A Classification of Search Terms

Nadine Schmidt-Mänz

Contents:

1. Analyses of Search Queries

Study	E1	E2	F	AV1	E3	W	B	A1	E4	A2	AV2
Group	Jansen	Spink	Hölscher	Silverstein	Spink	Zien	Cacheda	Spink	Spink	Spink	Spink
Observation Period	03/97	09/97	07/98	08/98	12/99	03/00	05/00	02/01	04/01	05/02	09/02
Length (Days)	1	1	31	43	1	66	16	1	1	1	1
Database	Logs	Logs	Logs	Logs	Logs	Ticker	Logs	Logs	Logs	Logs	Logs
Sessions	18.098	211.063	–	285.474.117	325.711	–	57.259	153.297	262.025	345.093	369.350
SQ in total	51.473	1.025.908	16.252.902	993.208.159	1.028.910	50.538.653	105.786	451.551	1.025.910	957.303	1.073.388
Unique SQ	–	531.416	–	153.645.050	–	–	35.518	–	–	–	–
Occurrence SQ (Ø)	–	1,9	–	6,5	–	–	3,0	–	–	–	–
ST in total	113.793	2.216.986	–	–	1.500.500	165.763.490	173.128	1.350.619	1.538.120	2.225.141	3.132.106
Unique ST	21.682	140.279	–	–	–	–	23.707	180.998	–	340.711	297.528
Occurrence ST (Ø)	2,2	15,8	–	–	–	–	7,3	7,4	–	7,5	10,5
Length SQ (Ø)	2,2	2,2	1,7	2,4	2,4	3,3	1,6	2,4	2,6	2,3	2,9
1-Term SQ (%)	–	26,6%	–	25,8%	29,8%	22,5%	–	25,0%	29,6%	33,0%	20,4%
SQ not repeated (%)	–	–	–	–	–	–	20,3	–	–	–	–
ST not repeated (%)	8,6%	6,8%	–	–	7,3%	–	6,3%	7,0%	7,4%	10,0%	5,6%
Complex SQ (%)	15,9%	9,3%	2,6%	20,4%	10,9%	35,6%	8,6%	4,3%	11,3%	4,6%	27,3%
Phrase search (%)	6,0%	5,1%	–	–	5,9%	10,4%	3,6%	0,0%	5,9%	0,0%	12,1%
Natural SQ (%)	–	–	–	–	1,0%	17,9%	–	–	0,28%	–	–
Search area (%)	0,1%	9,7%	–	–	–	–	0,2%	–	–	–	–
Top SQ	–	–	–	+	–	–	+	+	–	+	–
Top ST	–	+	–	–	+	–	+	+	+	+	+
First result page (%)	58,0%	66,3%	–	85,2%	69,9%	–	67,9%	54,1%	84,6%	76,3%	72,8%
Only one SQ per session	67,0%	48,5%	–	77,6%	60,4%	–	–	53,0%	55,4%	–	47,6%
SQ per session	2,8	2,3	–	2,0	1,9	–	–	2,9	2,3	2,8	2,9

1. Analyses of Search Queries

Study	FB	LY	MG	MS
Observatiton period	08/04-09/05	08/04-09/05	11/04-09/05	09/04-09/05
Length (days)	399	403	314	358
Database	Ticker	Ticker	Top 4000	Ticker
Sessions	-	-	-	-
SQ in total	132.833.007	189.930.859	4.407.566	4.089.731
Unique SQ	17.992.069	29.322.366	678.655	1.287.417
Occurrence SQ (Ø)	7,4	6,5	6,5	6,2
ST in total	241.833.877	344.242.099	7.333.343	7.853.501
Unique ST	6.296.833	11.232.710	430.338	627.507
Occurrence ST (Ø)	29,4	30,6	17,0	12,5
Length SQ (Ø)	1,8	1,7	1,6	1,8
1-term SQ (%)	50,1%	51,9%	58,7%	48,4%
SQ not repeated (%)	7,9% (58,3%)	9,3% (60,1%)	0,2%(1,0%)	17,9%(56,9%)
ST not repeated (%)	1,3% (49,1%)	1,8% (53,9%)	0,0% (0,7%)	3,4% (43,0%)
Complex SQ (%)	$< 3\%$	$< 3\%$	$< 3\%$	$< 3\%$
Phrase search (%)	2,1%	2,4%	–	2,5%
Natural SQ (%)	0,1%	0,1%	0,1%	0,2%
Search area (%)	65,8%	–	–	87,9%
Top SQ	+	+	+	+
Top ST	+	+	+	+
Only first result page (%)	–	–	–	–
Only one SQ per session	–	–	–	–
SQ per session	–	–	–	–

