

# Exploratory analysis needs theor[y|ies] – OR: Some answers to 14 questions



**Bettina  
Berendt**

K.U. Leuven,  
Belgium

[www.berendt.de](http://www.berendt.de)

## Agenda

2

### Research issues I

A brief view of our work on QLA

Research issues I  
General questions

Feasibility I  
Within academia

Feasibility II  
Transfer/collaboration academia – industry

### About me

Prof. Dr. Bettina Berendt

### Outcome:

### Background:

- \* Economics & Business Sci.
- \* Artif. Intellig., Inf. Systems
- \* Cognitive Sci.
- \* Computer Uses in Education

### Selected publications and talks

#### Overview of Research Areas

Note: Many of these areas overlap - for example, Web mining methods are used to understand and support authors in Digital Libraries, or visualization is employed to show search behaviour. So look into several categories to find the paper you're looking for!

- [Digital Libraries, Social Media, Blog Mining, Learner and Author Support / Knowledge Management](#)
- [Information Search and Ubiquitous Information](#)
- [Semantic Web Mining, Ontologies and Knowledge Discovery](#)
- [Web Usage Mining, Query Mining](#)
- [e-Commerce, Web Metrics, Evaluation of Information Systems](#)
- [Personalization and Privacy](#)
- [Information Visualization, Learning and Thinking with Pictorial Representations](#)
- [Distance Cognition](#)
- [Spatial Mental Models / Mental Images](#)
- [Complete publication list \(PDF\)](#)

Fertig

4

### Research issues I

A brief view of our work on QLA

## Questions we have asked

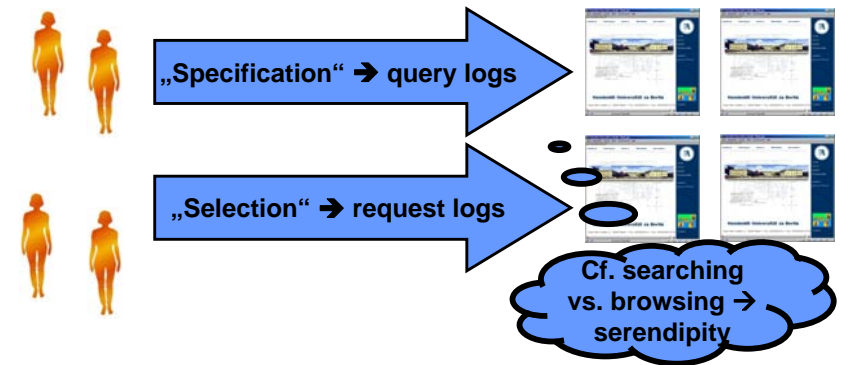
5

- How can we measure (+ improve) interface options' popularity, (search) effectiveness and (search) efficiency?
- How does cognitive style influence search strategy and success?
- Can interactive anatomical software help physiotherapists acquire functional knowledge?
- (How) do interfaces support processes known from the offline world?
- Should e-Commerce Web shops have offline stores too?
- People say they are worried about their privacy – but they act accordingly?
- How does language and culture influence Web usage?
- Are there systematic factors that put non-English native speakers at a disadvantage on the Web?
- [some remarks on description – prediction]

What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

6

→ (1) A fundamental choice on integrated data analysis

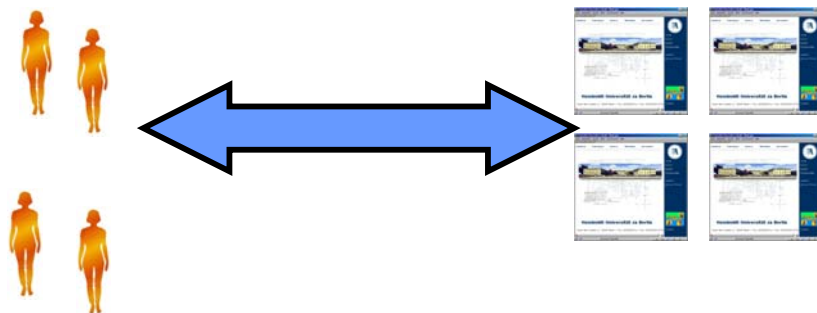


(joint work with E. Brenstein, A. Kralisch, B. Mobasher, S. Spiekermann, M. Spiliopoulou, F. Ritter, M. Teltzrow and other)

