Crowdsourcing für die Annotation von Webgenres

Katja Markert

Forschungszentrum L3S
Universität Hannover
und
School of Computing
University of Leeds

Hannover 2015

Genre

Genredefinition

"a distinctive type of communicative action, characterized by a **socially recognized** communicative **purpose** and common aspects of **form**." (Orlikowski 1994)

- Beispiele: Nachrichten, Leserbriefe, Tutorial
- Kontrast zu rein funktionalen Kategorien: persuasion, discussion
- Kontrast zu Themen: Sport, Wirschaft

Warum Genre?

- Digitale Aufbereitung großer Textbestände
- Suchmaschinen: Google Scholar, Google News; Archivsuche
- Genreadaption von Sprachmodellen: z.B. Wortartenannotation (Giesbrecht und Evert 2009)

Übersicht

Übersicht

Existierende Webgenrekorpora: Probleme

▸ Details

Korpus	Genres	Seiten	min	max	med	Annotation
KI-04	8	1205	126	205	145	einzeln
I-EN-S.	7	250	10	99	30	α=0.55
SANTINI	7	1400	200	200	200	einzeln
MGC	20	1536	55	227	77	α=0.56
HGC	7/32	1280	40	40	40	m. einzeln
KRYS I	70	6200	6	117	97	45-53%
Syracuse	292	3027	1	174	3	einzeln
Egbert 13	8/56	1000	0	99	1	64/43%

Kein allgemein anerkanntes Genreinventar

- manchmal vage (*article*) oder rein funktional (*entertainment*)

  ▸ Beispiele

- manchmal extrem detailliert (*How-to vs. Instruction*)

Keine Doppelannotation, nicht verlässliche Annotation

Existierende Webgenrekorpora: Probleme

Santini FAQ (200)	
110	hurricane
109	noaa
107	center
84	aoml
65	tropical
57	tax

Resultate auf Santinis (Lrec 2010)	
Bag of Words	95.86

Daher

- Irreführende Genre- und Themenkorrelationen, besonders bei designten Korpora (Sharoff, Wu, Markert 2010)
- Derzeitige automatische Genreklassifikation überbewertet? (Petrenz und Webber 2011)

Ziele des Leeds Web Genre Corpus (LWGC)

- **Verlässlicher** und **themendiversifizierter** Korpus
  - naive Annotierer: soziale Erkennbarkeit von Genres
  - Trennung von Entwicklern und Annotierern
  - schnell und billig $\longrightarrow$ leichte Erweiterbarkeit
- **LWGC-B(alanced):** Design durch fokussierte Suche $\longrightarrow$ Prototypen und gute Anzahl Seiten per Genre
- **LWGC-R(andom):** Zufallsauswahl $\longrightarrow$ Anwendbarkeit auf beliebige Webseiten, Abdeckung

- Kollaboration: Serge Sharoff, Zhili Wu und Noushin R. Al-Sheghi
- Förderung: Google Research Award, EPSRC
- Papiere: LREC 2010, ACL 2010, LREC 2014, TextGraphs 2014, eingereicht *Language Resources and Evaluation 2015*