2. A General Classification of Search Terms

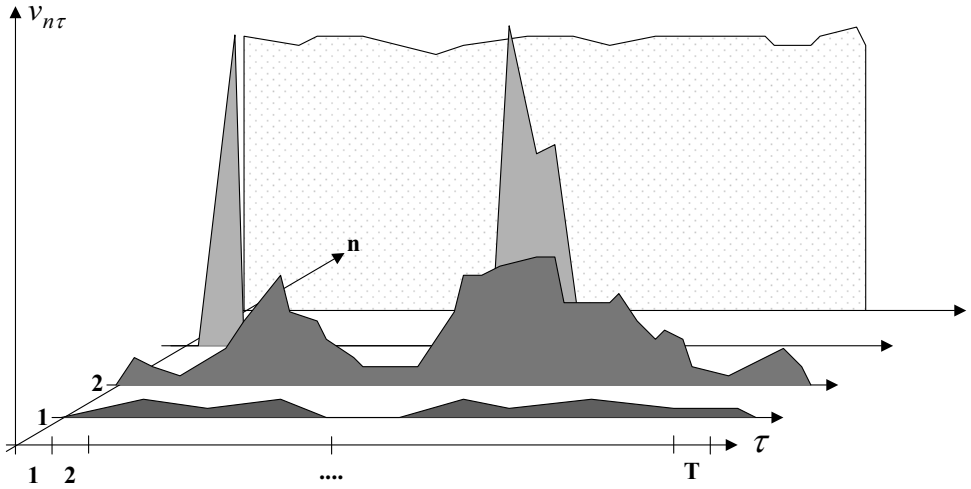

Figure 1: Examples for different v_{nt}

$\mathcal{V}$	1	...	τ	...	T	$\vec{V}^{(T)}$
1	v_{11}	...	$v_{1\tau}$	...	v_{1T}	$v_1^{(T)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	v_{n1}	...	$v_{n\tau}$	...	v_{nT}	$v_n^{(T)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N_T	v_{N_T1}	...	$v_{N_T\tau}$	...	v_{N_TT}	$v_{N_T}^{(T)}$
	$\|\|\vec{V}(1)\|\|_1$	...	$\|\|\vec{V}(\tau)\|\|_1$	...	$\|\|\vec{V}(T)\|\|_1$	$\Gamma_V^{(T)}$

Table 1: Matrix $\mathcal{V}_{N_T \times T}$

- Only search terms which occur more often than threshold f are frequent ($v_{n\tau} > f$).

- For the elements of matrix $\mathcal{A}^f_{N_T \times T}$ following is defined:

$$a^f_{n\tau} = \begin{cases} 1, & v_{n\tau} \geq f \\ 0, & else. \end{cases}$$

2. A General Classification of Search Terms

Definitions:

- **f**: Threshold for frequent terms, lower bound to cut off "white noise".
- **Mayflies**: Terms which occur frequently enough in one particular time interval ($v_{n\tau} > f$).
- **Evergreens**: Terms which occur frequently enough in nearly every time interval, e.g., in 90% of all time intervals.
- **Impulses**: Impulses occur after a desaster or very interesting news stories.
- **Events**: Some terms reoccur in a periodical manner or have a known arrival date (christmas).
- **Upper Bound**: Is defined by Evergreens.