What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

7

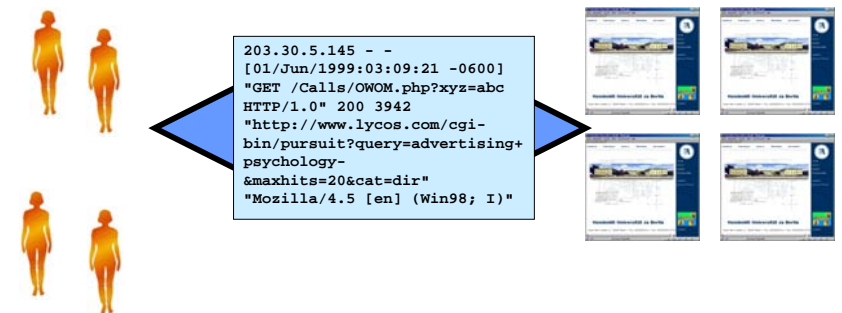
→ (1) A fundamental choice on integrated data analysis



(joint work with E. Brenstein, A. Kralisch, B. Mobasher, S. Spiekermann, M. Spiliopoulou, F. Ritter, M. Teltzrow and other)

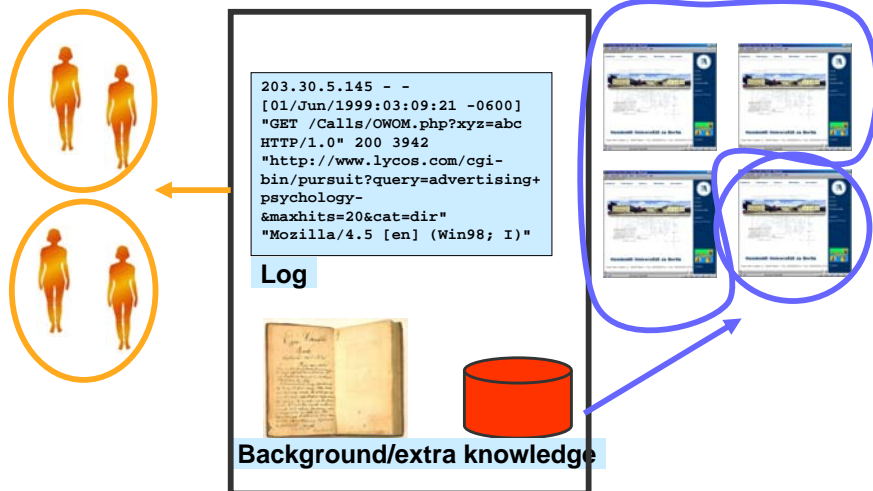
What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

8



What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

9



What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

10

What?

- \* Domain content ontology
- \* Domain behaviour ontology
- \* Interaction style ontology

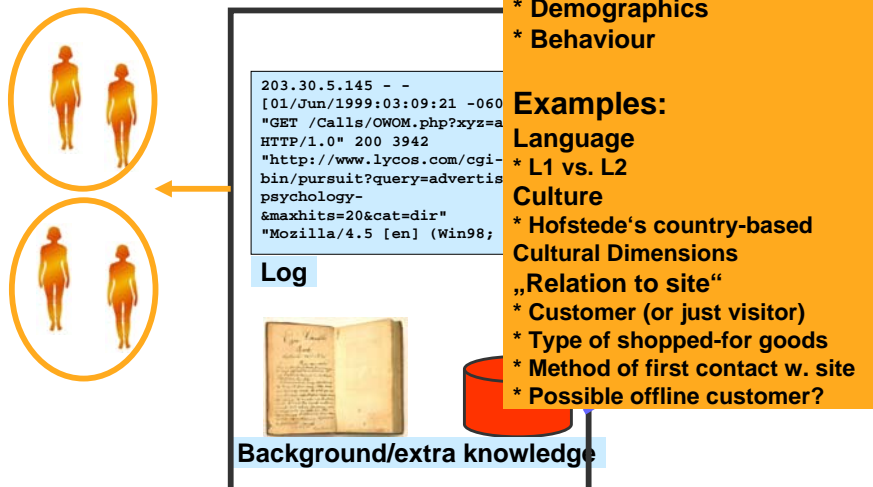
Examples:

