

Text Mining for Second Screen

Anwendungen 2

Ivan Demin

`ivan.demin@haw-hamburg.de`



HAW HAMBURG

5. Juni 2014

Agenda

- 1 Motivation
- 2 Learning to link with wikipedia
- 3 Generating Queries from User-Selected Text
- 4 Feeding the Second Screen
- 5 Ausblick

Second Screen

First Screen

- Dient zum Medienkonsum

Second Screen

- Dient zur Interaktion mit dem Gesehenen (z.B. Informationsabruf)

Fragen

- Design
- Kommunikationsstruktur
- Informationsbereitstellung



Second Screen mit Fernseher und Tablet-Computer

Doku Anwendungsfall



Winterspiele in Sotschi auf dem First Screen [1]

Doku Anwendungsfall



Geografische Informationen über Sotschi auf dem Second Screen [1] [2]

Doku Anwendungsfall



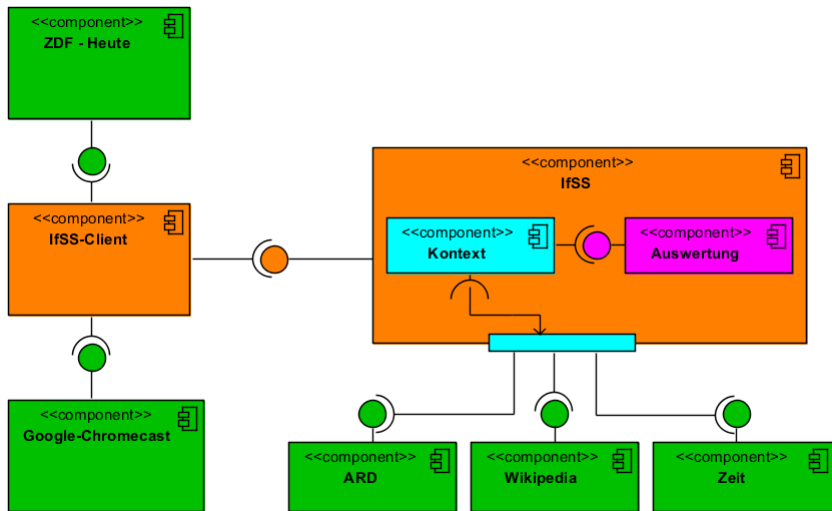
Zentraler Olympia-Park in Adler (Sotschi) [1] [3]

Doku Anwendungsfall



Kritische Informationen über Naturvernichtung um Sotchi [1] [4]

Information for Second Screen (IfSS)



Organisation der möglichen beteiligten Komponenten

Learning to link with wikipedia

2008, University of Waikato, New Zealand [MW08]

- Ansatz zum automatischen Querverweis von Dokumenten mit Wikipedia-Artikeln
- Trainingsdaten: 500 Wikipedia-Artikel mit rund 50.000 Verlinkungen
- Zwei Phasen: *Begriffsklärung*, *Link-Ermittlung*



Ausschnitt Dokumentphrasen mit assoziierten Wikipedia-Artikeln [MW08]

Phase: Begriffsklärung

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (DFS) is an **algorithm** for traversing or searching a **tree** **tree structure** or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before **backtracking**.

Formally, DFS is an **uninformed search** that progresses by expanding the first child node of the search **tree** that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search **backtracks**, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a **LIFO stack** for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	2.57%	63.26%
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
...		