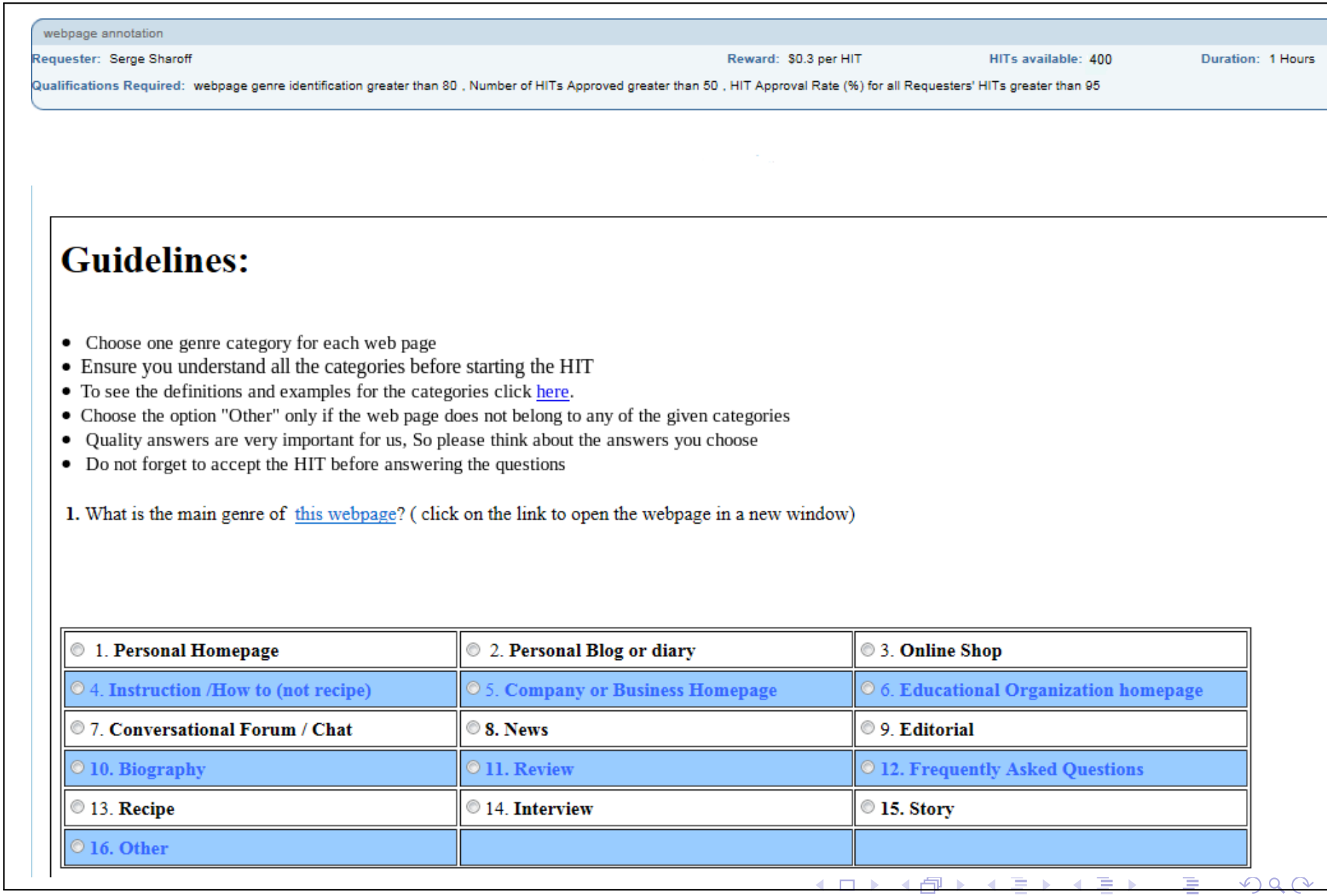
Übersicht

Anfangsinventar von 15 Genres

- Form und Funktion
- Soziale Erkennbarkeit
- Textorientiert

Personal Homepage (php)
Company Homepage (com)
Educational Org. Homepage (edu)
Personal Blog (blog)
Online shop (shop)
Recipe
Instruction/How-to (instr.)
News article (news)
Editorial
Forum
Biography (bio)
Frequently Asked Questions (FAQ)
Product Review (Review)
Interview
Story

Experimentanordnung auf Mechanical Turk



webpage annotation

Requester: Serge Sharoff Reward: $0.3 per HIT HITs available: 400 Duration: 1 Hours

Qualifications Required: webpage genre identification greater than 80 , Number of HITs Approved greater than 50 , HIT Approval Rate (%) for all Requesters' HITs greater than 95

Guidelines:

- Choose one genre category for each web page
- Ensure you understand all the categories before starting the HIT
- To see the definitions and examples for the categories click here.
- Choose the option "Other" only if the web page does not belong to any of the given categories
- Quality answers are very important for us, So please think about the answers you choose
- Do not forget to accept the HIT before answering the questions

1. What is the main genre of this webpage? (click on the link to open the webpage in a new window)

1. **Personal Homepage**	2. **Personal Blog or diary**	3. **Online Shop**
4. **Instruction /How to (not recipe)**	5. **Company or Business Homepage**	6. **Educational Organization homepage**
7. **Conversational Forum / Chat**	**8. News**	9. **Editorial**
10. Biography	**11. Review**	**12. Frequently Asked Questions**
13. **Recipe**	14. **Interview**	**15. Story**
16. Other		

Qualitätskontrolle

Verhindert Bots, Zufallsauswahl (Mason et al. 2012)

- **Systemqualifikationen:** mindestens 50 akzeptierte HITs, Akzeptanzquote $> 95\%$
- **Test vor Arbeitsbeginn:** 10 typische Genreseiten $> 80\%$
- **Kontrolle während Arbeit:** 1 Seite per HIT "honeypot"
- 5 Annotatoren per Webseite (Beigman 2009, Snow et al 2008)

Übersicht

LGWC-B(alanced): Charakteristika

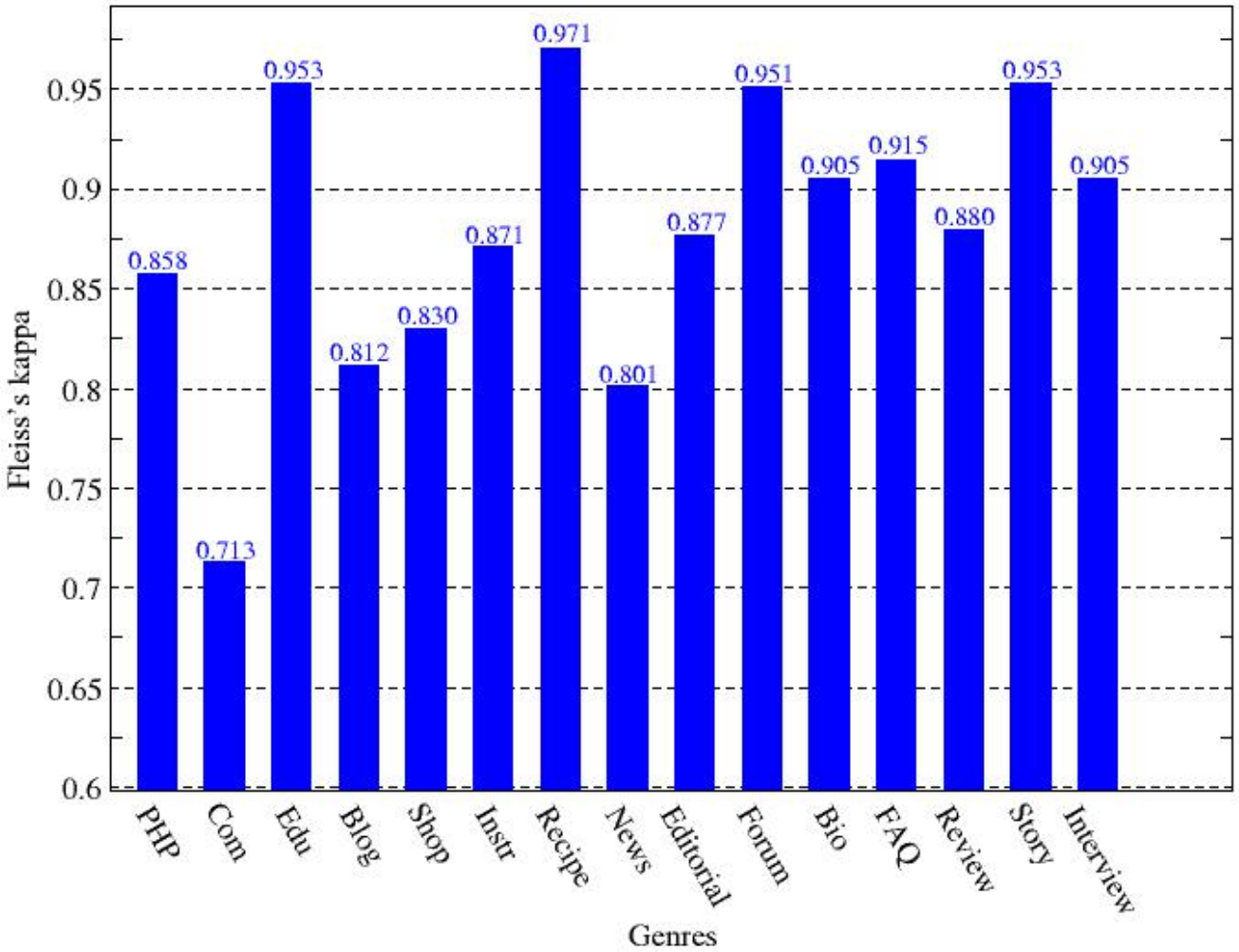
Genres	15
Design	3964 Seiten aus Yahoo und Open Directory
Herkunftsdivers.	Median Seite pro Site 1 für 13/15 Kategorien
Beispiel Bio	242 Seiten, 190 Sites
Speicherung	HTML + Screenshots
Dauer	7 Tage
Annotatoren	42
Kosten	820 Dollar
Kleinste Kat. Story	184
Größte Kat. Rezept	332