- Educational portals / search
- \* Requested (search) services
- ETD portal
- \* Target group
- Medical information site
- \* Search service: modality
- \* ICD-9 classification of diseases
- E-commerce
- \* Stage in buying process

**Background/extra knowledge**

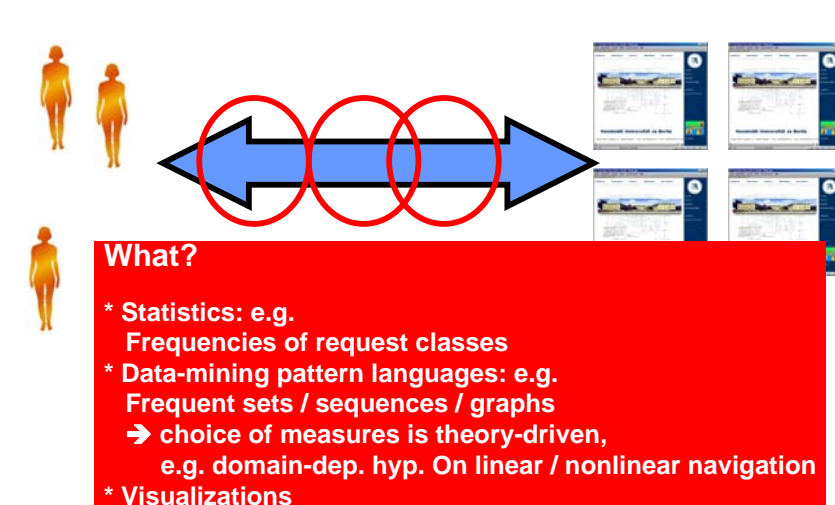
What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

11



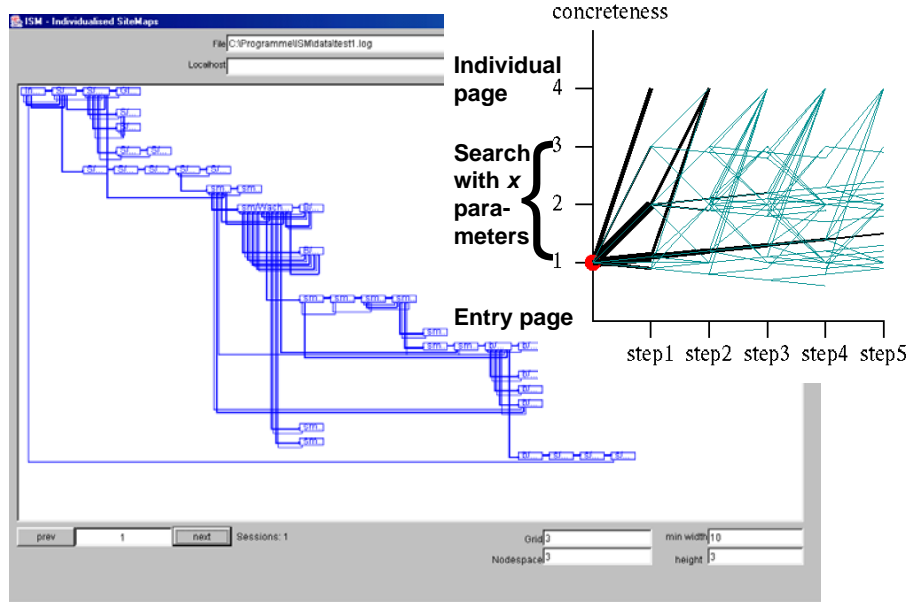
What approaches to log analysis are used in different fields?  
Which techniques are similar between / specific to fields/applications?