Disambiguierungsrepräsentation mit eindeutigen Links als Kontext [MW08]

Phase: Link-Ermittlung

Feature-Set

- ① Link-Wahrscheinlichkeit
- ② Verbundenheit
- ③ Konfidenz der Begriffsklärung
- ④ Allgemeingültigkeit
- ⑤ Lokation

Untersuchung

Versuchsaufbau

- Testdaten: 50 Nachrichtenartikel mit jeweils 250-300 Wörtern
- Untersucht durch: 88 Probanden
- Aufgabe: Korrektheit der generierter Links bestimmen

correct	76.4
incorrect (wrong destination)	0.9
incorrect (irrelevant and/or unhelpful)	19.8
incorrect (unknown reason)	2.9

Korrektheit von automatisch generierten Links [MW08]

Fazit

- Interessanter Ansatz zum Erlernen der Wikifizierung
- Mögliche Auflösung der Mehrdeutigkeit
- Erleichterung der Probandensuche durch *Mechanical Turk*¹

¹<https://www.mturk.com/mturk/welcome>

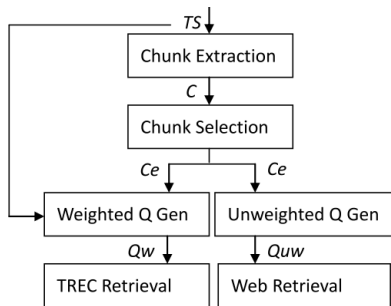
Generating Queries from User-Selected Text

2012, University of Massachusetts, Amherst [LC12]

- Ansatz zur Generierung von Anfragen aus ausgewählten Textpassagen

Bedeutung der Variablen

TS	=	text segment
C	=	chunk
C_e	=	effective chunks
Q_w	=	weighted query
Q_{uw}	=	unweighted query



Framework Design [LC12]

Chunk Selection

Häufigkeits-beruhender Ansatz

- Aufsummierung der Ergebnisse aus Websuche für einen Chunk
- Top k der häufig auftretenden Chunks als C_e verwendet

Maschinelles Lernansatz

- Identifizierung von C_e durch CRF-Modelle
- Jeder Chunk wird mit einem Label versehen

Feature-Set

- Wikipedia Titel
- Anzahl der Suchergebnisse
- Google ngram

Query Generation

Gewichtete Anfrage-Generation

- Aus TS wird TS_n ohne Stoppwörter berechnet
- Auf Basis der aufsummierten Suchergebnisse für ein Chunk
- Kombiniert mit einem Lernfaktor

Ungewichtete Anfrage-Generation

- Besten zwei bewerteten Chunks für die Suche verwendet

Untersuchung

Versuchsaufbau

- Testdaten: *TREC Gov2* (rund 25 Mio. Dokumente)
- Untersucht durch: Autoren
- Aufgabe: Korrektheit der gefundenen Dokumente bestimmen

Ergebnisse

- Bestes Ergebnis für: $TS_n + TopC(k)$ mit 20 % MAP

Fazit

- Merkmalsauswahl ist nicht nachvollziehbar
- Ansatz liefert keine guten Ergebnisse
- Ansatz ist für meine Belange nicht erstrebenswert

Feeding the Second Screen:

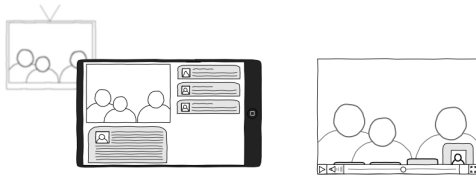
Semantic Linking based on Subtitles

2013, University of Amsterdam, Amsterdam [OMdR13]

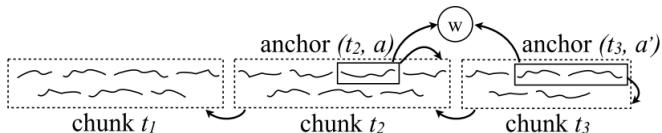
- Generierung von Wikipedia-Links aus Untertiteln

Ansatz

- 1 Link-Kandidaten finden
- 2 Sortieren
- 3 Umsortieren



Skizze Talkshow-Anwendungsfall [OMdR13]



Kontext-Graph als Repräsentation der Link-Kandidaten [OMdR13]