Schnell, günstig, gute Herkunftsdiversifizierung

▸ Details zu Herkunftsdiversifizierung

LWGC-B: Verlässlichkeit

Insgesamt: 88.2%, Kappa 87.4%.



LWGC-B: Themendiversifizierung

Schlüsselworte

Häufiger in einer Webseite als im Gesamtkorpus (Loglikelihood)

	Webseite	Gesamtkorpus
Worthäuf.	a	b
Häuf. anderer Wörter	c-a	d-b

LWGC-B FAQ (201)	
58	can
51	question
46	information
45	do
33	are
32	does
28	how
28	frequently

Santini FAQ (200)	
110	hurricane
109	noaa
107	center
84	aoml
65	tropical
57	tax

Weitere Beispiele zur Themendiversifizierung

LWGC-B Story (184)	
72	said
37	then
34	could
33	old
....	...
28	king
27	thought
25	stood
25	went

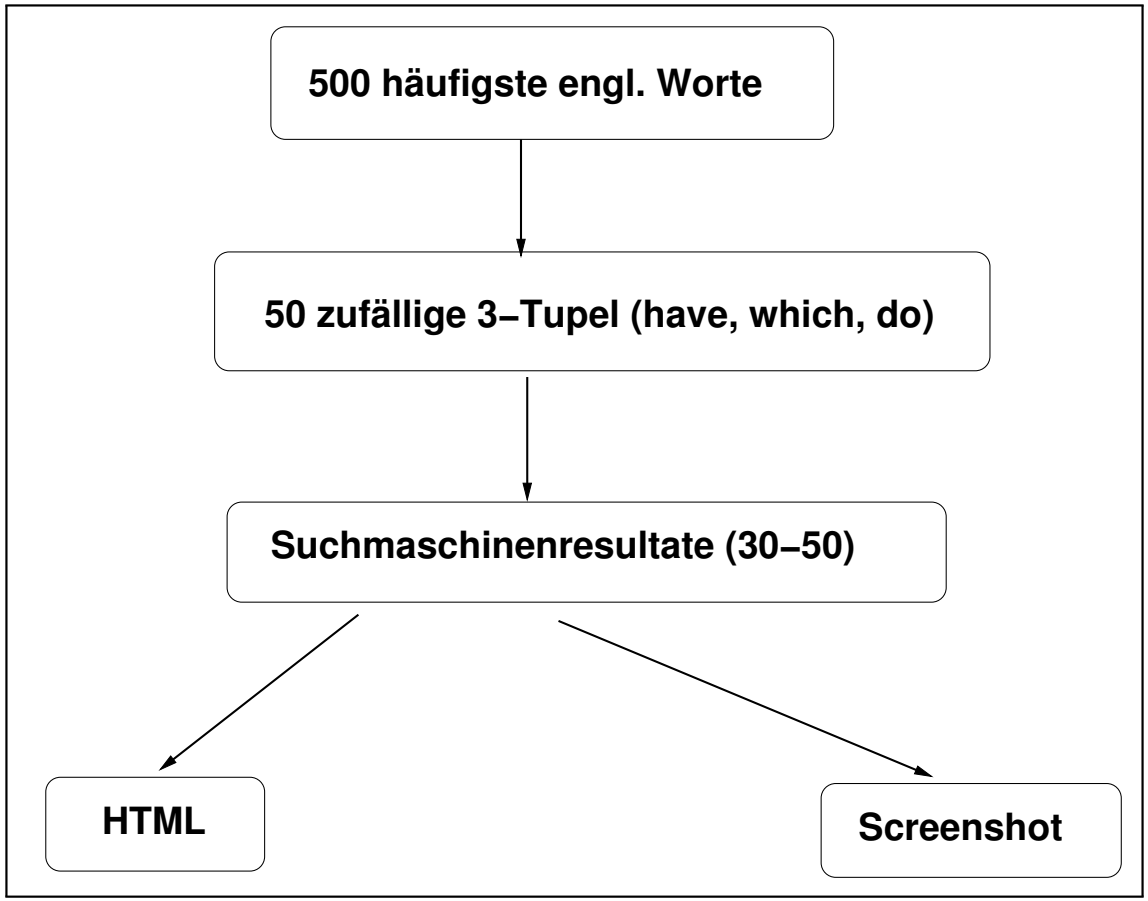
LWGC-B Bio (242)	
63	biography
39	became
27	had
26	music
....	...
23	died
21	published
21	born
20	married

LWGC-B Blog (244)	
70	posted
50	january
47	comments
46	blog
....	...
23	april
23	am
23	about
22	october

- Funktionswörter
- Genrenamen
- sehr spezifische Strukturen der Genres (*comments*, *posted*)
- Aufnahme von funktionstypischen Merkmalen: past tense, Diskurspartikel
- Wenige Themenkorrelationen

