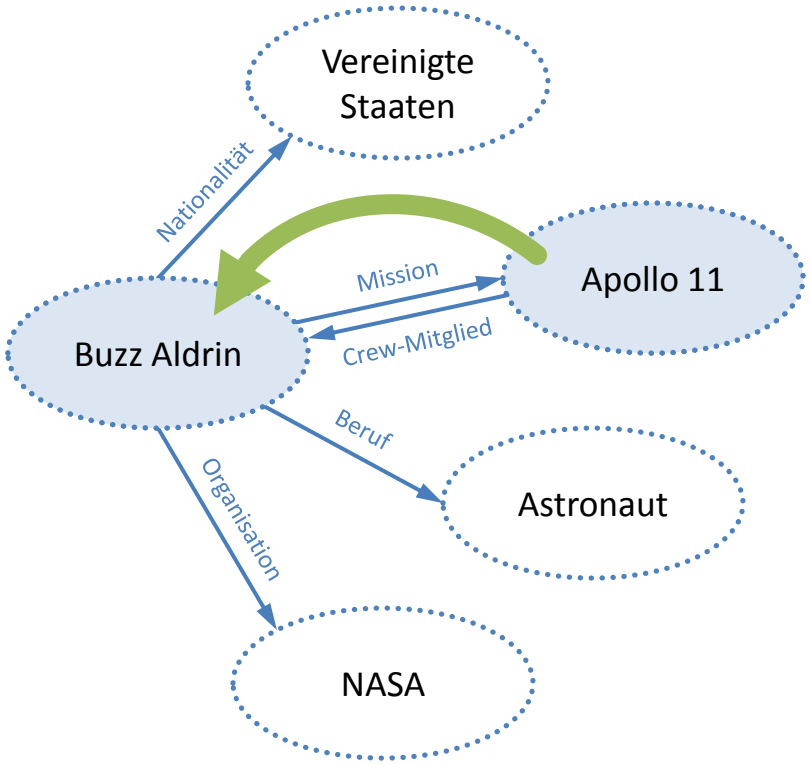
- Explizite Semantische Analyse (ESA)

  [Gabrilovich und Markovitch, 2007]