2. A General Classification of Search Terms

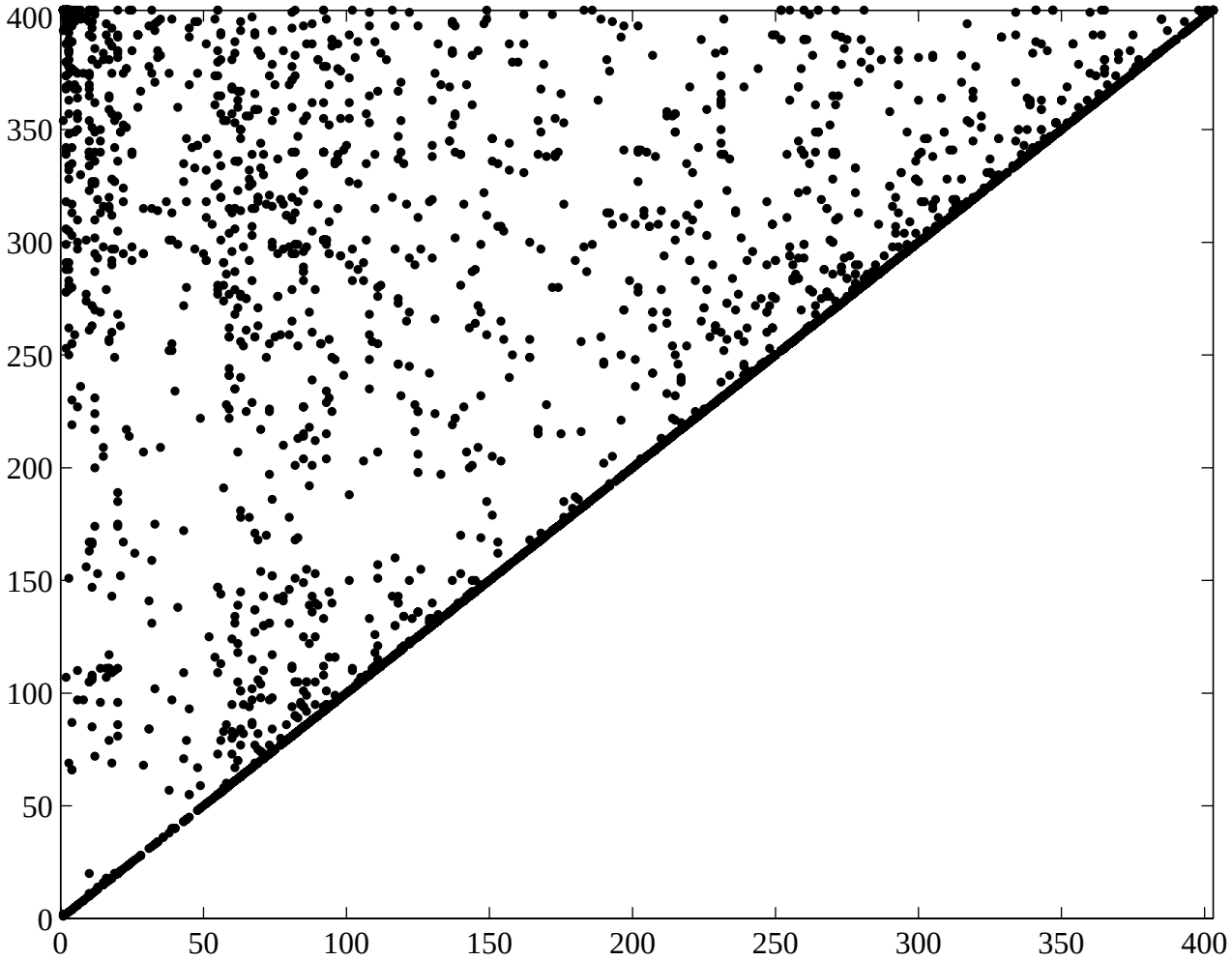

Figure 2: Distribution of terms: first vs last occurence

3. Patterns in Online Searching Behavior