LWGC-R(andom): Textkollektion

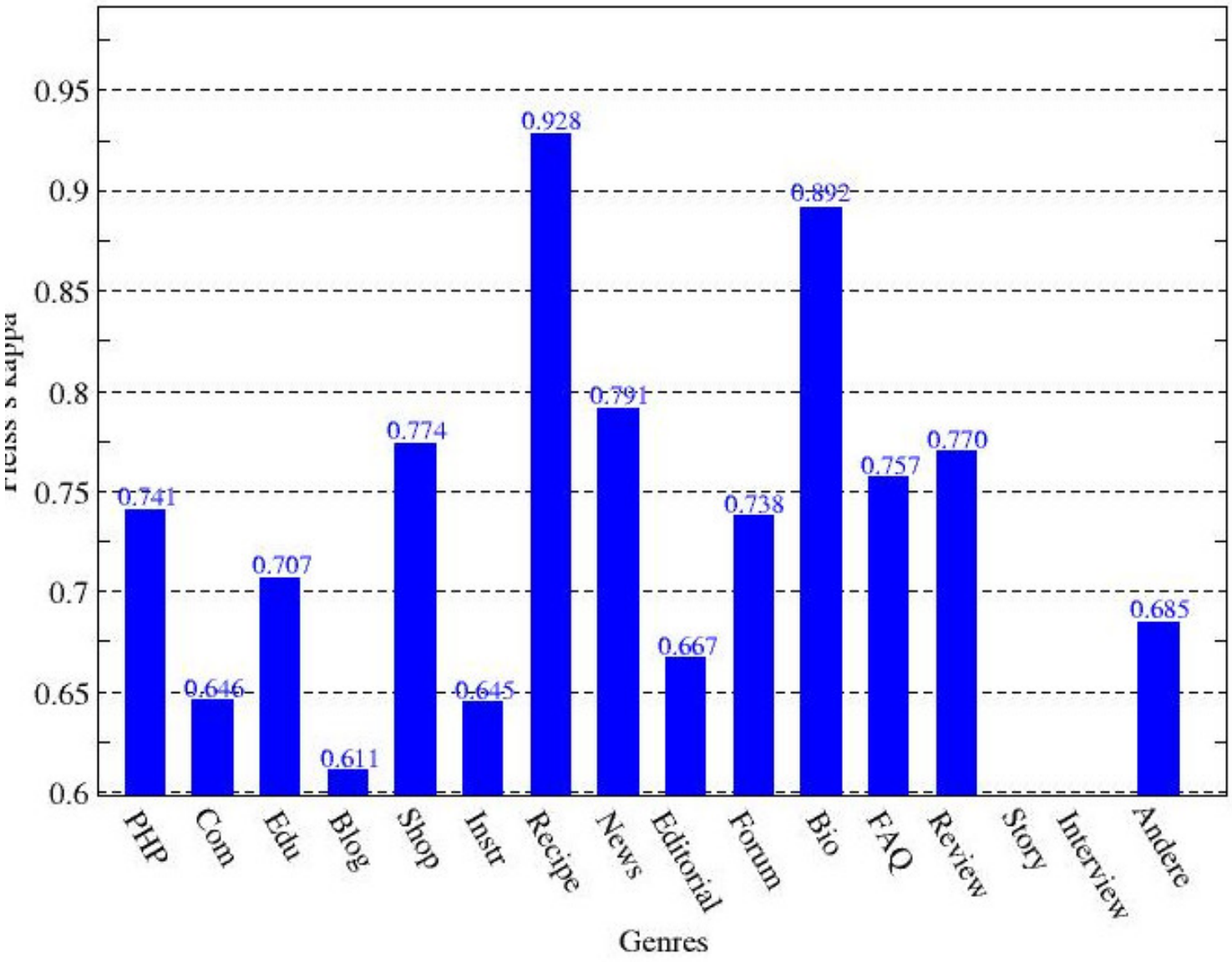
Random conjunctive queries:



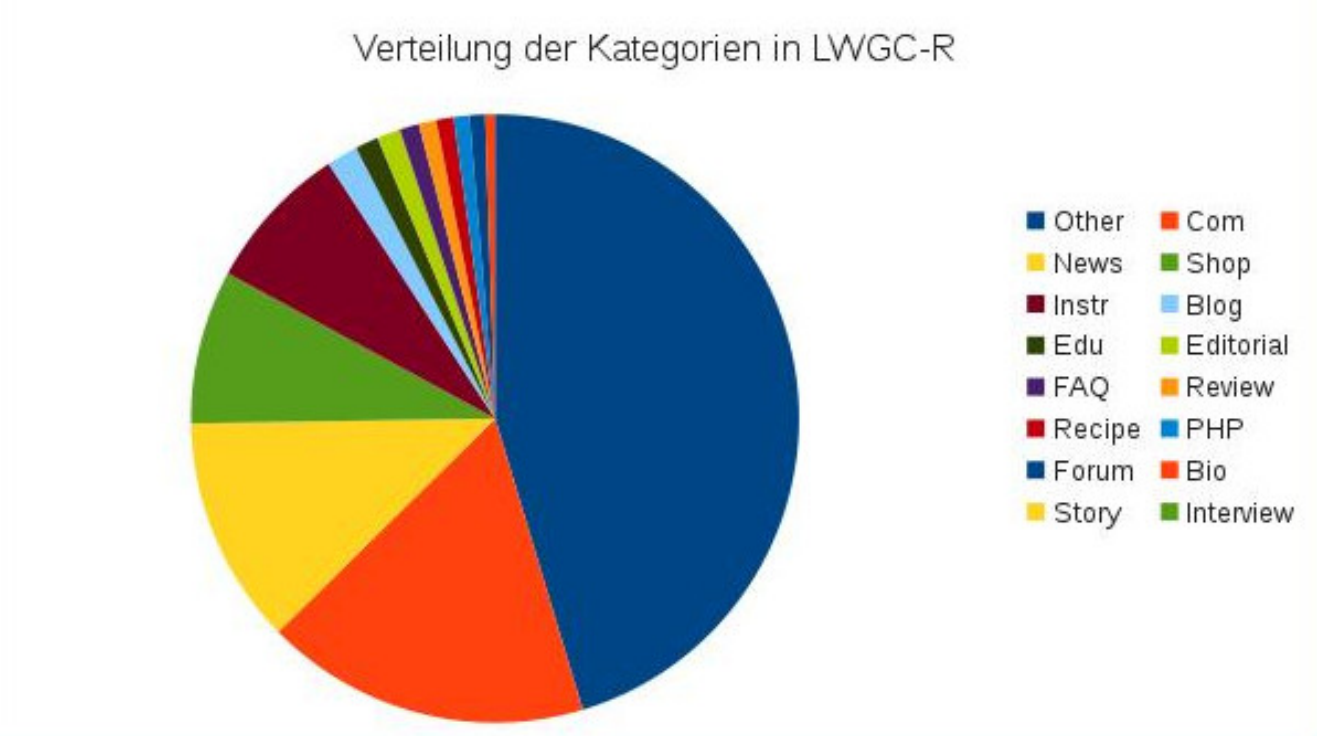
Nicht völlig zufällig; Tendenz zu populären Seiten

