Kaskadierendes Verfahren zur Detektierung von Suchsitzungen in Anfrage-Log-Dateien

Bachelorarbeit

Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Tino Rüb, Matrikel-Nr. 21059
tino.rueb@uni-weimar.de
Betreut durch: Prof. Dr. Benno Stein

Einführung

Suchen

„Die Suche über eine computerbasierte Schnittstelle ist ein **iterativer Prozess**, der sich über **mehrere Suchanfragen** erstrecken kann. Dieser Prozess folgt dem klassischen Prinzip **Versuch und Irrtum**.“

[01] Swanson, 1977.

Einführung

Suchen

„Die Suche über eine computerbasierte Schnittstelle ist ein **iterativer Prozess**, der sich über **mehrere Suchanfragen** erstrecken kann. Dieser Prozess folgt dem klassischen Prinzip **Versuch und Irrtum**.“

[01] Swanson, 1977.

Veränderungsmuster

zweier aufeinanderfolgender Anfragen:

- Wiederholung
- Generalisierung
- Spezialisierung
- Umformulierung
- Neu

[02] Lau und Horvitz, 1999.

Einführung

Definition einer Suchsitzung

„Eine oder mehrere aufeinanderfolgende Suchanfragen, die in einem **relativ kurzen Zeitraum** erfolgen, und die einem **spezifischen Informationsbedarf** des Benutzers entsprechen.“

[03] Silverstein et al., 1999.

Einführung

Definition einer Suchsitzung

„Eine oder mehrere aufeinanderfolgende Suchanfragen, die in einem **relativ kurzen Zeitraum** erfolgen, und die einem **spezifischen Informationsbedarf** des Benutzers entsprechen.“

[03] Silverstein et al., 1999.

AnonID	Query	ClickURL	ItemRank	QueryTime
...				
4253773	istanbul archeology			2006-05-07 12:02:54
4253773	istanbul archeology	http://www.kulturturizm.gov.tr	19	2006-05-07 12:03:35
4253773	istanbul archeology			2006-05-07 18:24:07
4253773	constantinople			2006-05-07 18:24:40
4253773	constantinople	http://www.roman-empire.net	10	2006-05-07 18:24:40
4253773	acid - base foods	http://www.basica.de	5	2006-05-08 03:33:12
4253773	imodium	http://www.drugs.com	4	2006-05-08 12:42:48
4253773	imodium	http://www.drugdigest.org	6	2006-05-08 12:42:48
...				

Einführung

Definition einer Suchsitzung

„Eine oder mehrere aufeinanderfolgende Suchanfragen, die in einem **relativ kurzen Zeitraum** erfolgen, und die einem **spezifischen Informationsbedarf** des Benutzers entsprechen.“

[03] Silverstein et al., 1999.

Verfahren der Literatur

einfache Merkmale

- zeitlicher Abstand
  - systemorientiert, benutzerorientiert
- lexikalische Ähnlichkeit
  - Levensthein-Distanz, n-m-Gramm-Überlappung

Verfahren der Literatur

einfache Merkmale

- zeitlicher Abstand
  - systemorientiert, benutzerorientiert
- lexikalische Ähnlichkeit
  - Levensthein-Distanz, n-m-Gramm-Überlappung

erweiterte Merkmale

- explizite semantische Analyse [ESA]
- Quantifizierung der Ergebnisse einer Suchmaschine [SR]

[04] Lucchese et al., 2010.

[05] Metzler et al., 2007.

Verfahren der Literatur

einfache Merkmale

- zeitlicher Abstand
  - systemorientiert, benutzerorientiert
- lexikalische Ähnlichkeit
  - Levensthein-Distanz, n-m-Gramm-Überlappung

erweiterte Merkmale

- explizite semantische Analyse [ESA] [04] Lucchese et al., 2010.
- Quantifizierung der Ergebnisse einer Suchmaschine [SR] [05] Metzler et al., 2007.

Kombination mehrerer Merkmale

- heuristisch
  - geometrische Methode [06] Gayo-Avello, 2009.
- maschinelle Lernverfahren
  - Dempster-Shaver, Neuronales Netz, Multiplen Linearen Regression...
  - Logistische Regression [07] Jones und Klinkner, 2008.

Motivation und Zielsetzung

Motivation

- höherer Informationsgehalt als einzelne Anfrage
- Intention des Benutzer lässt sich besser bestimmen

Motivation und Zielsetzung

Motivation

- höherer Informationsgehalt als einzelne Anfrage
- Intention des Benutzer lässt sich besser bestimmen

Zielsetzung

- bessere Aussagesicherheit
- performantes Verfahren

Motivation und Zielsetzung

Motivation

- höherer Informationsgehalt als einzelne Anfrage
- Intention des Benutzer lässt sich besser bestimmen

Zielsetzung

- bessere Aussagesicherheit
- performantes Verfahren

Ideen

- Ausgangspunkt: die verschiedene Merkmale benötigen unterschiedliche hohe „Kosten“ für ihre jeweilige Quantifizierung
- 1) nur wenn „günstige“ Merkmale keine sichere Aussage zulassen
  ⇒ schrittweise „teurere“ Merkmale quantifizieren
- 2) ist die Quantifizierung aller Anfragen im Falle eines sub-samples zwingend notwendig ?

Datenaufbereitung

Irrelevante Inhalte

- robots und softwareagents
- Meta-Suchmaschinen
- uninteressant bzgl. der Fragestellung

Datenaufbereitung

Irrelevante Inhalte

- robots und softwareagents
- Meta-Suchmaschinen
- uninteressant bzgl. der Fragestellung