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13

Idee: Neuer Schritt

- Linked Open Data (LOD)
  - DBpedia: 3,6 Millionen Entitäten

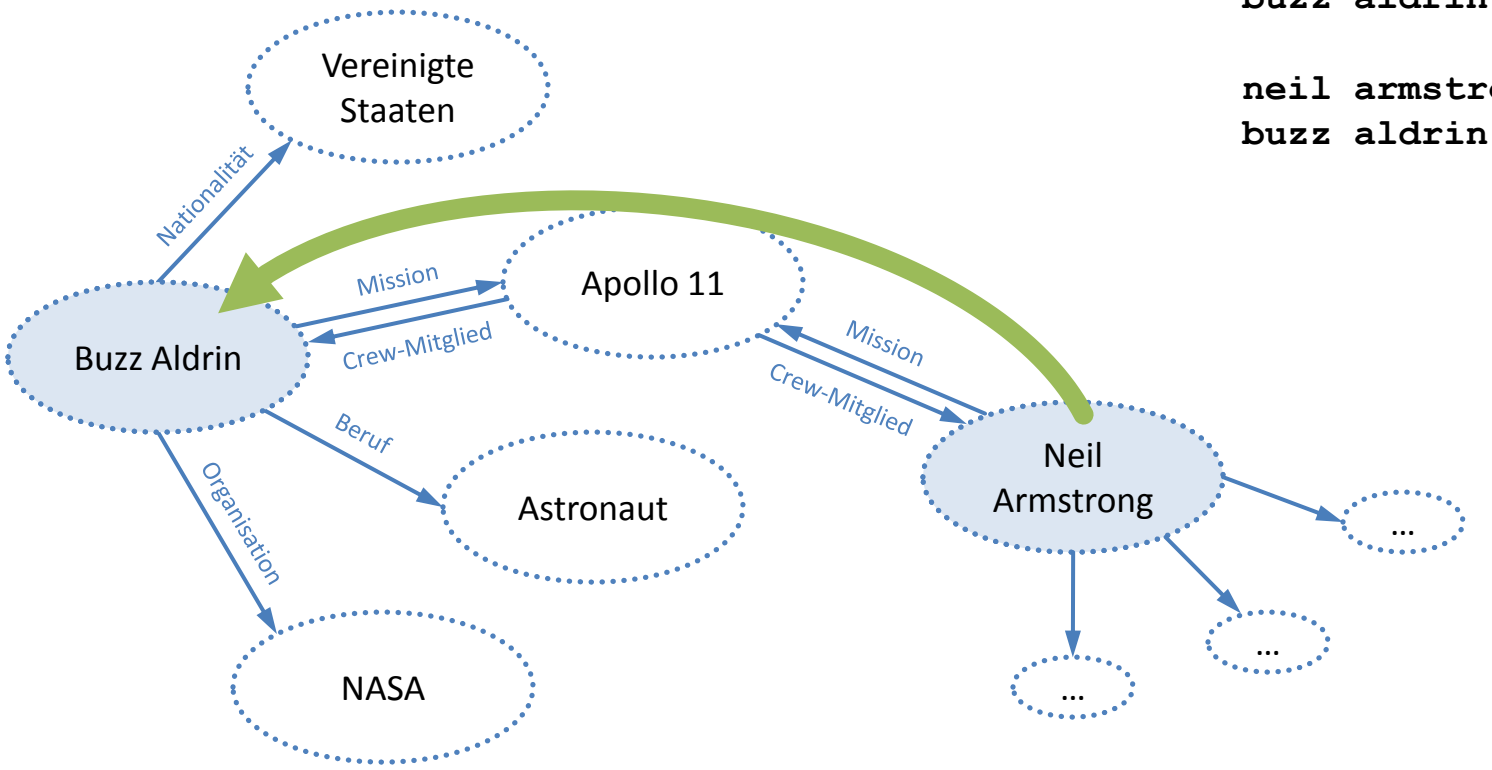



```
apollo 11
buzz aldrin
```

✓

Idee: Neuer Schritt

- Linked Open Data (LOD)
  - DBpedia: 3,6 Millionen Entitäten




```
apollo 11
buzz aldrin
```