LWGC-R: Verlässlichkeit

Insgesamt: 78.5%, Kappa 0.71.



LWGC-R: Abdeckung der Kategorien



- Zufallsauswahl erschwert Kollektion nicht sehr populärer Kategorien wie story oder interview
- Abdeckung des Genreinventars auf LWGC-R: 56.2%

Erhöhung der Abdeckung

- Annotation für 438 Seiten, die in LWGC-R als "other" klassifiziert wurden.
- Kappa: 0.65
- Erhöht Abdeckung auf 76%.

encyclopedic	97
index/list	56
dictionary	22
quotes	13
song lyrics	3
other	247

LWGC-R: Themendiversifizierung

LWGC-R Blog (16)	
5	february
4	reply
4	november
4	do
....	...
3	posted
3	december
3	comment
3	by

LWGC-R Shop (79)	
18	products
18	product
15	see
15	price
....	...
10	supplies
10	amazon
10	reviews
10	item

LWGC-R News (117)	
41	news
24	said
19	james
15	season
....	...
11	colorredteam
10	lebron
10	nba
10	league

- Meistens ok
- Sind beliebte Seiten (Amazon) überrepräsentiert oder sollte das so sein?
- Temporaler Aspekt: temporale Diversifizierung?

Übersicht

Zusammenfassung

- **Erstes verlässlich annotiertes** Korpus für Webgenres
- Crowdsourcing **billig, schnell** und kann auf viele Genres sowie Webseiten angewandt werden
- **Leeds Web Genre Corpus Balanced**: viele Seiten per Kategorie, gute Themendiversifizierung, sehr hohe Verlässlichkeit
- **Leeds Web Genre Corpus Random**: unbalanciert, Seiten außerhalb des jetzigen Genreinventars, hohe Verlässlichkeit

Ausblick

- Erweiterung: Genres, temporal (Webarchive), multilingual
- Herausgabe: TEI, Copyright
- Gattungsanalyse: Grammatik, Diskurs, Layout und Bild
- Maschinelle Genreerkennung ▸ Details
- Digitale Gattungsgeschichte
- Genreadaption von computerlinguistischen Algorithmen

Existierende Webgenrekorpora: Probleme

▸ zurück

Korpus	Genres	Seiten	min	max	med	Sammlung	Annotation
KI-04	8	1205	126	205	145	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
SANTINI	7	1400	200	200	200	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
MGC	20	1536	55	227	77	Zuf.+Des.	α=0.56
HGC	7/32	1280	40	40	40	Design	m. einzeln
KRYS I	70	6200	6	117	97	Design	45-53%
Syracuse	292	3027	1	174	3	Design	einzeln
Egbert 13	8/56	1000	0	99	1	Zufall	64/43%

Kein allgemein anerkanntes Genreinventar

- manchmal sehr vage (*article*) oder rein funktional (*entertainment*, *discussion*) ▸ Beispiele
- manchmal extrem detailliert (*How-to vs. Instruction*, Egbert und Biber 2013)

Existierende Webgenrekorpora: Probleme

zurück

Korpus	Genres	Seiten	min	max	med	Sammlung	Annotation
KI-04	8	1205	126	205	145	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
SANTINI	7	1400	200	200	200	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
MGC	20	1536	55	227	77	Zuf.+Des.	α=0.56
HGC	7/32	1280	40	40	40	Design	m. einzeln
KRYS I	70	6200	6	117	97	Design	45-53%
Syracuse	292	3027	1	174	3	Design	einzeln
Egbert 13	8/56	1000	0	99	1	Zufall	64/43%

Manchmal sehr wenige Seiten per Genre

Existierende Webgenrekorpora: Probleme

zurück

Korpus	Genres	Seiten	min	max	med	Sammlung	Annotation
KI-04	8	1205	126	205	145	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
SANTINI	7	1400	200	200	200	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
MGC	20	1536	55	227	77	Zuf.+Des.	α=0.56
HGC	7/32	1280	40	40	40	Design	m. einzeln
KRYS I	70	6200	6	117	97	Design	45-53%
Syracuse	292	3027	1	174	3	Design	einzeln
Egbert 13	8/56	1000	0	99	1	Zufall	64/43%

Design oder Zufall sollten sich ergänzen (Rehm et al 2008)

Existierende Webgenrekorpora: Probleme

zurück

Korpus	Genres	Seiten	min	max	med	Sammlung	Annotation
KI-04	8	1205	126	205	145	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
SANTINI	7	1400	200	200	200	Design	einzeln
I-EN-S.	7	250	10	99	30	Zufall	α=0.55
MGC	20	1536	55	227	77	Zuf.+Des.	α=0.56
HGC	7/32	1280	40	40	40	Design	m. einzeln
KRYS I	70	6200	6	117	97	Design	45-53%
Syracuse	292	3027	1	174	3	Design	einzeln
Egbert 13	8/56	1000	0	99	1	Zufall	64/43%