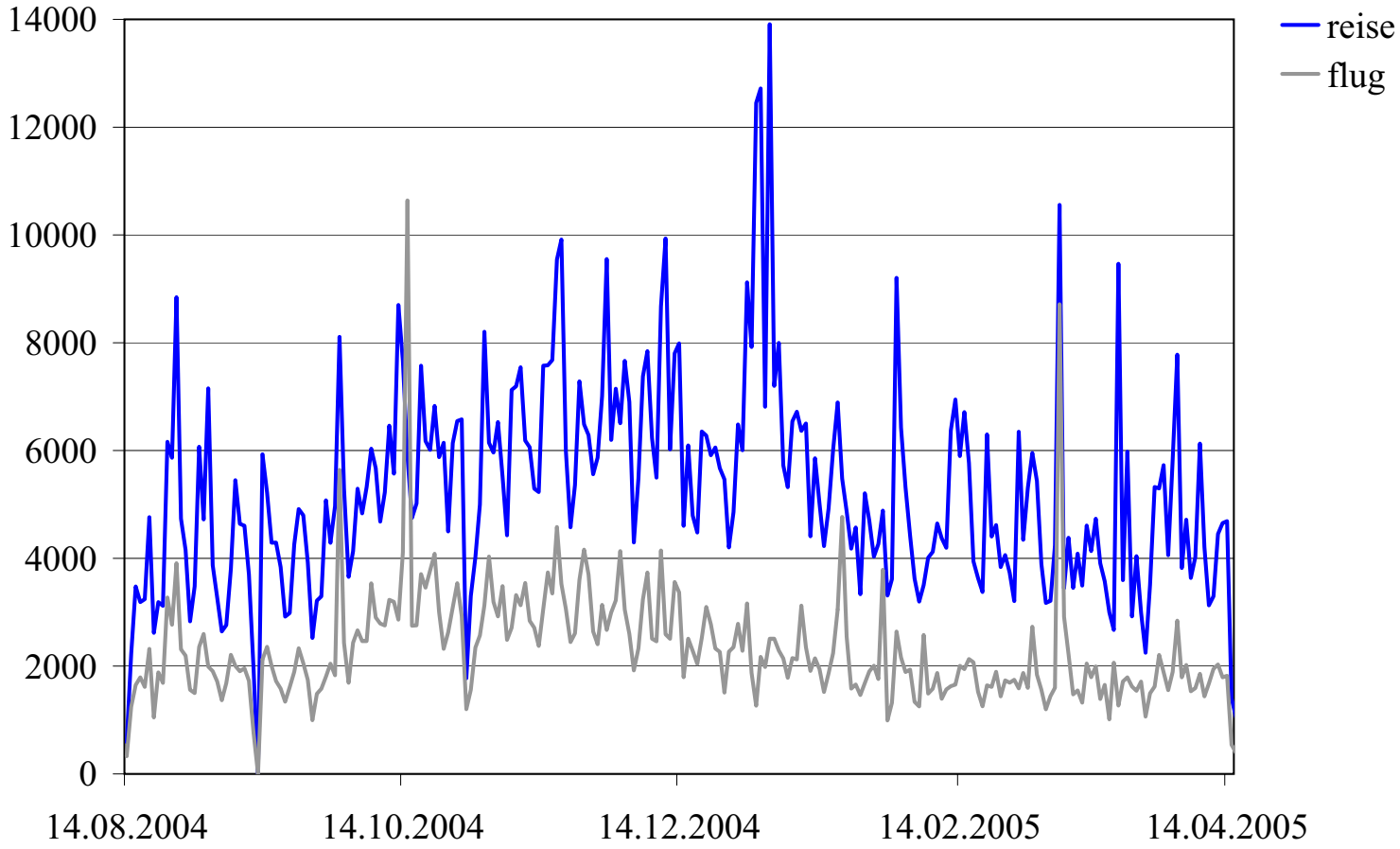

Figure 3: Evergreens

3.Patterns in Online Searching Behavior

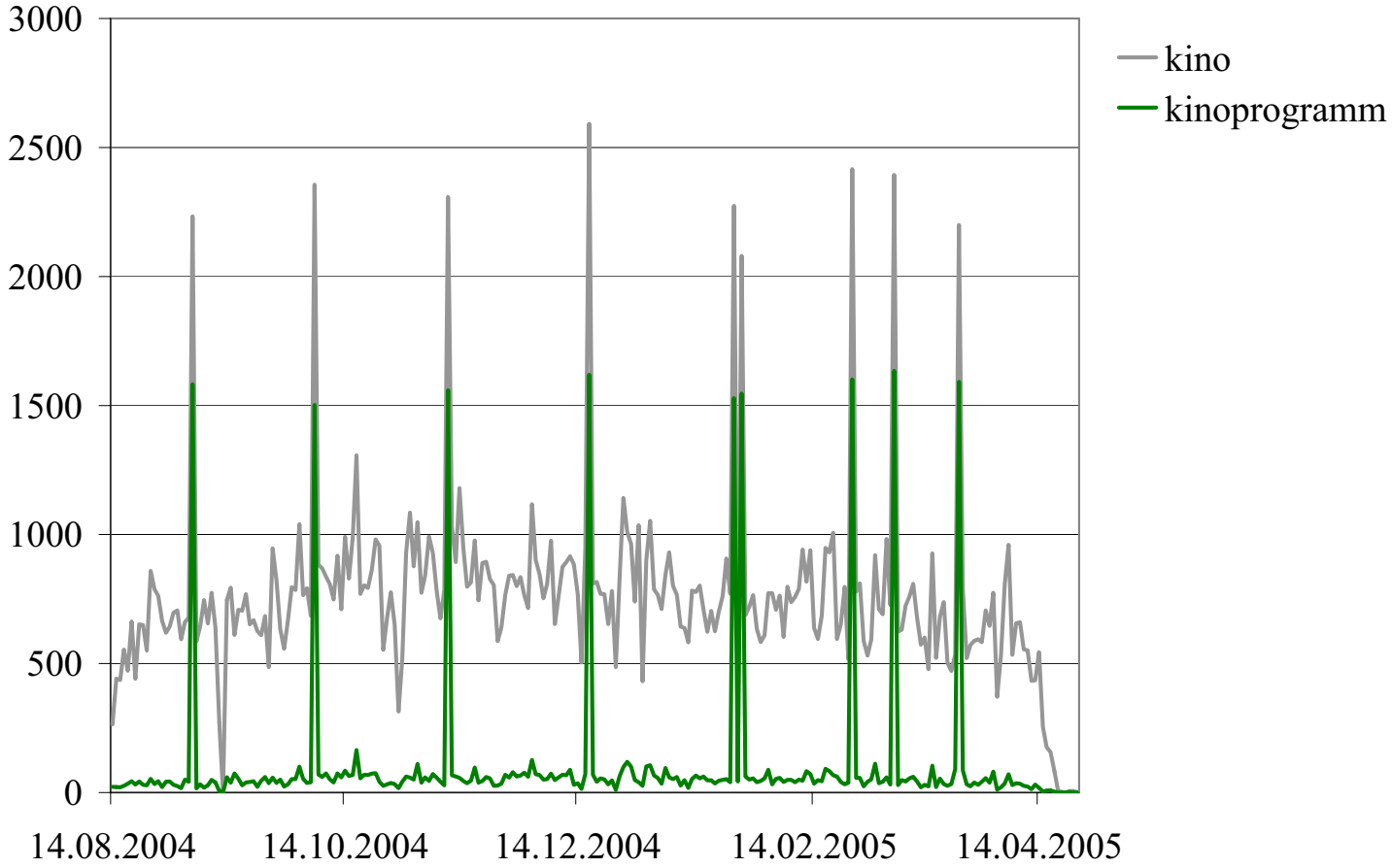

Figure 4: Event I

3. Patterns in Online Searching Behavior