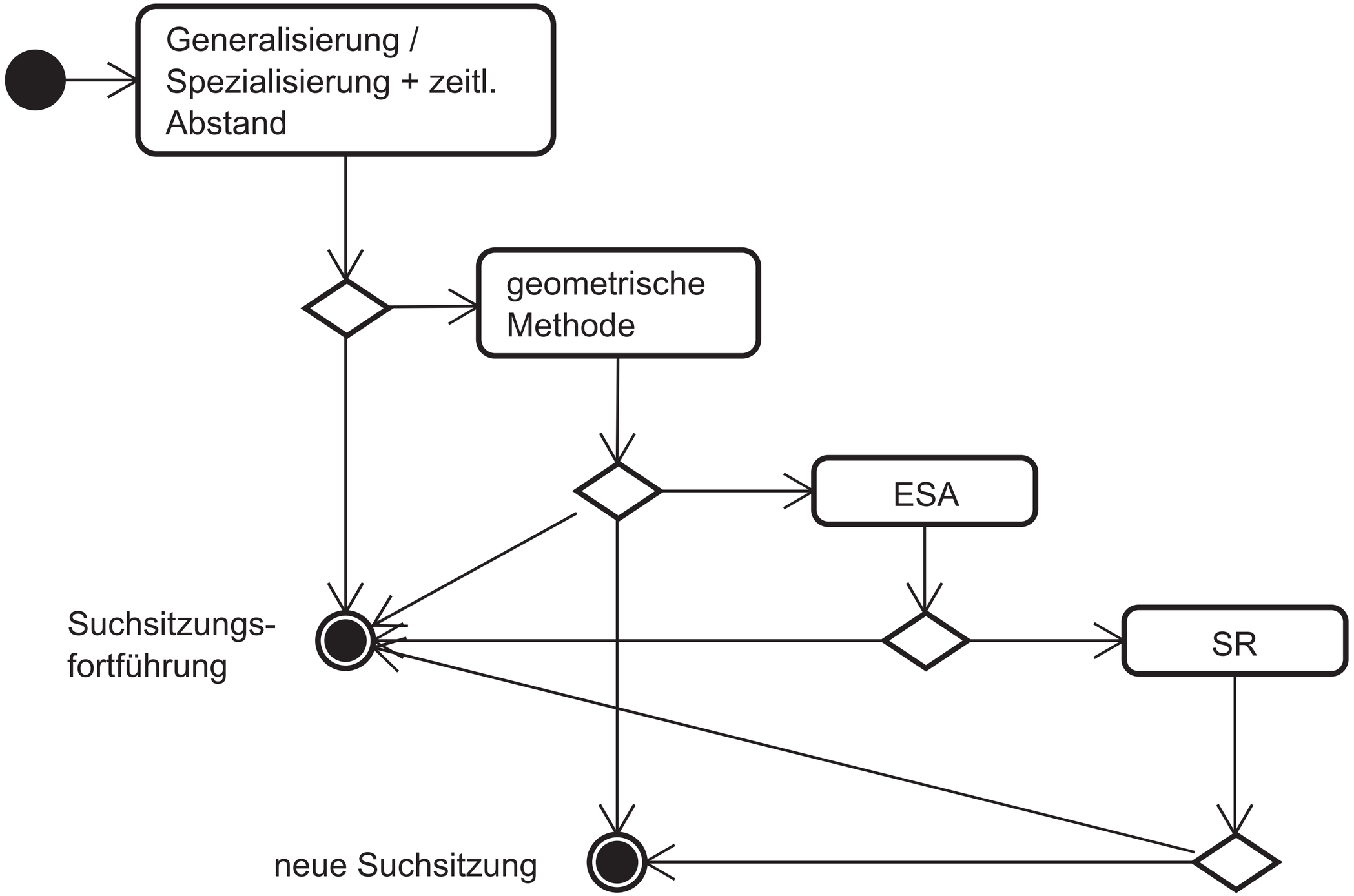
Angewendete Filter-Regeln

	Beschreibung	Aktionen	Benutzer
1.	nur eine Aktion	0,16 %	9,49 %
2.	im Durchschnitt weniger als 10 Sekunden zwischen je 2 Aktionen	2,48 %	12,17 %
3.	Median der Anfragenlänge größer als 100	< 0,01 %	0,04 %
	Gesamt	2,64 %	21,70 %

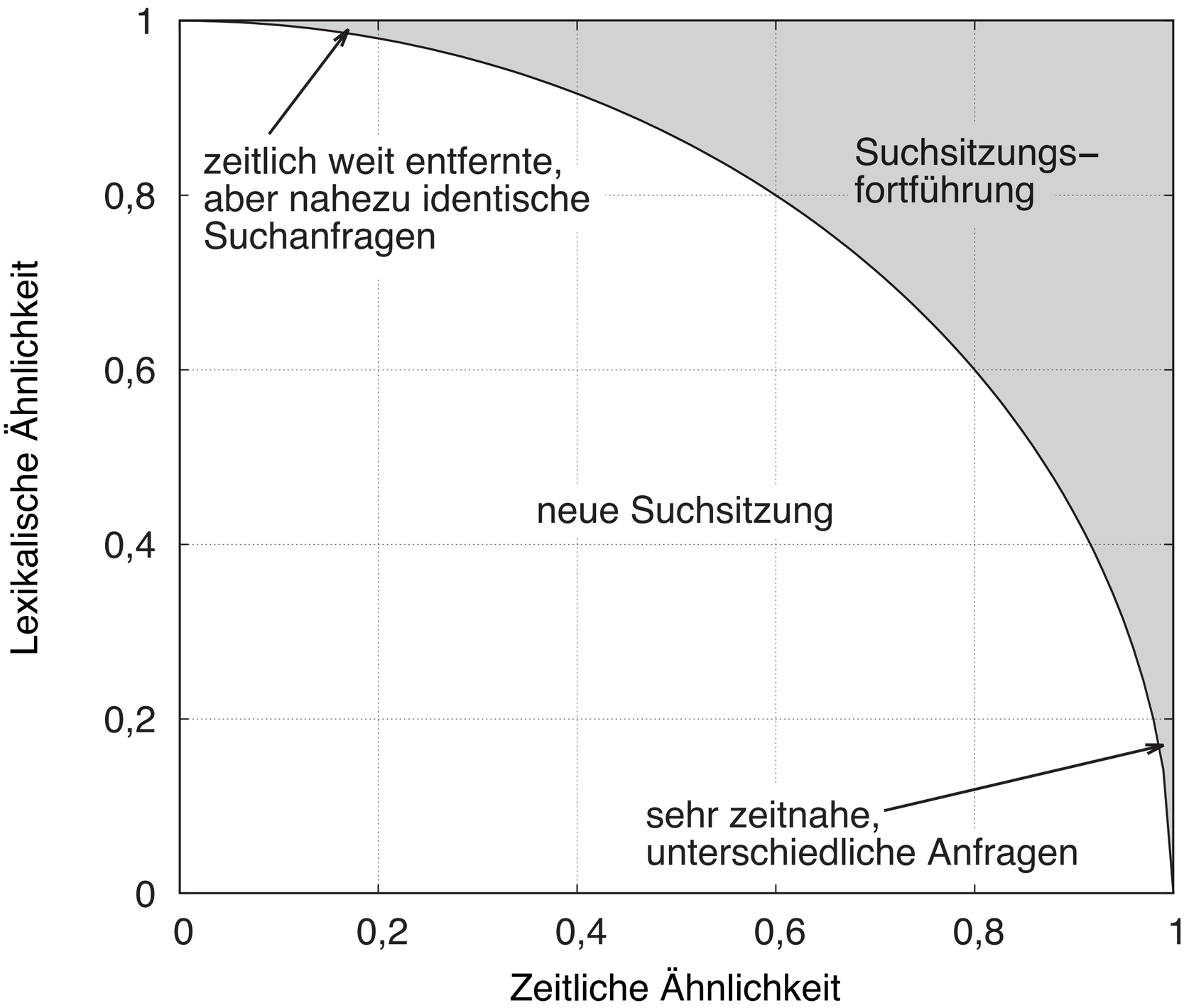
Entfernt werden:

- 959.641 Aktionen von insgesamt 36.389.567
- 130.292 Benutzer von insgesamt 600.477