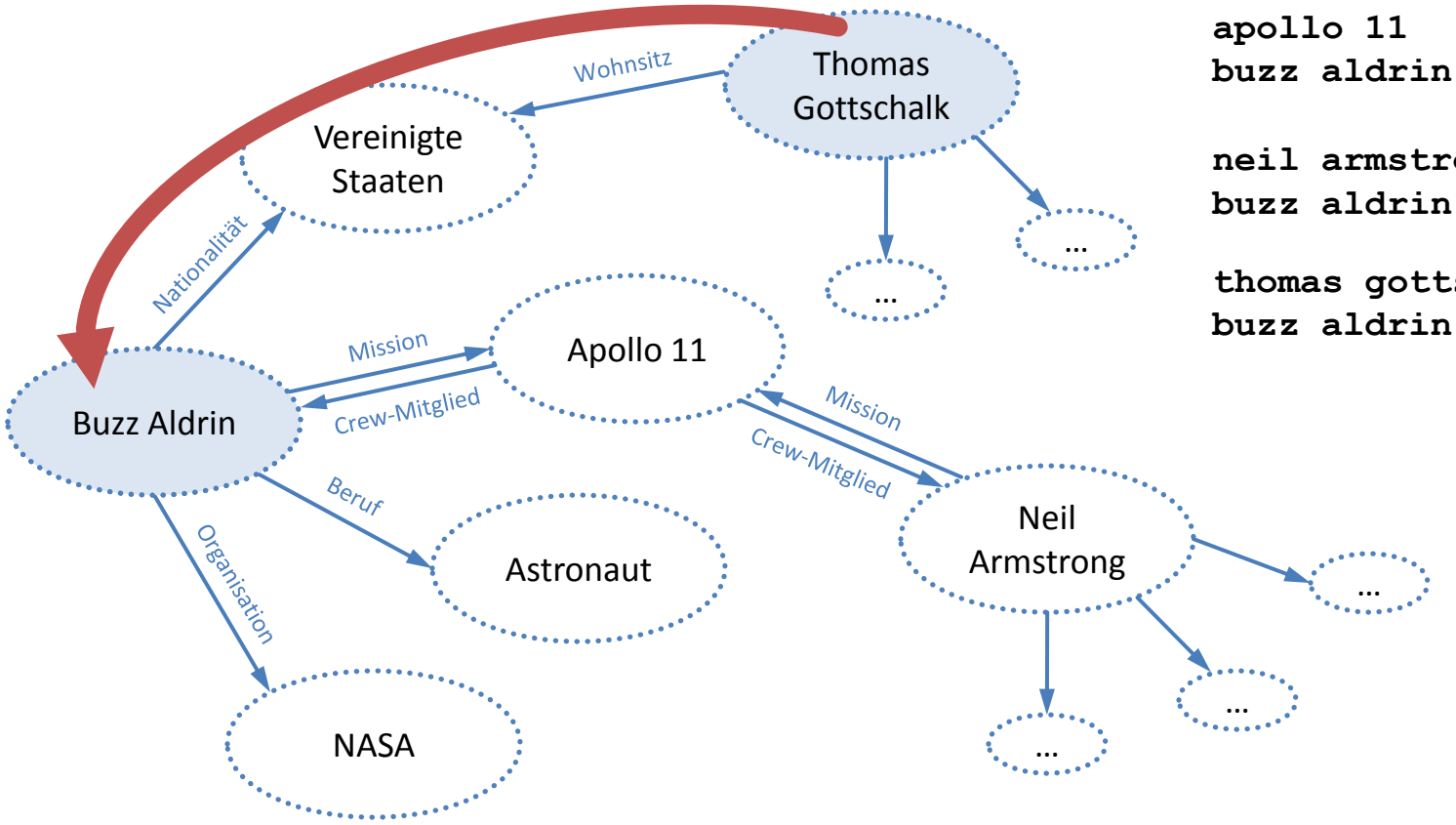
✓

```
neil armstrong
buzz aldrin
```

✓

Idee: Neuer Schritt

- Linked Open Data (LOD)
  - DBpedia: 3,6 Millionen Entitäten




```
apollo 11
buzz aldrin
```
✓

```
neil armstrong
buzz aldrin
```
✓

```
thomas gottschalk
buzz aldrin
```
✗

Schritt 4

- Linked Open Data (LOD)

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13

Schritt 5

- Anfrage-Erweiterung

Top 10 London Attractions - Visit **London**
www.visitlondon.com › ... › Attractions in London
Read about **London's top 10** most popular tourist **attractions**, including the **London** Eye, the National Gallery and Tower of **London**. You can also book **attraction** ...
↳ Accessibility - The British Museum - EDF Energy London Eye - Tate Modern

Buckingham Palace - Places To Go in London - Visit London
www.visitlondon.com › ... › Tourist Attraction › Historic Site/House
Buckingham Palace serves as both the office and London residence of Her...
Mon, Jun 4 Diamond Jubilee Concert at Buckingham Palace
Tue, Jun 5 Diamond Jubilee Carriage Procession at Buckingham Palace
Jun 30 - Oct 7 Buckingham Palace Summer Opening 2012

ID	Anfrage	Zeit
786	london top 10 attractions	2012-04-20 19:11:58
786	buckingham palace	2012-04-20 19:12:42
786	tower bridge	2012-04-20 19:14:36
786	presidential election	2012-04-22 21:29:53
786	presidential election france	2012-04-22 21:30:06
786	houses of parlaiment	2012-04-23 17:49:32
786	houses of parliament	2012-04-23 17:49:34
786	big ben	2012-04-23 17:53:13

- Top-10 Suchergebnisse → Kosinusähnlichkeit der Snippets

Ergebnisse

- Precision & Recall → F-Measure

	Hagen et al.		**Verbesserte Kaskade**	
	F-Measure	Zeit je Paar	F-Measure	Zeit je Paar
Lexikalische Ähnlichkeit	0,809	0,0113 ms	0,800	0,0009 ms
Geometrische Methode	0,871	0,19 ms	0,889	0,16 ms
Explizite Semantische Analyse	0,877	0,20 ms	0,891	0,20 ms
Linked Open Data	-	-	0,887	90,12 ms
Anfrage-Erweiterung	0,881	536,04 ms	0,894	541,40 ms

Missionserkennung

ID	Anfrage	Zeit	Missions-ID
786	london top 10 attractions	2012-04-20 19:11:58	1
786	buckingham palace	2012-04-20 19:12:42	1
786	tower bridge	2012-04-20 19:14:36	1
786	presidential election	2012-04-22 21:29:53	2
786	presidential election france	2012-04-22 21:30:06	2
786	houses of parlaiment	2012-04-23 17:49:32	1
786	houses of parliament	2012-04-23 17:49:34	1
786	big ben	2012-04-23 17:53:13	1

