Using Wikipedia Page View Statistics to Measure Issue Salience

Simon Munzert
University of Konstanz

WEBDATANET 2015 | Salamanca

May 27th, 2015

Outline

Motivation

Existing measures of issue salience

- MIP surveys (RePass 1971, Miller et al. 1976)
- media coverage (Baumgartner and Jones 1993, Epstein and Segal 2000)
- Internet search data (Ripberger 2011, Mellon 2014)

Existing measures of issue salience

- MIP surveys (RePass 1971, Miller et al. 1976)
- media coverage (Baumgartner and Jones 1993, Epstein and Segal 2000)
- Internet search data (Ripberger 2011, Mellon 2014)

|	survey data (MIP)	media data	Internet search data
pros	individual-level measure, long historical record, direct measure	low costs, frequency, flexibility, long historical record	low costs, frequency, flexibility, broad coverage
cons	costs, sparseness, limited coverage, validity concerns	aggregate-level, bridging assumptions, media variable 'burnt'	aggregate-level measure, lack of transparency, limited time series

An alternative approach

Strategy

- identify set of appropriate *Wikipedia* articles (subject to validation)
- monitor daily-level page view counts
- derive measure of salience

Rationale

- *Wikipedia* main web information source of choice
- people search for certain topics/issues out of interest/attention
- measure short-term attention and long-term salience towards issues using page view counts of selected *Wikipedia* articles

Example
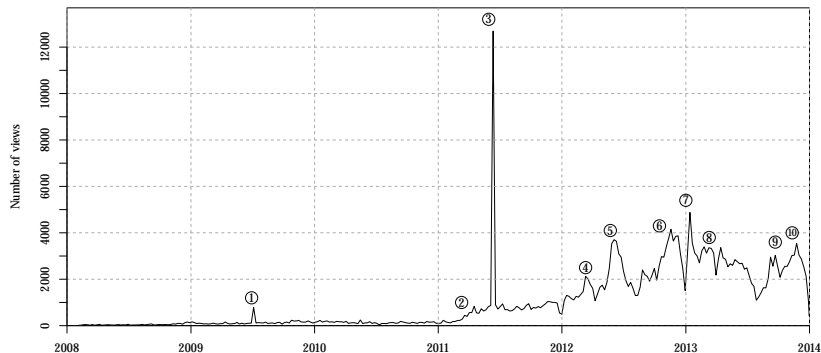
Wikipedia page view counts for the article 'Energiewende'.

Assumptions

1. Aggregate page view data reflect the average importance of issues among *Wikipedia* users.
2. Public issue salience of *Wikipedia* users and the general public is equivalent.
3. *Wikipedia* provides relevant information for the topic of interest.
4. Changes in page view data reflect changes in issue importance.

Validation strategy

- **face and content validity**: selection and investigation of articles
- **criterion validity**: comparison with survey [and Google Trends] data
- **predictive validity** using parliamentary interpellations [to be done]

Data

- MIP data: German survey data (*Politbarometer* study), bi-weekly since 2006, pre-coded categories

- *Wikipedia* data: daily-level raw counts, February 2008 to December 2013, taken from http://stats.grok.se

- focus on the four most popular issues: 'Unemployment', 'Europe/EU', 'Economy', 'Financial crisis'

Selected artices

Issue	Article name	# views	avg. views	max. views
Unemployment	Arbeitslosengeld_II	2,140,971	986	7,973
	Arbeitslosigkeit	1,354,892	624	2,434
	Arbeitslosenversicherung	1,248,687	575	4,938
	Arbeitslosenstatistik	991,693	457	2,515
	Bundesagentur_für_Arbeit	770,609	355	1,206
	Agenda_2,010	717,928	331	8,046
	Sozialhilfe_(Deutschland)	711,950	328	1,090
	Arbeitslosengeld	630,090	290	990
	Hartz_IV	455,756	210	1,598
	Arbeitsmarkt	293,630	135	16,727
Europe/EU	Europäische_Union	11,706,678	5,392	19,922
	Europa	10,650,888	4,906	27,434
	Euro	7,694,512	3,544	41,217
	Europäische_Kommission	1,283,189	591	5,553
	Eurozone	1,219,839	562	6,342
	Europäischer_Rat	992,840	457	5,166
	Europäische_Wirtschafts...	953,526	439	8,692
	Europäischer_Binnenmarkt	551,133	254	2,353
	Eurokrise	102,919	47	1,606