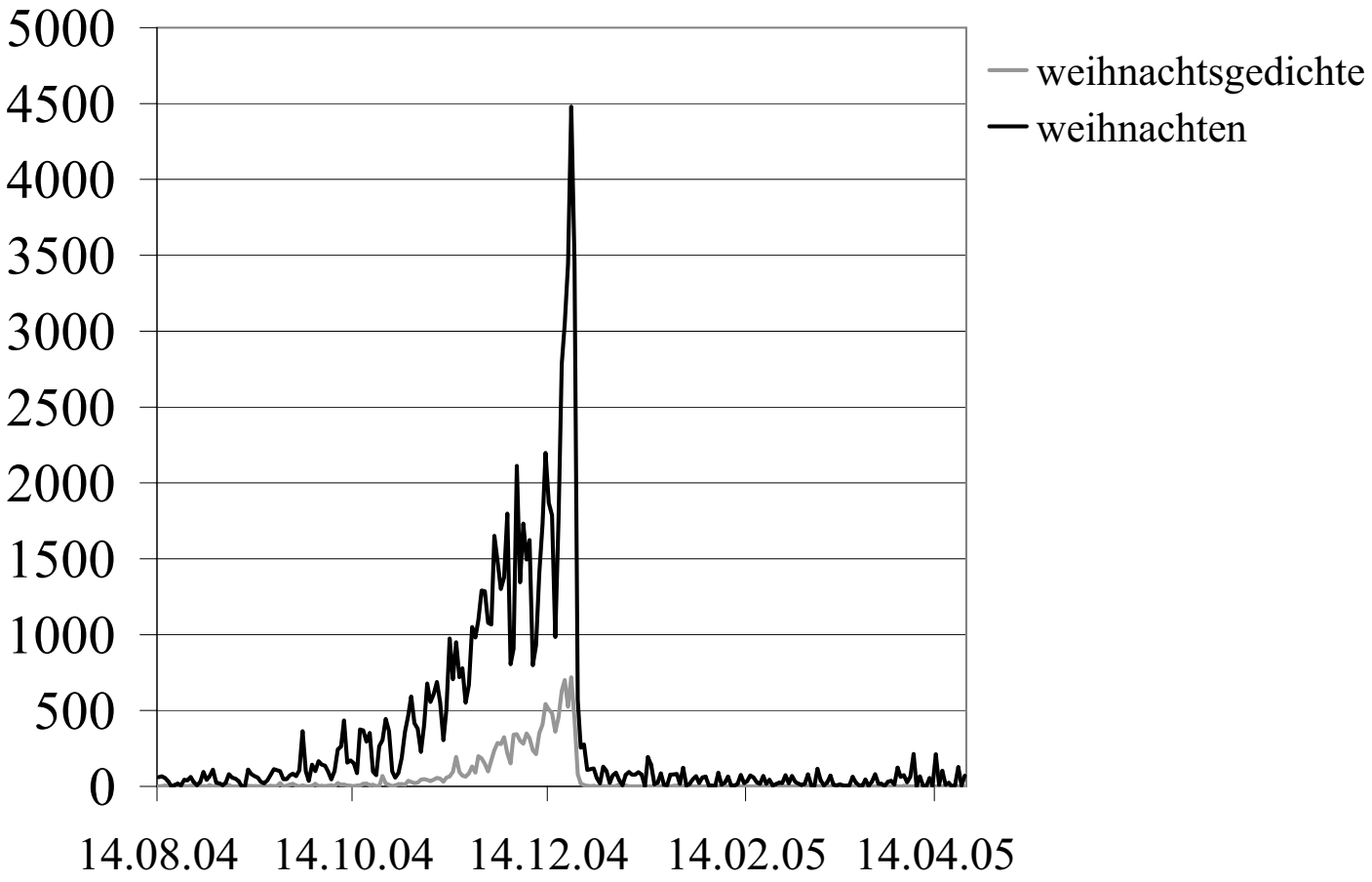

Figure 5: Event II

3. Patterns in Online Searching Behavior

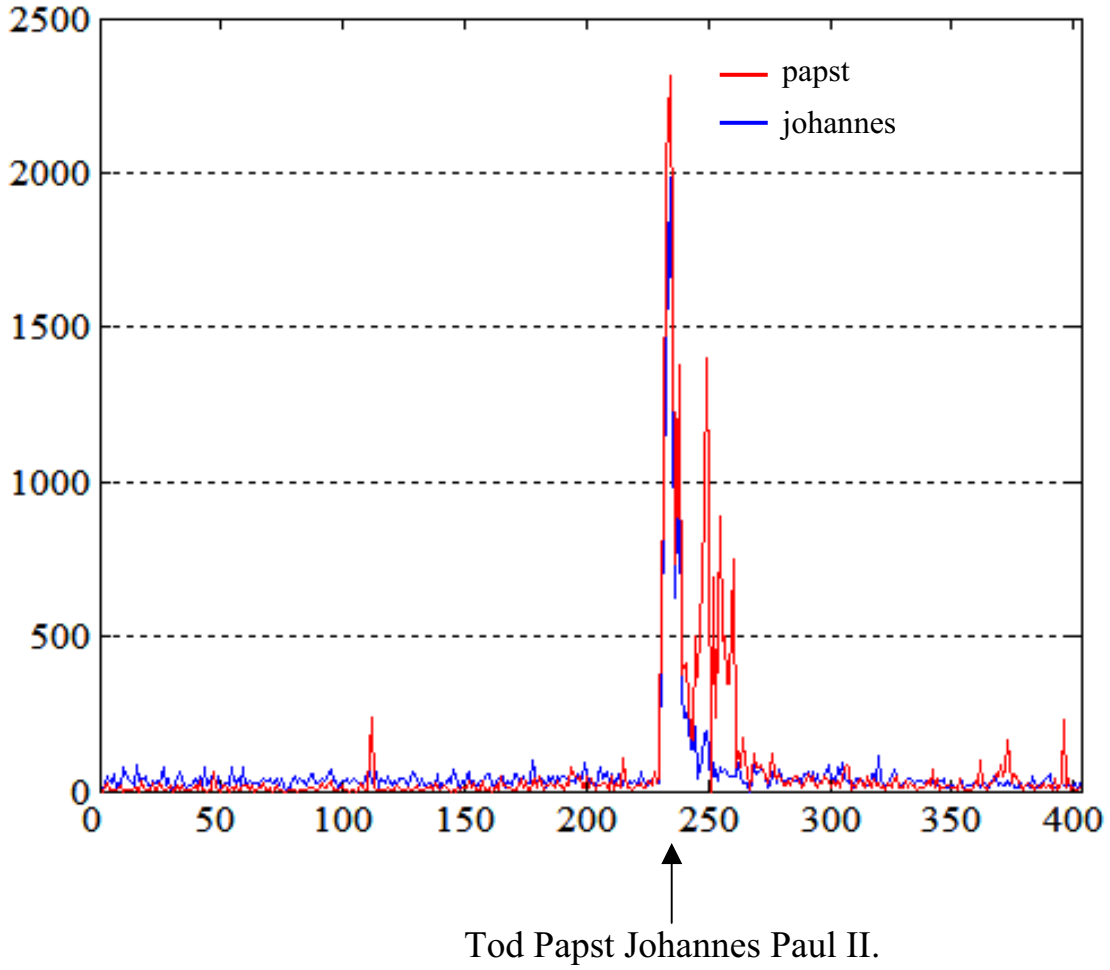

Figure 6: Impulse II