	Precision	Recall	F-Measure
Stringvergleich	0,712	0,460	0,641
Geometrische Methode	0,718	0,546	0,676
Explizite Semantische Analyse	0,715	0,546	0,673
Anfrage-Erweiterung	0,727	0,579	0,692
Baseline	0,075	0,057	0,070

Zusammenfassung & Ausblick

- Zusammenfassung
  - Neuer Gold-Standard zur Missionserkennung
  - Verbesserung der bestehenden Kaskade
  - Linked Open Data noch nicht ausgereift
  - Verfahren zur Missionserkennung

- Ausblick
  - LOD-Schritt verbessern
  - Semantischer Zusammenhang über Wikipedia
  - Untersuchung des Missionsverhaltens von Nutzern
  - Einbindung in das ChatNoir-Projekt

Ähnlichkeitsmerkmale

Zeit		5 Minuten	[Silverstein et al., 1999]
		30 Minuten	[Radlinski und Joachims, 2005]
		Variabel	[Murray et al., 2006]
	Aber:	Maximale Genauigkeit: 70 %	[Jones und Klinkner, 2008]
Syntax		Termüberlappung	[Jansen et al., 2007]
		Jaccard-Koeffizient	[Zhang und Moffat, 2006]
		Levenshtein-Distanz	[Sun et al., 2010]
Zeit + Syntax		< 60 min + Termüberlappung	[Seco und Cardoso, 2006]
		Geometrische Methode	[Gayo-Avello, 2009]
Semantik		Explizite Semantische Analyse	[Gabrilovich und Markovitch, 2007]
		Linked Open Data	[Hollink et al., 2011]
		Anfrage-Erweiterung	[Shen et al., 2005]

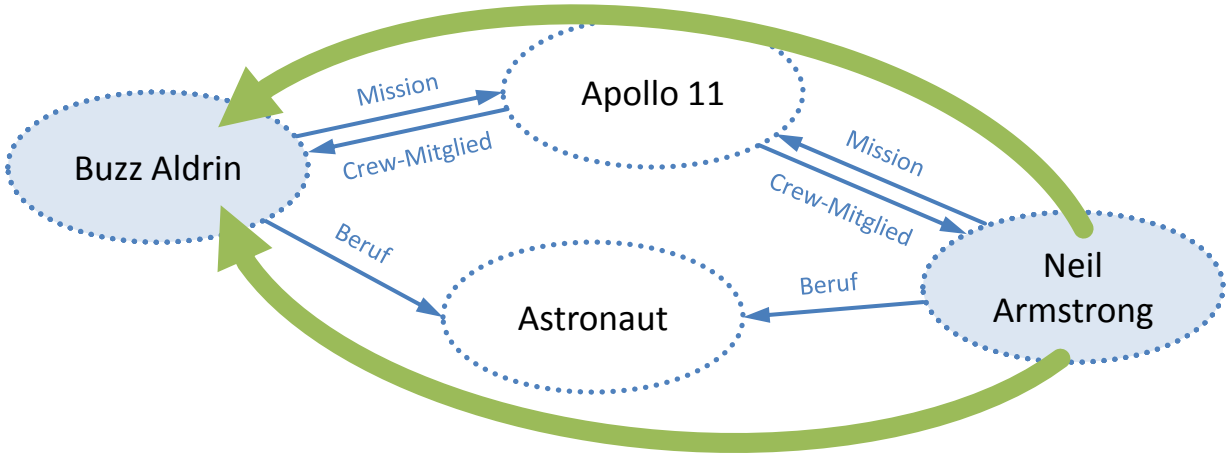
Potenzial des LOD-Schrittes

- 100 zusammenhängende Dinge
  - Steve Ballmer und Bill Gates
  - Volkswagen und Martin Winterkorn
- 50 eher nicht zusammenhängende Dinge
  - Ford Taurus und Taurusgebirge
  - Great Barrier Reef und Schweden

		ESA	LOD
Zusammenhang vorhanden	✓	14	52
	✗	86	48
Zusammenhang nicht vorhanden	✓	47	45
	✗	3	5