Kontext-Graph

Feature-Set

- ① Fixpunkt
- ② Zielpunkt
- ③ Fixpunkt + Zielpunkt
- ④ Kontext

Untersuchung

Versuchsaufbau

- Testdaten: Untertitel aus sechs Talkshow-Episoden (rund 36k Wörter)
- Untersucht durch: Trainierten Annotationsspezialisten
- Aufgabe: Korrektheit der generierter Links bestimmen

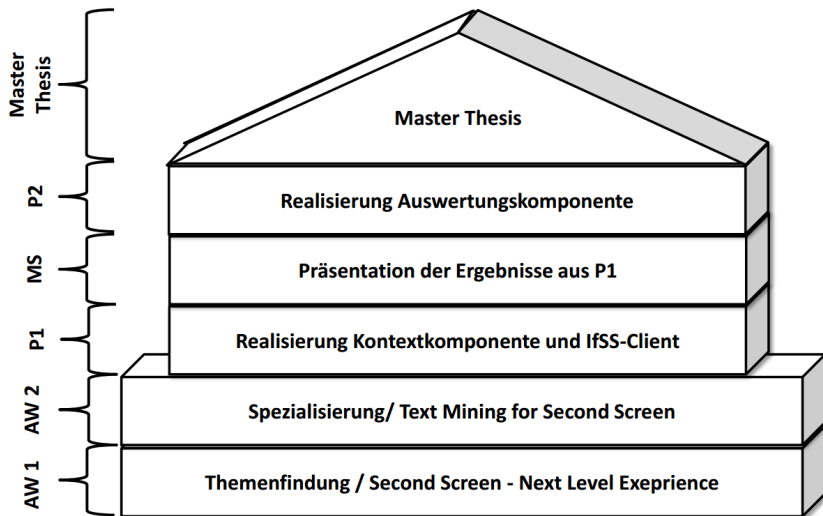
	Average classification time per chunk (in ms)	R-Prec	MAP
1. Baseline retrieval model	54	0.5753	0.6235
2. Learning to rerank approach	99	0.7177	0.7884
<i>Learning to rerank (L2R) + one context graph feature</i>			
3. L2R+ <i>DEGREE</i>	104	0.7375	0.8252
4. L2R+ <i>DEGREECENTRALITY</i>	108	0.7454	0.8219
5. L2R+ <i>PAGERANK</i>	119	0.7380	0.8187
<i>Learning to rerank (L2R) + three context graph features</i>			
6. L2R+ <i>DEGREE</i> + <i>PAGERANK</i> + <i>DEGREECENTRALITY</i>	120	0.7341	0.8204

Korrektheit von automatisch generierten Links [OMdR13]

Fazit

- Interessante Erweiterung des Ansatzes aus Paper 1
- Modellierung vom Kontext als Graphen nachahmenswert
- Training vom Annotationsspezialisten nicht nachvollziehbar

Ausblick



Fahrplan Master

Vielen Dank für die Aufmerksamkeit!
Fragen? 😊

- [LC12] Chia-jung Lee and W Bruce Croft.
Generating Queries from User-Selected Text.
2012.
- [MW08] David Milne and Ian H. Witten.
Learning to link with wikipedia.
In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, page 509, New York, New York, USA, October 2008. ACM Press.
- [OMdR13] Daan Odijk, Edgar Meij, and Maarten de Rijke.
Feeding the second screen: semantic linking based on subtitles.
pages 9–16, May 2013.

- [1] Sotchi Dokumentation - <http://dokumentation.zdf.de/ZDF/zdfportal/blob/31160078/2/data.jpg> - Abruf 30.03.2014
- [2] Artikel Sotchi - <http://de.wikipedia.org/wiki/Sotchi> - Abruf 30.03.2014
- [3] Aufbau Olympisches Lager - http://commons.wikimedia.org/wiki/File:Sochi_2014_olympic_coastal_cluster_map-en.svg - Abruf 30.03.2014
- [4] Kritischer Artikel über Sotchi - <http://zdf.de/ZDF/zdfportal/xml/object/31550864> - Abruf 30.03.2014