Kaskadierender Ansatz

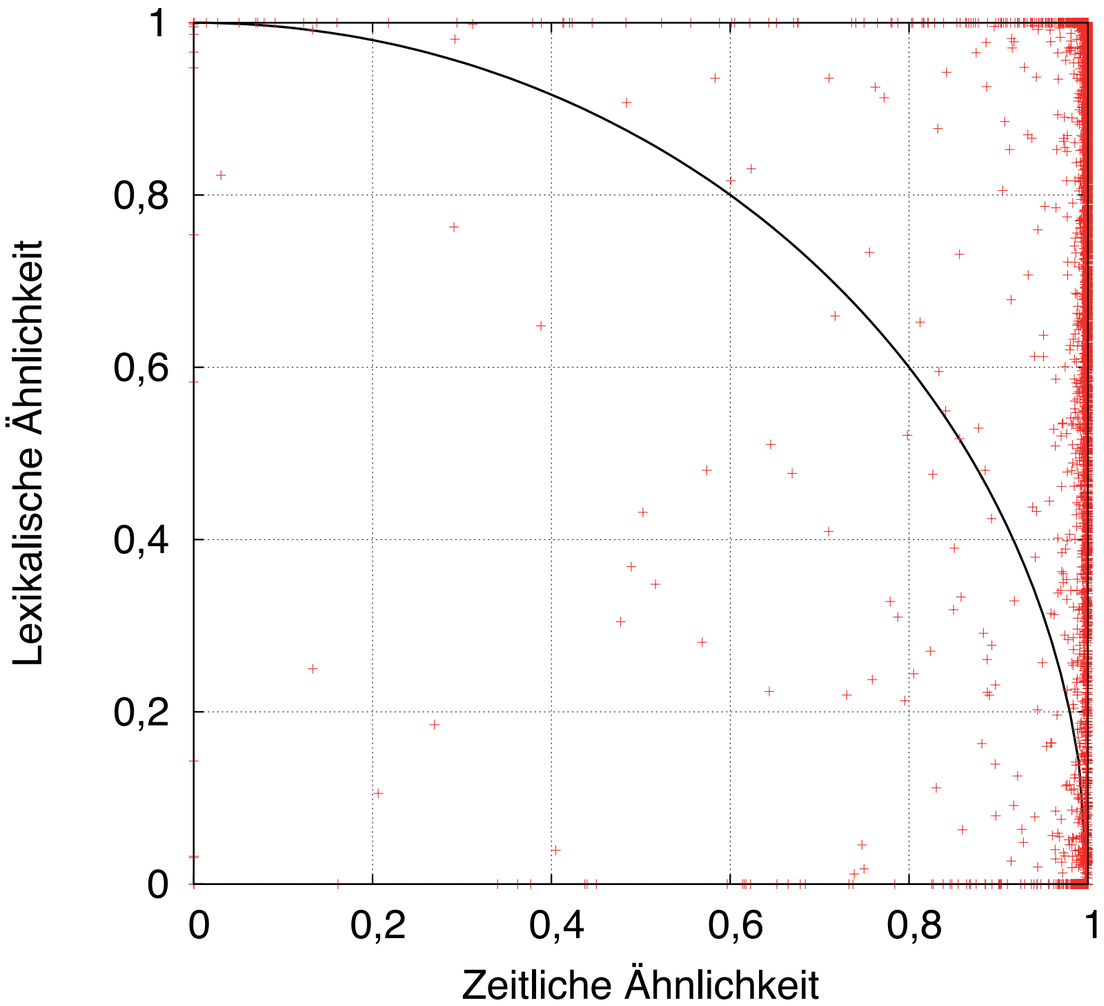


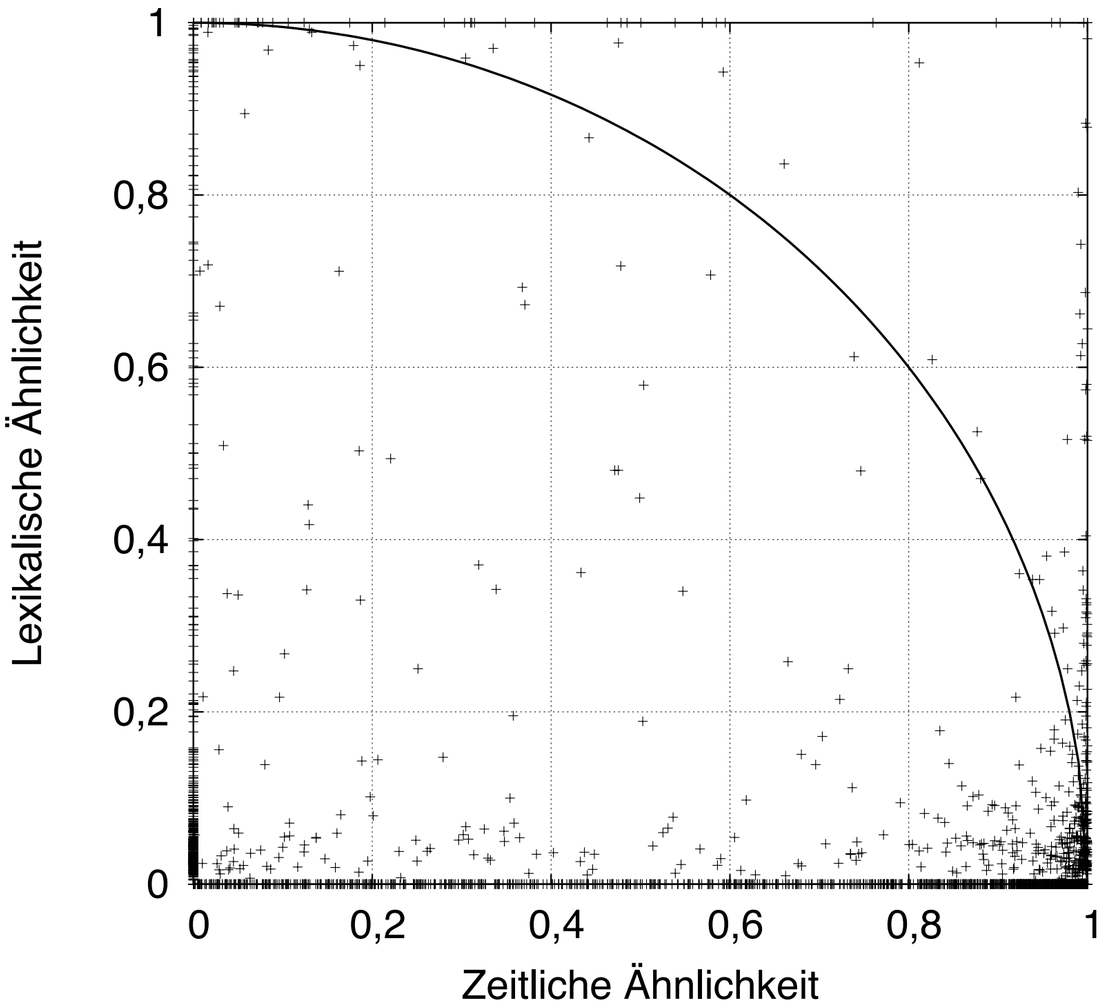