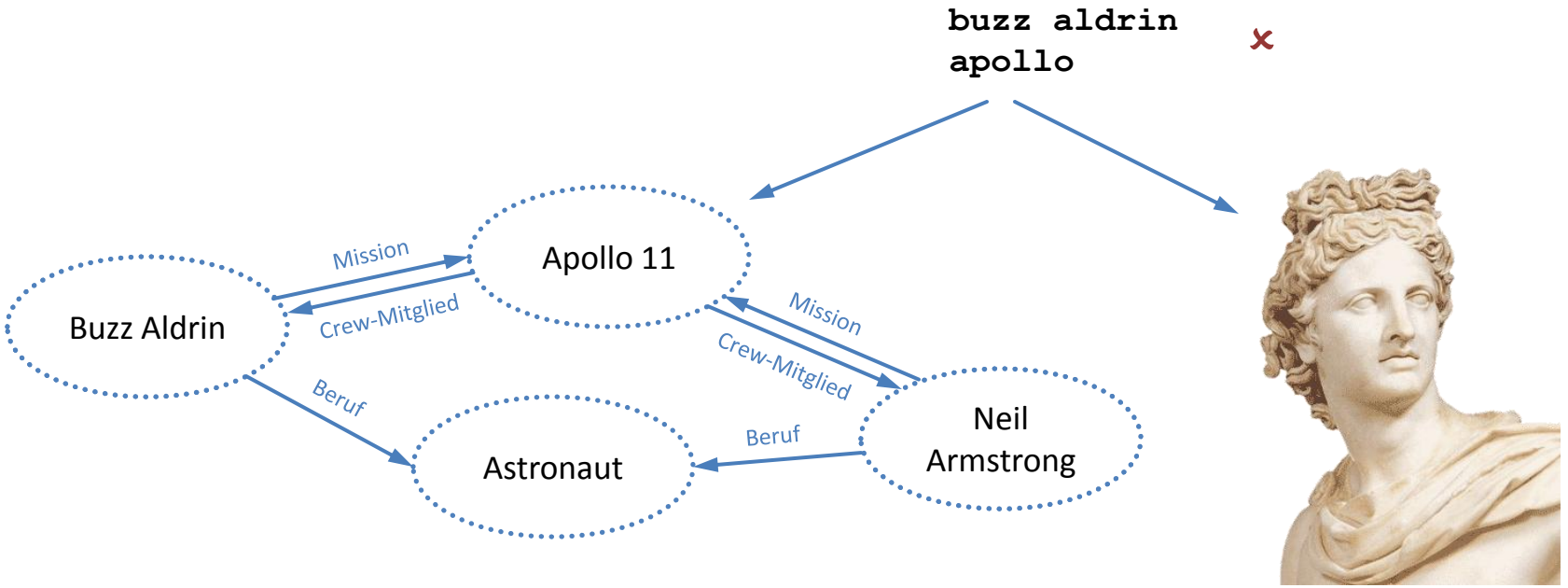
Schwächen des LOD-Schrittes

- Gesamtgewicht über mehrere Pfade?
  - Siebformel nach Poincaré und Sylvester



Schwächen des LOD-Schrittes

- Gesamtgewicht über mehrere Pfade?
  - Siebformel nach Poincaré und Sylvester
- Abbilden von Anfragen auf Entitäten



Gütemaße

N_{shift}	Anzahl der vom Verfahren gesetzten Sitzungsgrenzen
N_{true_shift}	Anzahl der Sitzungsgrenzen im Gold-Standard
$N_{shift\&correct}$	Anzahl übereinstimmender Sitzungsgrenzen

$$precision = \frac{N_{shift\&correct}}{N_{shift}}$$

$$recall = \frac{N_{shift\&correct}}{N_{true\_shift}}$$

$$F - Measure_{\beta} = \frac{(1 + \beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Zusammenfassung & Ausblick

- Zusammenfassung
  - Neuer Gold-Standard zur Missionserkennung
  - Verbesserung der bestehenden Kaskade
  - Linked Open Data noch nicht ausgereift
  - Verfahren zur Missionserkennung

- Ausblick
  - LOD-Schritt verbessern
  - Semantischer Zusammenhang über Wikipedia
  - Untersuchung des Missionsverhaltens von Nutzern
  - Einbindung in das ChatNoir-Projekt