Selected artices

Issue	Article name	# views	avg. views	max. views
Financial crisis	Finanzkrise_ab_2,007	1,270,856	585	4,287
	Basel_II	954,732	440	2,354
	Lehman_Brothers	774,715	357	20,163
	Finanzkrise	540,238	249	5,567
	Investmentbank	470,135	217	8,345
	Bad_Bank	277,068	128	6,029
	Bankenkrise	170,047	78	1,860
	Deutsche_Bankenkrise	77,008	35	3,435
Economy	Bruttoinlandsprodukt	5,035,363	2,319	37,026
	Konjunktur	2,773,927	1,278	14,191
	Wirtschaft	1,594,081	734	4,488
	Wirtschaft_Deutschlands	866,686	399	1,412
	Wirtschaftswachstum	810,545	373	1,390

Seasonality
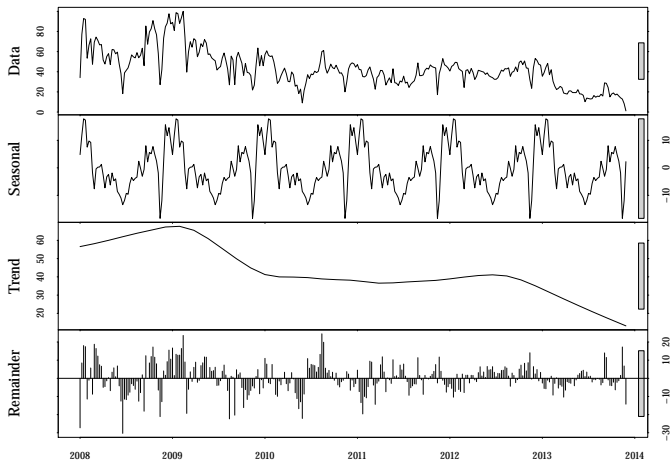
Seasonal trend decomposition by LOESS for views of the German *Wikipedia* article on unemployment benefits.

Criterion validity

Issue	Article	Coefficient	N-W s.e.	R2
	Agenda__2010	0.53	0.31	0.07
	Arbeitslosengeld	0.45	0.10	0.16
	Arbeitslosengeld__II	0.09	0.21	0.00
	Arbeitslosenstatistik	-0.30	0.09	0.09
Unemployment	Arbeitslosenversicherung	0.28	0.11	0.04
	Arbeitslosigkeit	0.40	0.10	0.10
	Arbeitsmarkt	0.78	1.74	0.00
	Bundesagentur__für__Arbeit	-0.15	0.22	0.01
	Hartz__IV	-0.37	0.08	0.18
	Sozialhilfe__Deutschland	-0.80	0.21	0.18
	Euro	-0.43	0.14	0.04
	Eurokrise	0.35	0.15	0.08
	Europa	0.05	0.17	0.00
	Europäische__Kommission	0.30	0.32	0.01
Europe/EU	Europäische__Union	-0.43	0.15	0.04
	Europäische__Wirtschafts__und__Währungsunion	0.41	0.50	0.02
	Europäischer__Binnenmarkt	0.75	0.21	0.11
	Europäischer__Rat	0.55	0.23	0.07
	Eurozone	1.08	0.10	0.63
	Bad__Bank	0.89	0.19	0.14
	Basel__II	0.75	0.22	0.25
Financial crisis	Finanzkrise	2.06	0.27	0.64
	Finanzkrise__ab__2007	0.55	0.27	0.12
	Investmentbank	1.81	0.69	0.16
	Lehman__Brothers	2.28	0.68	0.35
	Bruttoinlandsprodukt	1.50	0.35	0.12
	Konjunktur	0.66	0.21	0.16
Economy	Wirtschaft	0.28	0.14	0.04
	Wirtschaft__Deutschlands	-0.70	0.14	0.23
	Wirtschaftswachstum	0.46	0.17	0.06