Keine Doppelannotationen oder nicht verlässliche Annotationen

Kappa	Höhe
$<$0.4	Schlecht
0.4 – 0.6	Moderat
0.6 – 0.8	Substantiell
0.8 – 1.0	Perfekt

Beispiele von Genreinventaren

zurück

KI-04 Nutzerbefragung, Mischung Funktion und Form

Article	Link Collection
Discussion	Non-pers. homepage
Download	Personal Homepage
Help	Shop

I-EN: Rein funktional, Expertenmeinung

Discussion	Information
Propaganda	Instruction
Recreation	Regulations
Reporting	Unknown

Beispieldefinitionen

Keine langen Definitionen, keine linguistische Terminologie

- Review: an evaluation of a publication, a product or a service such as a movie, a video game, a musical composition or a book
- Frequently asked questions: questions (with answers) commonly asked about a particular topic, in list form

Herkunftsdiversifizierung:LWGC-B

Genre	Seiten	Sites	max pro Site	med pro Site
Php	304	288	9	1
Com	264	264	1	1
Edu	299	299	1	1
Blog	244	215	9	1
Shop	292	209	23	1
Instruction	231	142	15	1
Recipe	332	116	8	1
News	330	127	12	1
Editorial	310	69	11	3
Forum	280	106	11	1
Bio	242	190	15	1
Faq	201	140	8	1
Review	266	179	15	1
Story	184	24	38	7
Interview	185	154	11	1

zurück

Überwachtes maschinelles Lernen: Resultate

zurück

Überwachter SVM auf LWGC-B in Kreuzvalidierung. 8 Teilmengen für Training, 1 für Validation, 1 für Testing.

Merkmale	Resultat
Baseline	8.38
Bag of Words	**88.98**
Buchstaben-4-grams	87.96
POS-3grams	70.28
POS-2grams	68.94
POS	60.14
Stilistik+HTML tags	55.47

Erstaunlicherweise **simple Wortfrequenzen** auch auf Genrekorpus mit Themen- und Herkunftsdiversität stark

Warum funktioniert Bag of Words für Webgenres?

LWGC-B FAQ (201)	
58	can
51	question
46	information
45	do
33	are
32	does
28	how
28	frequently
26	services

LWGC-B Bio (242)	
63	biography
39	became
27	had
26	music
....	...
23	died
21	published
21	born
20	married

LWGC-B Forum (280)	
201	posts
164	forum
143	join
137	thread
....	...
93	pm
82	quote
68	am
67	post

- Funktionswörter (alleine 75% Akk)
- sehr spezifische Strukturen der Genres (*thread*)
- Web erleichtert manches: Genrenamen (alleine 57% Akk.)

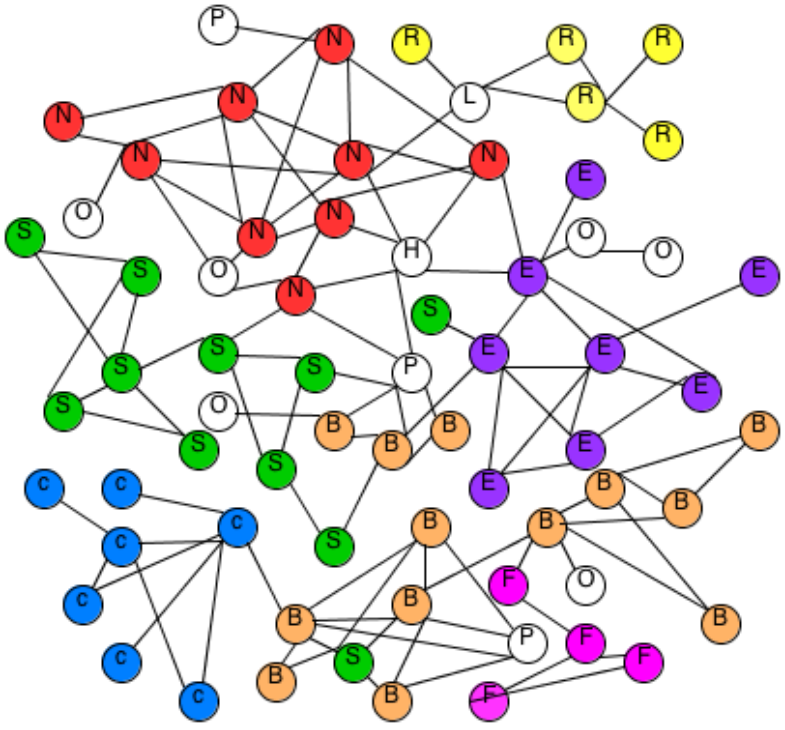
Netzwerkausnutzung

zurück

Bisher: Webseiten werden nur aufgrund ihres Inhalts klassifiziert.

Homophilie

- Annahme: Verlinkte Webseiten gehören zur gleichen Kategorie
- Oft benutzt für Themenklassifikation
- Neu und nicht notwendigerweise wahr für Genreklassifikation

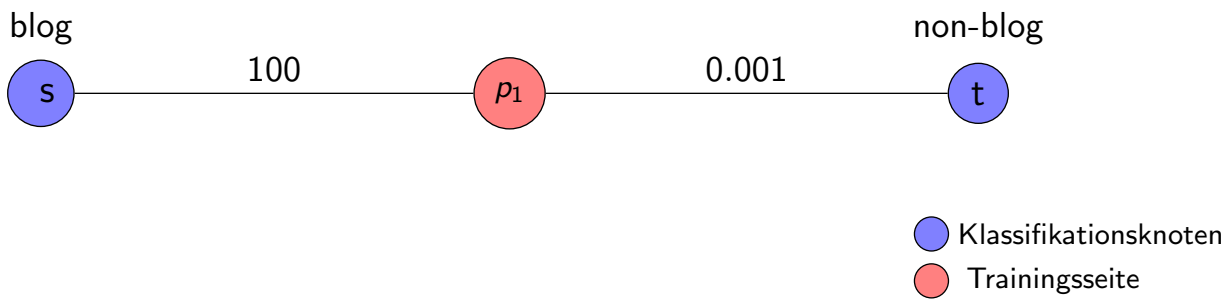


Mit einem halbüberwachten Minimum Cut Netzwerkalgorithmus ist eine kleiner, aber signifikante Verbesserung der Genreerkennung möglich (von 88.98% auf 90.11%) Details