4. Clustering Search Terms

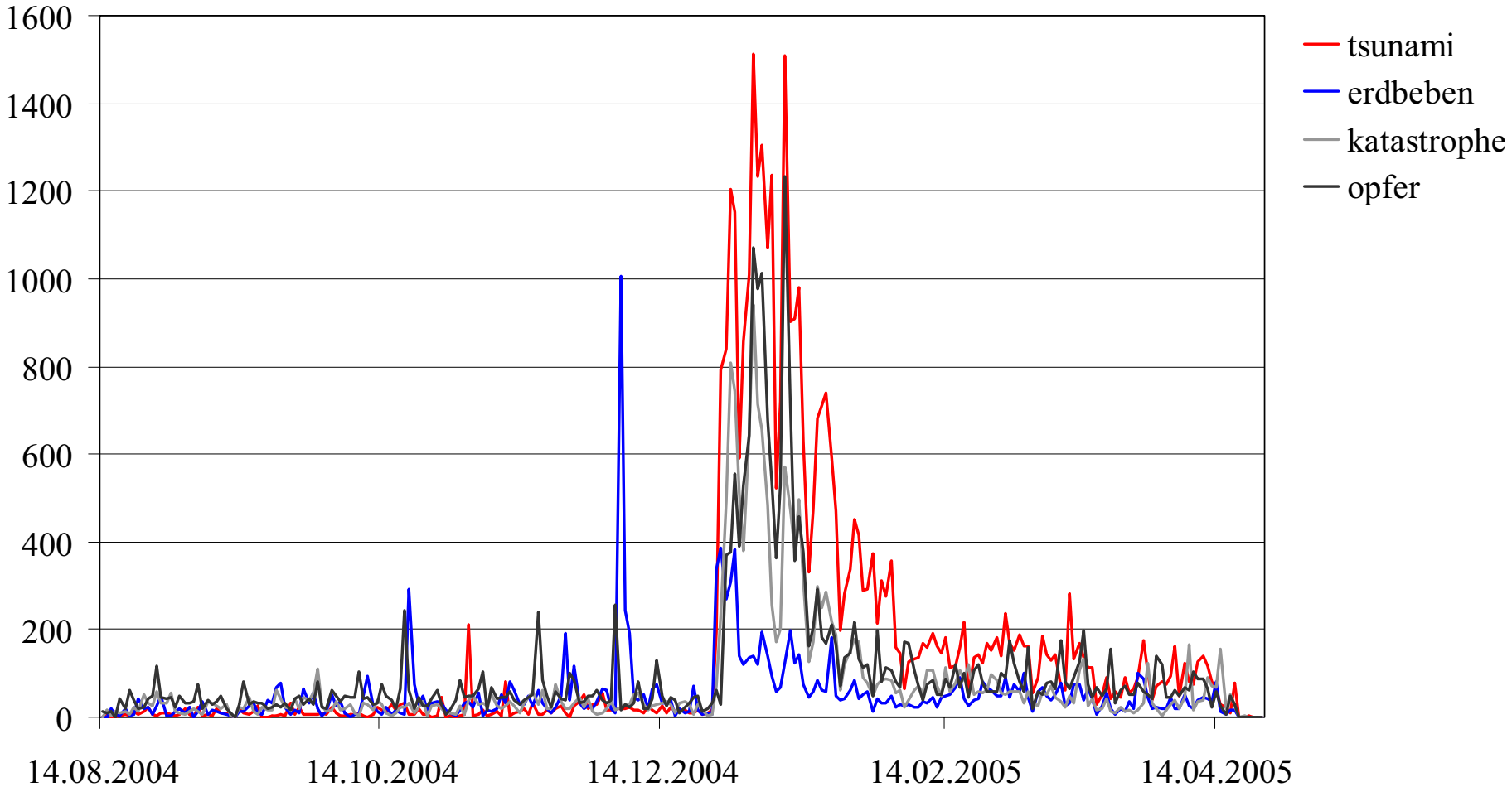

Figure 7: Impulse I

4. Clustering Search Terms

Identifier	Formula
Tf	$w_{n\tau} = v_{n\tau}$
01	$w_{n\tau} = a^f_{n\tau}$
Sum	$w_{n\tau} = v^f_{n\tau} / v^{(T)}_n$
Max	$w_{n\tau} = v^f_{n\tau} / \max_\tau\{v_{n\tau}\}$
MaxLogIdf	$w_{n\tau} = v^f_{n\tau} * \log_{10}(\frac{N_T}{b^f_\tau}) / \max_\tau\{v_{n\tau}\}$