Criterion validity II

Issue	Article	Coefficient	N-W s.e.	R2
Unemployment	Agenda__2010	0.52	0.15	
	Arbeitslosengeld	0.25	0.15	
	Arbeitslosengeld__II	0.80	0.20	
	Arbeitslosenstatistik	-0.24	0.13	
	Arbeitslosenversicherung	-0.65	0.20	0.60
	Arbeitslosigkeit	0.61	0.22	
	Arbeitsmarkt	1.34	1.89	
	Bundesagentur__für__Arbeit	0.26	0.29	
	Hartz__IV	-0.10	0.08	
	Sozialhilfe__Deutschland	-1.22	0.27	
Europe/EU	Euro	-0.41	0.20	
	Eurokrise	-0.15	0.09	
	Europa	-0.01	0.14	
	Europäische__Kommission	0.08	0.20	
	Europäische__Union	-0.09	0.21	0.66
	Europäische__Wirtschafts__und__Währungsunion	0.32	0.17	
	Europäischer__Binnenmarkt	-0.04	0.27	
	Europäischer__Rat	-0.10	0.29	
	Eurozone	1.11	0.14	
Financial crisis	Bad__Bank	0.27	0.10	
	Basel__II	0.03	0.09	
	Finanzkrise	1.81	0.24	0.75
	Finanzkrise__ab__2007	0.36	0.11	
	Investmentbank	0.13	0.31	
	Lehman__Brothers	0.25	0.24	
Economy	Bruttoinlandsprodukt	1.29	0.45	
	Konjunktur	0.71	0.30	
	Wirtschaft	-0.09	0.18	0.52
	Wirtschaft__Deutschlands	-0.85	0.13	
	Wirtschaftswachstum	-0.19	0.36	

Conclusion

- proof of concept: *Wikipedia* article views can serve as an aggregate indicator of public issue attention (== salience?) if combined

- promising measure for (a) single-issue analyses and (b) testing hypotheses on short-term attention (McCombs and Zhu 1995)

Conclusion

- proof of concept: *Wikipedia* article views can serve as an
  aggregate indicator of public issue attention (== salience?) if
  combined

- promising measure for (a) single-issue analyses and (b) testing
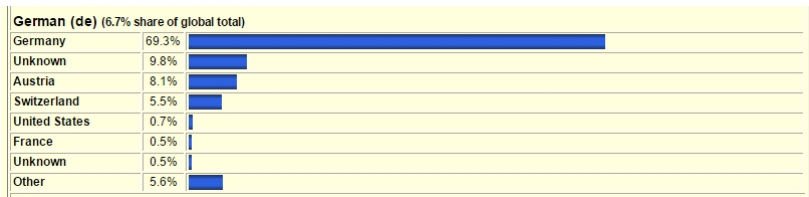 hypotheses on short-term attention (McCombs and Zhu 1995)

To-do list

- ☐ extend test on criterion validity with external criterion
- ☐ assess reliability with out-of-sample testing
- ☐ expand set of issues
- ☐ establish more efficient combination procedure
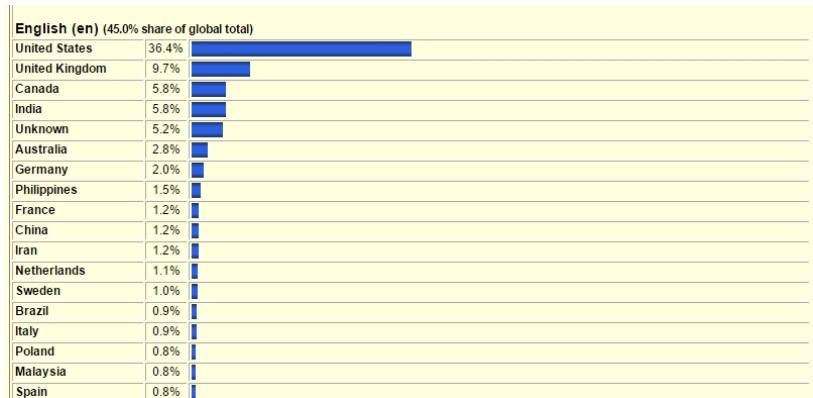- ☐ investigate other Wikis

Thank you for your attention!