Geometrische Methode



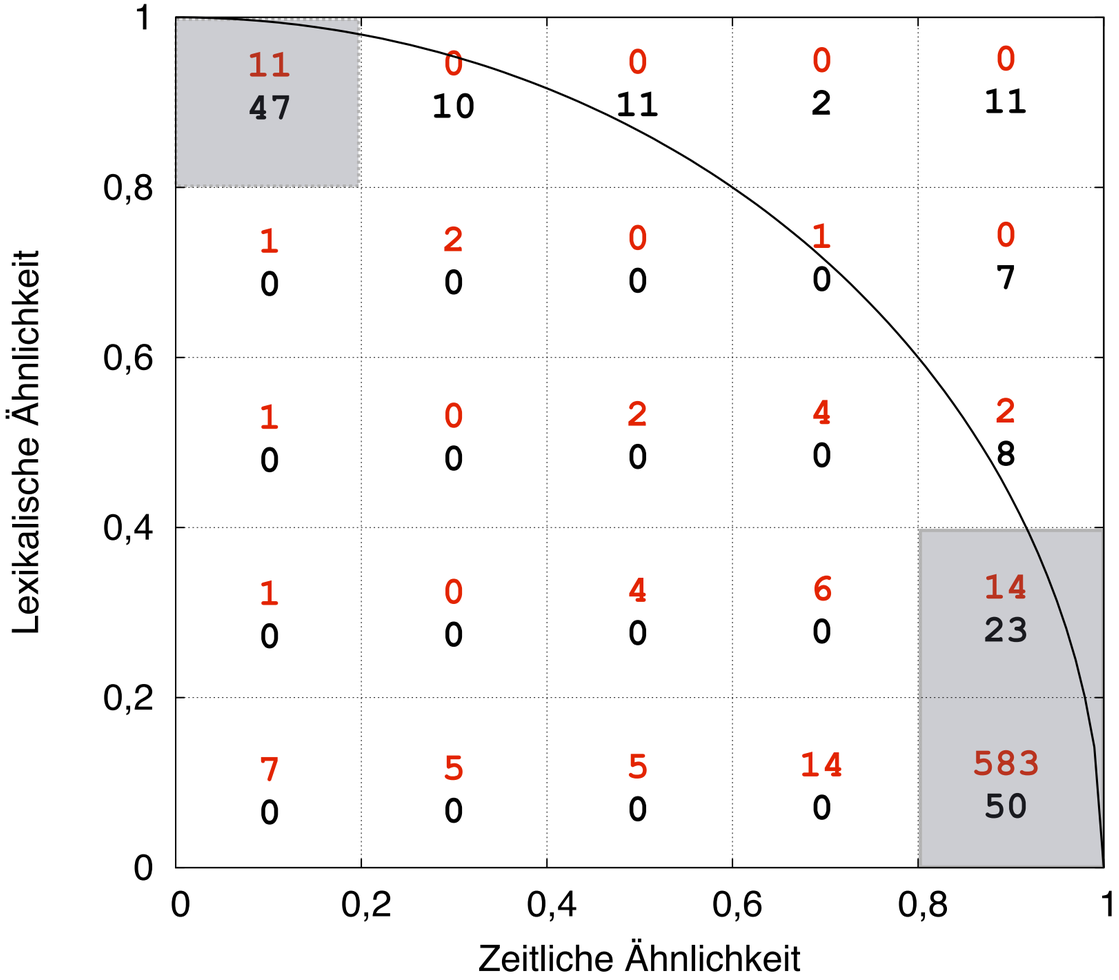
nach [06] Gayo-Avello, 2009.