Minimum Cut Algorithmus

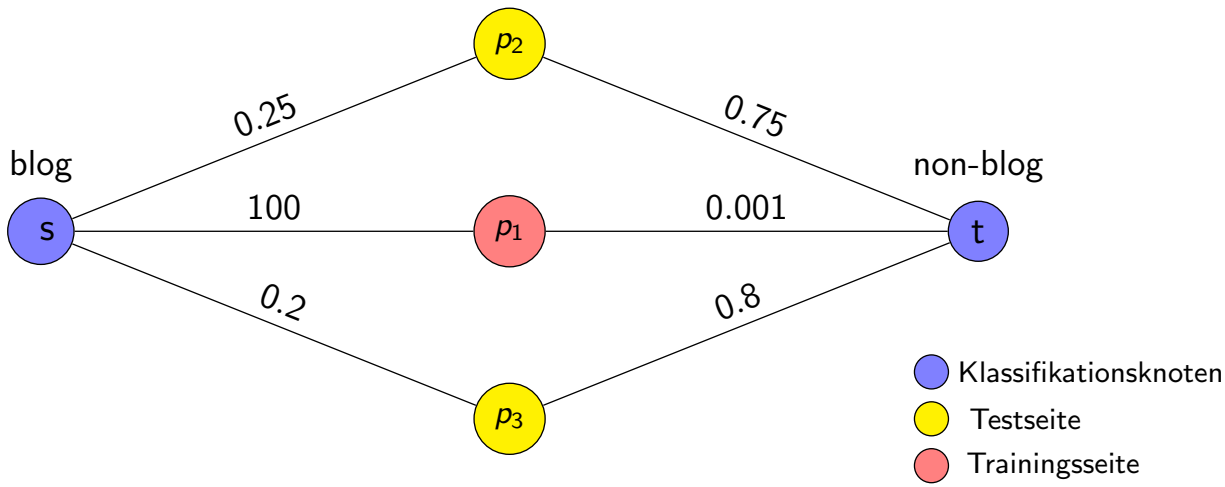
zurück

- S (source) und T (sink oder target) sind Klassifikationsknoten
- Trainingsseiten sind mit hohem konstantem Gewicht mit zugehörigem Klassifikationsknoten verbunden



Minimum Cut Algorithmus

Testseiten sind mit Klassifikationsknoten durch Gewichte verbunden, die durch überwachten Unigram-SVM bestimmt werden.