Simon Munzert
simon.munzert@uni.kn
@simonsaysnothin

User origin, German Wikipedia

User origin, English Wikipedia

English (en) (45.0% share of global total)

United States	36.4%	
United Kingdom	9.7%	
Canada	5.8%	
India	5.8%	
Unknown	5.2%	
Australia	2.8%	
Germany	2.0%	
Philippines	1.5%	
France	1.2%	
China	1.2%	
Iran	1.2%	
Netherlands	1.1%	
Sweden	1.0%	
Brazil	0.9%	
Italy	0.9%	
Poland	0.8%	
Malaysia	0.8%	
Spain	0.8%	

Weekly seasonality

Wikipedia page view counts (red curve) time series for the 'Energiewende' article in 2013. The time series exhibits considerable seasonality because page view statistics use to drop at weekends.
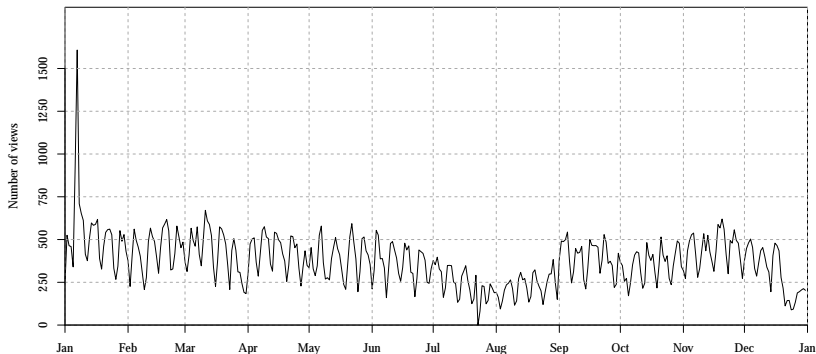
Top 25 MIP mentions

Category labels, German	Category labels, English	# mentions	% mentions
Arbeitslosigkeit	Unemployment	23,847	22.1
EG, Europa, Binnenmarkt	EU, Europe, internal market	11,422	10.6
internationale Bankenkrise	International banking crisis	9,442	8.7
Wirtschaftslage	Economy	8,689	8.0
Politische Moral	Political morality	5,924	5.5
soziales Gefälle	Social inequality	5,732	5.3
sonstige innenpolitische Dinge	Other domestic issues	4,976	4.6
Inflation, Preise, Zinsen	Inflation, prices, interest rates	3,982	3.7
Renten und Alte	Pensions and the elderly	3,947	3.7
Staatsverschuldung	State debt	3,843	3.6
Ausländer	Foreigners	3,052	2.8
Kindergartenplätze	Child care	2,761	2.6
Löhne und Einkommen	Wages and incomes	2,395	2.2
Umweltschutz	Environment	2,372	2.2
Kernkraftwerke	Nuclear power	2,275	2.1
Sonstiges, allgemein	Other, general	2,171	2.0
Steuern, Steuerreform	Taxes, tax reform	2,125	2.0
Gesundheitswesen	Healthcare	2,061	1.9
Sozialpolitik	Social policy	1,566	1.4
Streit in der Regierung/ Koalition	Government crisis	1,277	1.2
Moral, Werte	Moral issues	1,270	1.2
Energiewende/erneuerbare Energien	Energy transition	817	0.8
Energieversorgung, -problem	Energy supply	796	0.7
Innere Sicherheit	Inner security	778	0.7
Pessimismus, Jammern	Pessimism	587	0.5