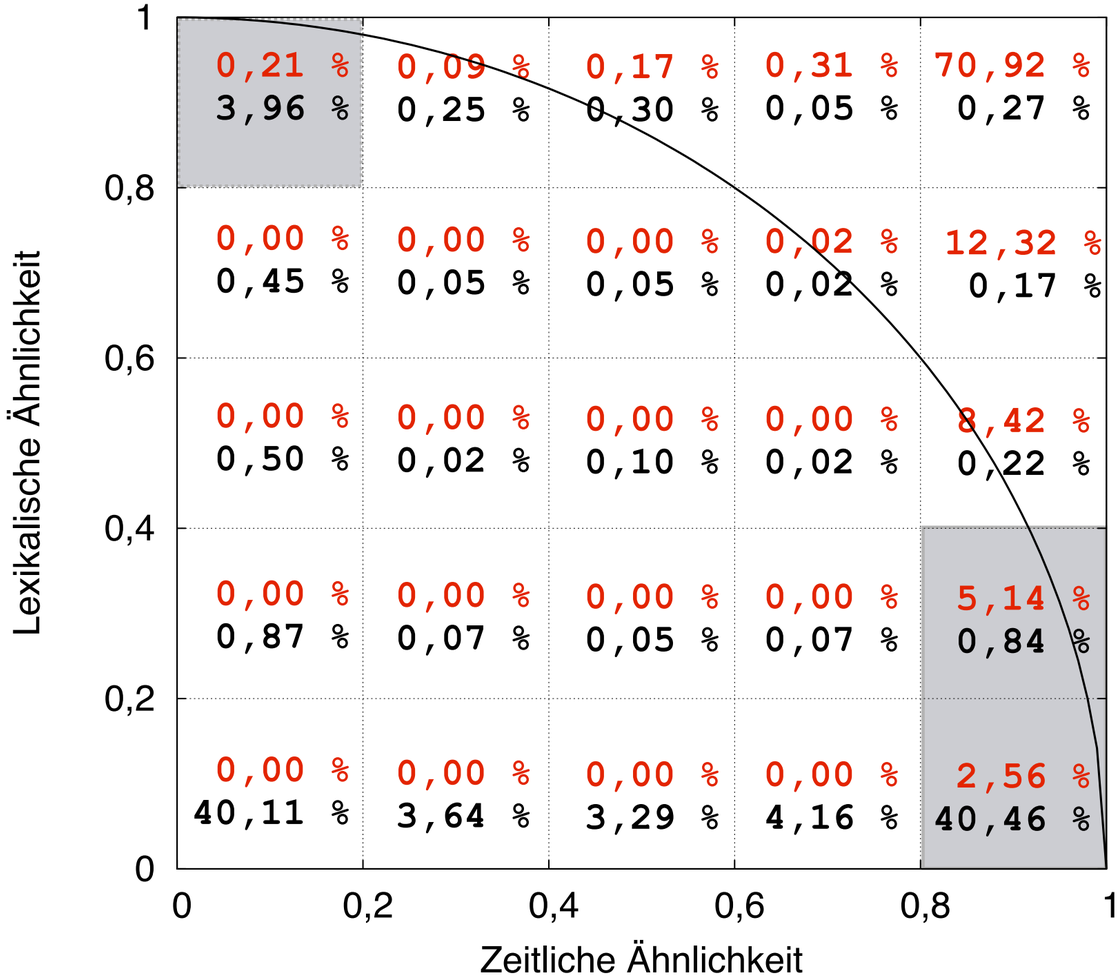
Verteilung der Geometrische Methode



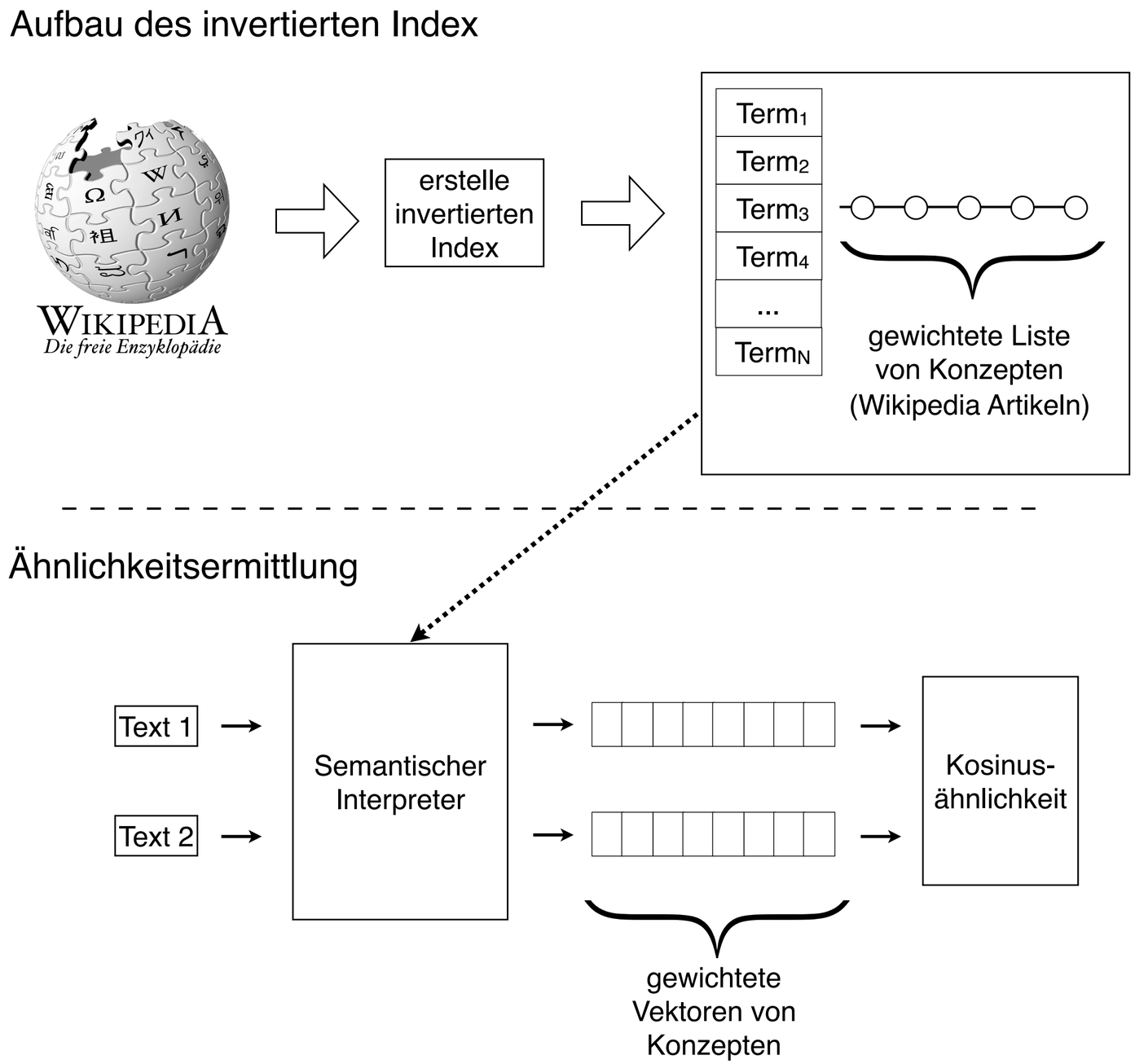


Detailbetrachtung der geometrischen Methode

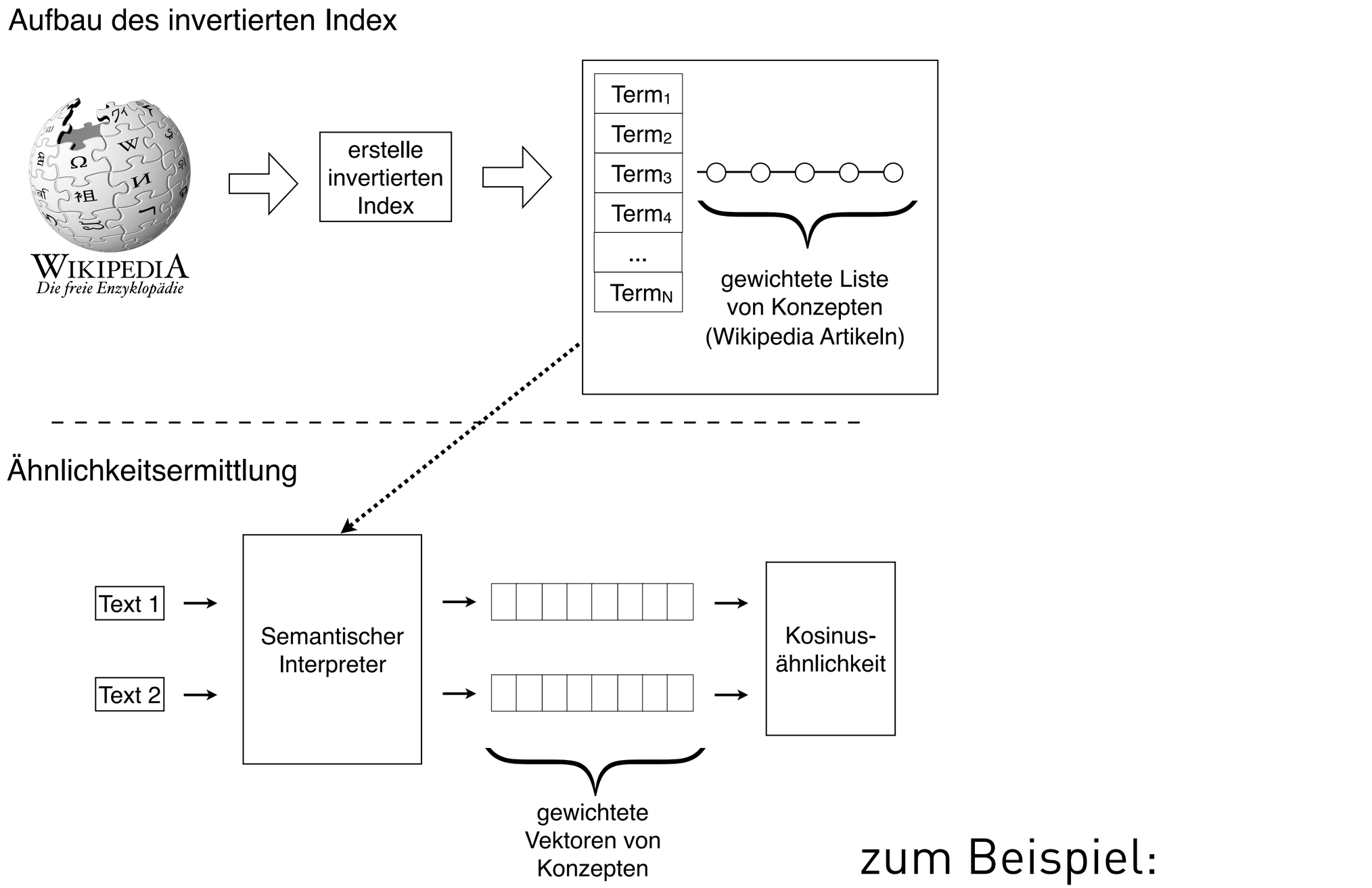