Table 2: Calculations of weighted row vectors $\vec{W}^f_n$

Bezeichnung	Formel
Cosinus	$sim_{nn'} = \frac{\vec{W}^f_n * \vec{W'}^f_{n'}}{\|\|\vec{W}^f_n\|\|_2 * \|\|\vec{W}^f_{n'}\|\|_2}$
Jaccard	$sim_{nn'} = \frac{\vec{W}^f_n * \vec{W'}^f_{n'}}{\|\|\vec{W}^f_n\|\|_2 * \|\|\vec{W}^f_{n'}\|\|_2 - \vec{W}^f_n * \vec{W'}^f_{n'}}$
Euklid	$sim_{nn'} = \frac{1}{\|\|\vec{W}^f_n - \vec{W}^f_{n'}\|\|_2 + 1}$

Table 3: Similarity measures to calculate the similarity matrix $\mathcal{S}$

- l_p-Norm ($||\mathbf{x}||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$) für $p \in \mathbb{N}$
- $\vec{W}^f_n = (w^f_{n1}, \ldots, w^f_{n\tau}, \ldots, w^f_{nT})$ und $sim_{nn'} \in [0; 1]$

4. Clustering Search Terms

1. Determine Mayflies and Evergreens.
2. Calculate the similarity matrix $\mathcal{S}$ for all terms which are not Mayflies or Evergreens.
3. Every term n represents a cluster n.
4. Add to every cluster n the K nearest neighbor terms n' $(n \neq n')$
   with $sim_{nn'} > \varpi$, and
   all terms with the same similarity of the last added term n'.
5. Delete all clusters with only one element, all redundant clusters and all clusters which are a subset of any other cluster.

4. Clustering Search Terms

Cluster with tf/cosinus-measure for $f = 400,\ K = 40,\ \varpi = 0,8$, (standard deviation):

andreas athen olympia ulrich (175184)

archaik hellas homer ilias klassik odyssee sparta (87,8)

eintragung keywords linkpopularität meta optimieren optimierung platzierung plazierung position positionierung

fahne flagge (500,8)

fotos last minute reisen urlaub (193789,2)

goethe johann wolfgang (5710)

professionelle ranking suchbegriff suchbegriffe suchdienste suchmaschienen suchmaschinen suchmaschinenanmeldung suchmaschinenanmeldungen suchmaschineneintrag suchmaschinenoptimierung webseiten (252,3)

handwerker heimwerker heizung sanitär solar (1651,7)

krankenkasse krankenversicherung krankenversicherungen leistungsvergleich privatkrankenkassen privatkrankenversicherung privatkrankenversicherungen tarifvergleich trarifvergleich vergleiche (2963,0)

leistungsvergleich privatkrankenkasse privatkrankenkassen privatkrankenversicherung tarifvergleich testvergleich trarifvergleich (10,7)

silvester sylvester (397,8)

suchmaschinen suchmaschinenanmeldung suchmaschineneintrag suchmaschinenoptimierung webseiten webseitenoptimierung (3923,2)

weihnacht weihnachten weihnachtsbaum weihnachtsbilder weihnachtsgedichte weihnachtsgrüße weihnachtskarten weihnachtsmann (545,9)

5. Conclusions

- General classification of terms:
  - Simplified navigation in search engines
  - Categories in portals
  - Hot spots
- Caching strategies with Evergreens
- Recommendations of time-similar terms
- Events to predict occurrences of terms and co-occurrences of time-similar terms
- Optimization of Adword-booking
- Influence of news on online searching behavior
- Further research?

Contact:

Nadine Schmidt-Mänz

Institut für

Entscheidungstheorie und Unternehmensforschung

Universität Karlsruhe (TH)

home: marketing.wiwi.uni-karlsruhe.de

mail: nadine.maenz@wiwi.uni-karlsruhe.de

cell: +49 176 600 17 0 15

phone: +49 721 608 4770

skype: n.maenz