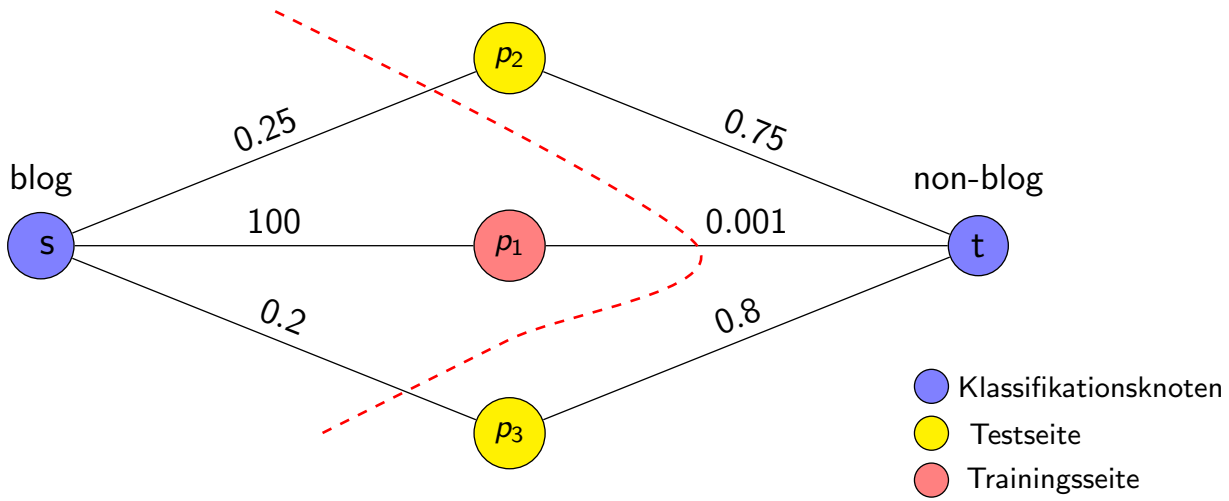


Minimum Cut Algorithmus

Mincut

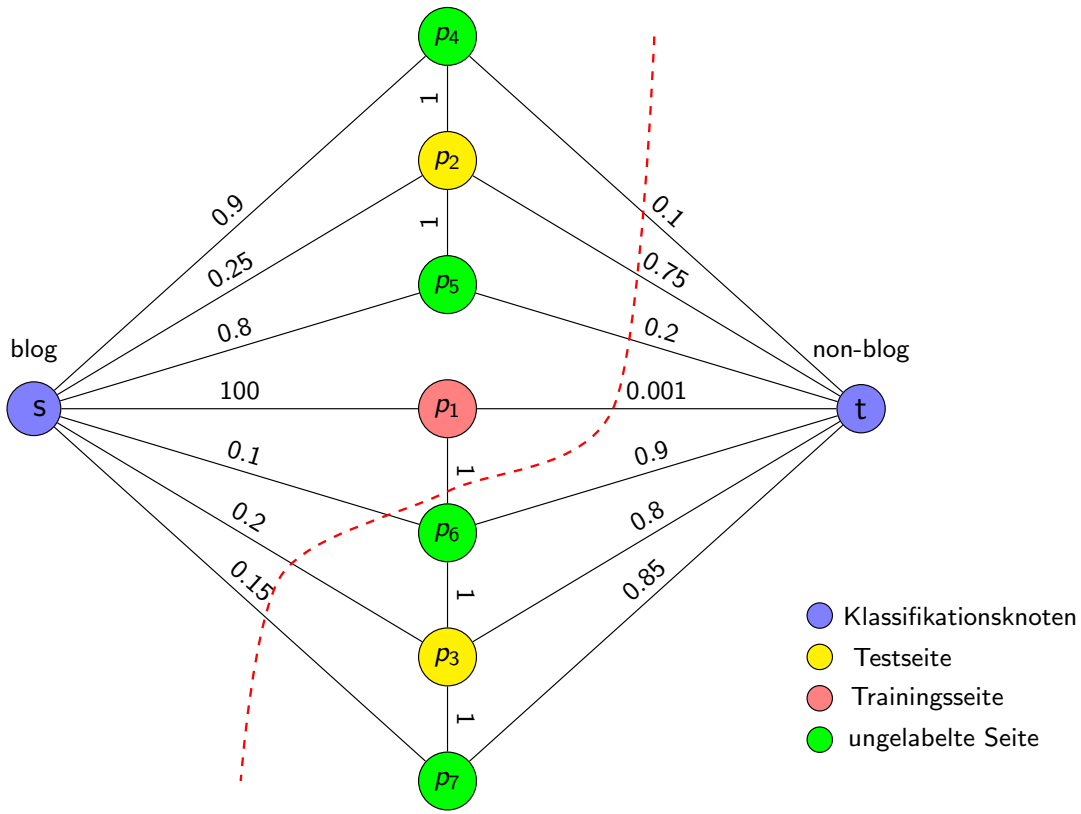
Teile Graph in zwei Hälften, so dass Summe der getrennten Gewichte minimiert wird (Boykoff et al 2001)

- Ohne Links zwischen Seiten entspricht Resultat dem des Standard-SVM



Halbüberwachter min-cut Algorithmus

- Links zwischen Seiten inkorporieren Homophilie (Gewicht 1)
- Ungelabelte Daten nötig: Kinder von Trainings- oder Testseiten



Wahl der ungelabelten Seiten

Wähle ähnliche Nachbarn ▸ Mincut mit allen Daten

$$\cos(\vec{w}, \vec{u}) = \frac{\vec{w} \cdot \vec{u}}{\| \vec{w} \| \| \vec{u} \|} = \frac{\sum_{i=1}^{n} w_i \times u_i}{\sqrt{\sum_{i=1}^{n} (w_i)^2} \times \sqrt{\sum_{i=1}^{n} (u_i)^2}} \quad (1)$$

Kosinus	# ungelabelter Seiten	Durchschn # Nachbarn
≥ 0	103,372	40.65
≥ 0.4	50,232	17.52
≥ 0.6	13,919	3.77
$\geq$ **0.8**	**3,772**	**0.98**
≥ 0.9	1,732	0.44

Resultat nach automatischer Nachbarnselektion

90.11%: signifikant besser als überwachter Algorithmus

▸ zurück

Min-cut Resultate mit allen Nachbarn

back

halbüberwachter min-cut			
Kat	R	P	F
php	0.323	0.717	0.445
forum	0.978	0.478	0.642
review	0.752	0.526	0.619
news	0.526	0.657	0.584
com	0.309	0.856	0.454
shop	0.625	0.750	0.682
instruction	0.679	0.614	0.645
recipe	0.990	0.596	0.744
blog	0.786	0.573	0.663
bio	0.952	0.491	0.648
editorial	0.907	0.283	0.432
faq	0.567	0.482	0.521
edu	0.814	0.806	0.810
story	0.982	0.625	0.764
interview	0.878	0.389	0.539
insgesamt = 59.68%			

überwacht			
Genre	R	P	F
php	0.938	0.798	0.862
forum	0.943	0.974	0.958
review	0.872	0.859	0.866
news	0.894	0.782	0.835
com	0.920	0.874	0.897
shop	0.849	0.950	0.897
instruction	0.866	0.889	0.877
recipe	0.988	0.988	0.988
blog	0.865	0.841	0.853
bio	0.884	0.926	0.905
editorial	0.765	0.926	0.837
faq	0.866	0.879	0.872
edu	0.950	0.969	0.959
story	0.864	0.941	0.901
interview	0.827	0.785	0.805
Insgesamt = 88.98%			

Resultate nach automatischer Nachbarnselektion

back

halbüberwachter min-cut			
Genre	R	P	F
php	**0.928**	**0.850**	**0.887**
forum	0.925	0.977	0.951
review	0.895	0.832	0.862
news	**0.897**	**0.798**	**0.845**
com	0.897	0.891	0.894
shop	**0.860**	**0.965**	**0.910**
instruction	**0.870**	**0.914**	**0.892**
recipe	**0.994**	**0.991**	**0.993**
blog	**0.889**	**0.879**	**0.884**
bio	**0.905**	**0.948**	**0.926**
editorial	**0.800**	**0.932**	**0.861**
faq	0.902	0.841	0.870
edu	**0.957**	**0.963**	**0.960**
story	**0.902**	**0.943**	**0.922**
interview	**0.870**	**0.809**	**0.839**
Insgesamt = **90.11%**			

überwacht			
class	R	P	F
php	0.938	0.798	0.862
forum	0.943	0.974	0.958
review	0.872	0.859	0.866
news	0.894	0.782	0.835
com	0.920	0.874	0.897
shop	0.849	0.950	0.897
instruction	0.866	0.889	0.877
recipe	0.988	0.988	0.988
blog	0.865	0.841	0.853
bio	0.884	0.926	0.905
editorial	0.765	0.926	0.837
faq	0.866	0.879	0.872
edu	0.950	0.969	0.959
story	0.864	0.941	0.901
interview	0.827	0.785	0.805
overall accuracy = 88.98%			