Explizite Semantische Analyse [ESA]



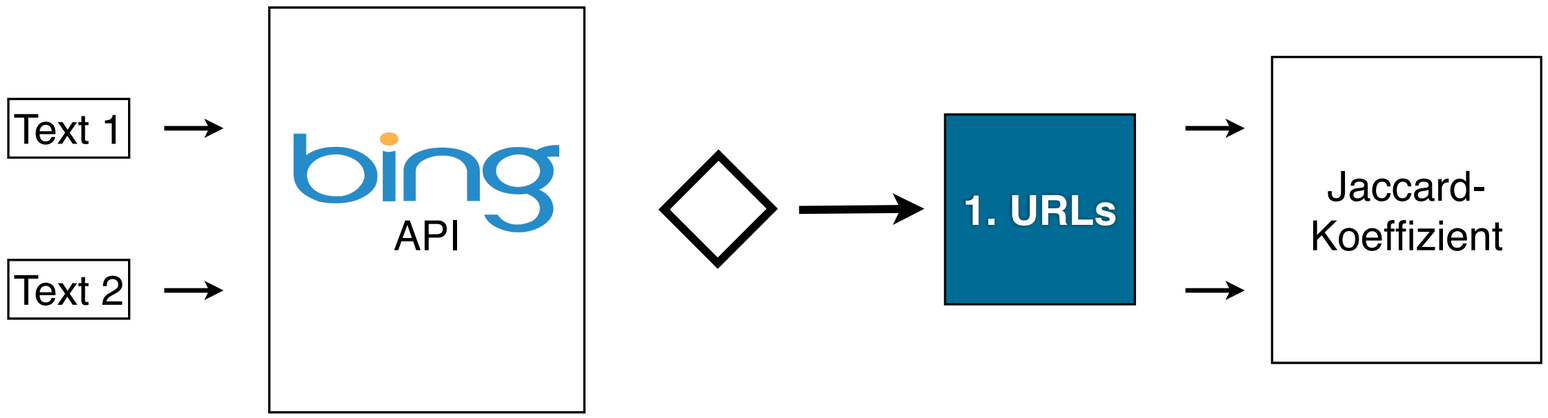
Explizite Semantische Analyse [ESA]



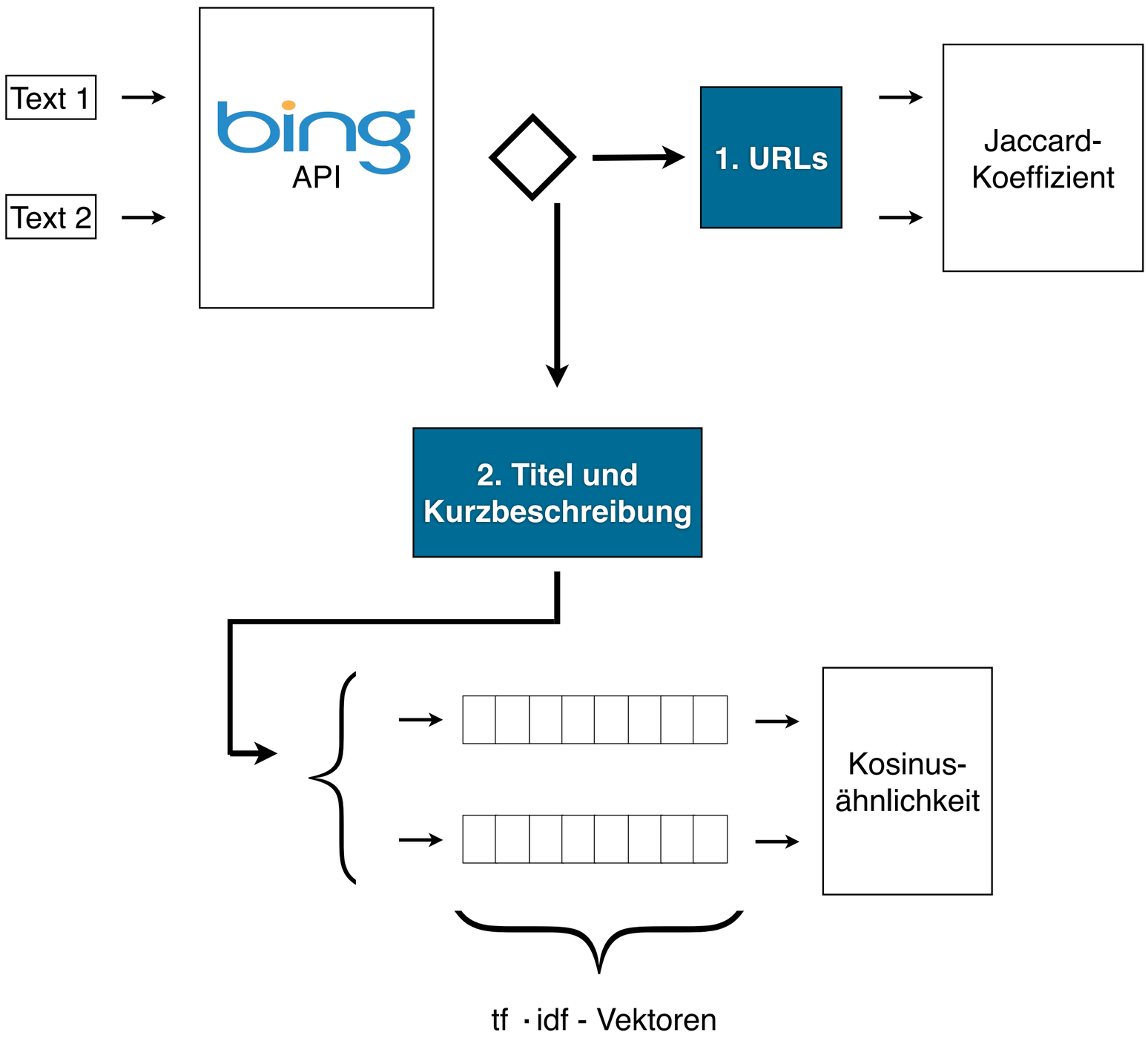
zum Beispiel:

```
4253773 istanbul archeology
4253773 constantinople
```

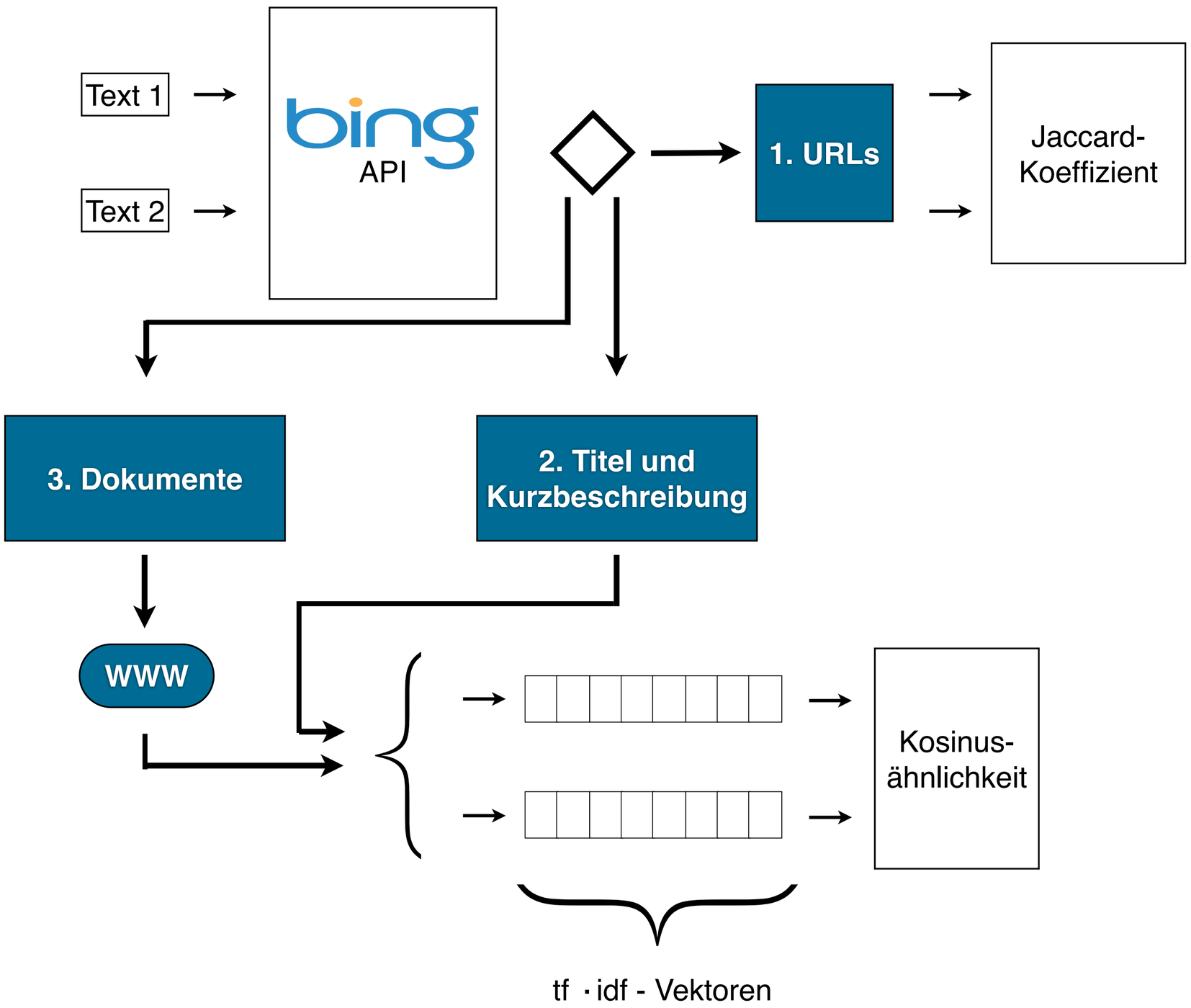
Quantifizierung der Ergebnisse einer Suchmaschine



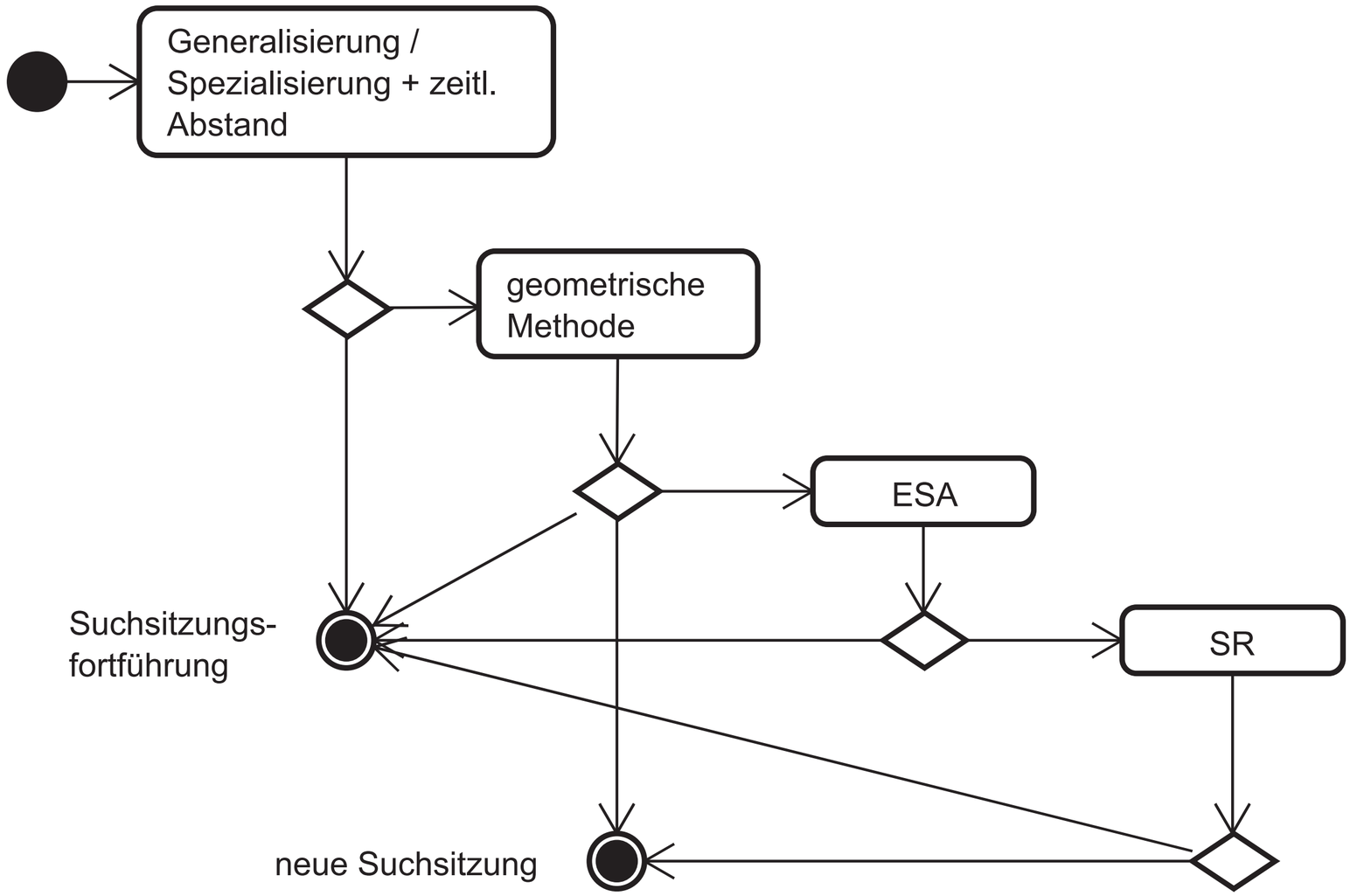
Quantifizierung der Ergebnisse einer Suchmaschine



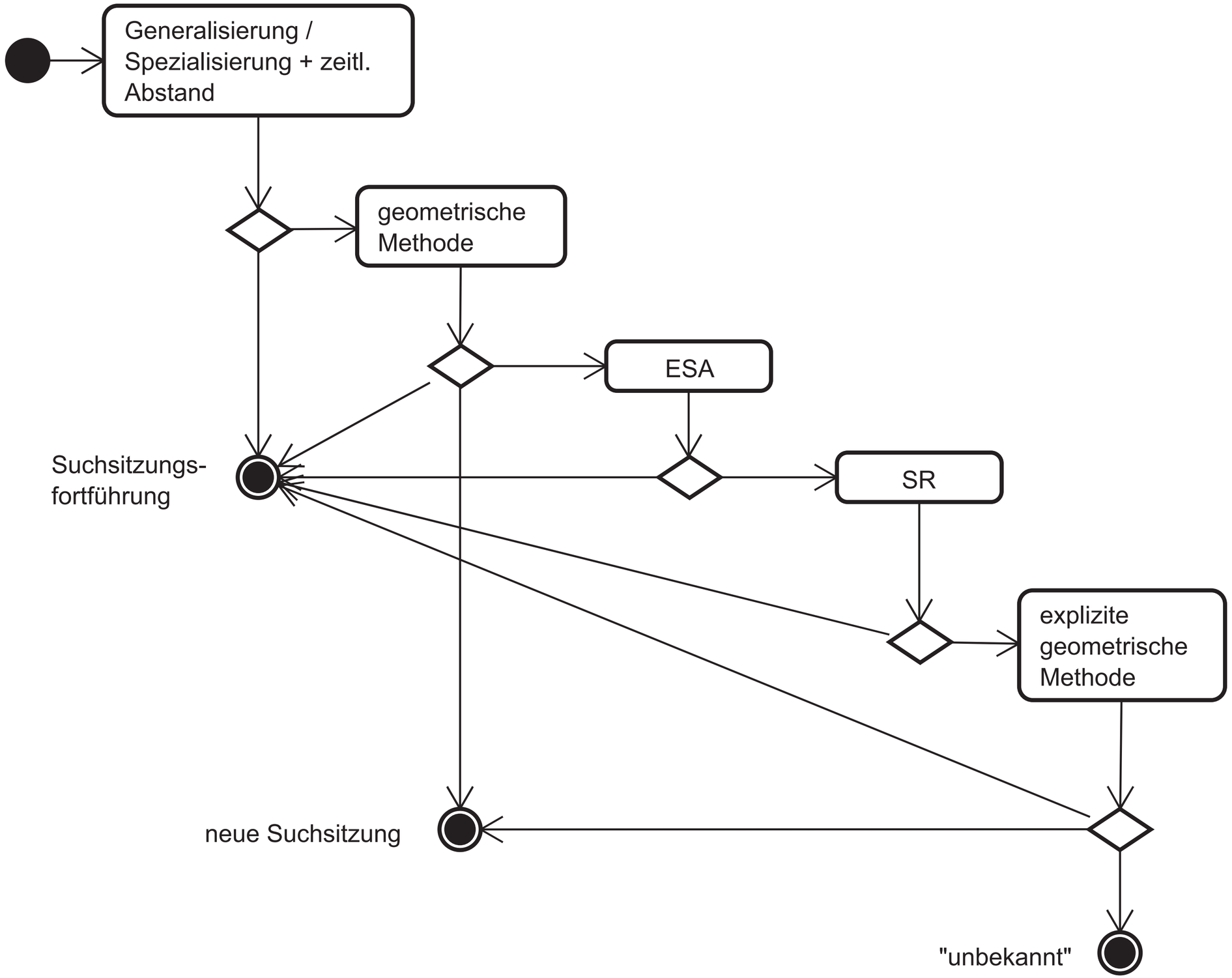
Quantifizierung der Ergebnisse einer Suchmaschine



Kaskadierendes Verfahren



Kaskadierendes Verfahren (subsample)



Ergebnisse

	precision	recall	$F_{\beta=1,5}$
geometrische Methode	0,86726	0,94306	**0,91836**
kaskadierendes Verfahren	0,86175	0,96757	**0,93234**
kaskadierendes Verfahren (subsample)	0,96799	0,97884	**0,97548**

Ergebnisse

	precision	recall	$F_{\beta=1,5}$
geometrische Methode	0,86726	0,94306	**0,91836**
kaskadierendes Verfahren	0,86175	0,96757	**0,93234**
kaskadierendes Verfahren (subsample)	0,96799	0,97884	**0,97548**

	Generalisierung / Spezialisierung	geometrische Methode	ESA	SR	expl. geom. Methode
Anteil	40,49 %	35,15 %	2,05 %	0,85 %	21,45 %
Laufzeit	1	x 2,25	x 2,43	≫ x 10	-

Zusammenfassung und Ausblick

Zusammenfassung

- generell Verbesserung der Ergebnisse
- für Teil von 77,23 % Anfragen (58,4 % der Suchsitzungen) sehr gute Ergebnisse
- Die ESA ist ca. 60 % Deckungsgleich mit den Ergebnissen auf Basis der SR

Zusammenfassung und Ausblick

Zusammenfassung

- generell Verbesserung der Ergebnisse
- für Teil von 77,23 % Anfragen (58,4 % der Suchsitzungen) sehr gute Ergebnisse
- Die ESA ist ca. 60 % Deckungsgleich mit den Ergebnissen auf Basis der SR

Ausblick

- Multitasking und hierarchische Relationen
- ESA: zusätzliche Dokumentenkollektionen (Wikitionary, WordNet Database, ...)
- weitere Ideen zu Bestimmung und Berücksichtigung der Aussagesicherheit

Vielen Dank!