12



## Frequent patterns – usable across fields

13

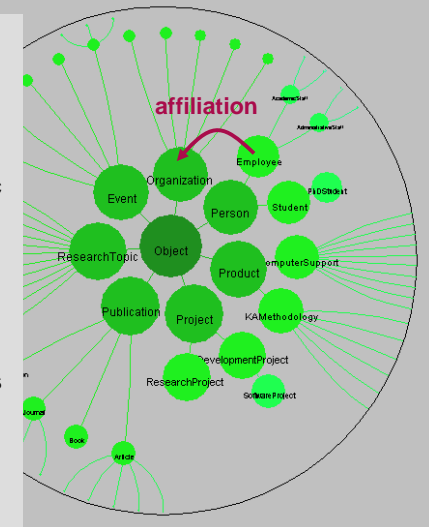


## Semantics of requests

14

### Step 1: Domain ontology

- community portal [ka2portal.aifb.uni-karlsruhe.de](http://ka2portal.aifb.uni-karlsruhe.de)
- ontology-based:
  - Knowledge base in F-Logic
  - Static pages: annotations
  - Dynamic pages: generated from queries
  - Queries also in F-Logic
  - Logs contain these queries



[Oberle, Berendt, Hotho, & Gonzalez, Proc. AWIC 2003]

## Semantics of requests

15

### Step 2: Modelling requests and sessions-as-sets

RESEARCHER  
PERSON  
PROJECT  
PUBLICATION  
RESEARCHTOPIC  
EVENT  
ORGANIZATION  
RESEARCHINTEREST  
LASTNAME  
TITLE  
ISABOUT  
EVENTS  
EVENTTITLE  
WORKSATPROJECT  
AUTHOR  
AFFILIATION  
ISWORKEDONBY  
PROGRAMCOMMITTEE  
EMPLOYS  
NAME  
RESEARCHGROUPS  
EMAIL

An example query with **concepts** and **relations**:

```
FORALL N, PEOPLE <-PEOPLE:
Employee[affiliation->> "http://www.anInstitute.org"]
and PEOPLE:Person[lastName->>N].
```

Query =  
feature vector of concepts + relations



Session =  
feature vector of concepts + relations,  
summed over all queries in the session

Clustering,  
Association rules,  
Classification, ...

## 2. Semantics of sequences

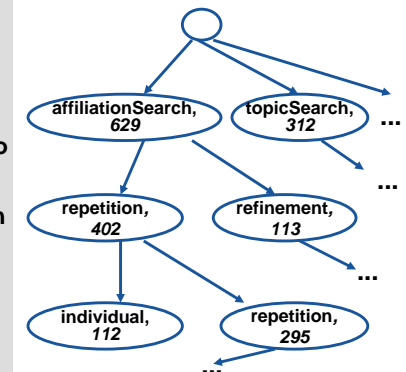
16

### Step 3: Strategy pattern discovery

An ontology of navigation strategies

- Define strategy templates as regular expressions
  - Of requests (mapped to ontological entities)
  - Of transitions (between ontological entities)

Ex. [.search .\* individual]
- Discover strategies by learning a strategy trie



[Berendt & Spiliopoulou, VLDB Journal, 2000]

[Berendt, Data Mining and Knowledge Discovery, 2002]

## Semantics of sequences

### Step 4: Strategy pattern evaluation

17

#### Use strategy patterns' statistics to

- Derive descriptive measures of patterns
  - support, confidence
  - popularity, effectiveness, efficiency
- Apply inferential statistics to compare patterns

Table 4. Further actions in the analyzed patterns

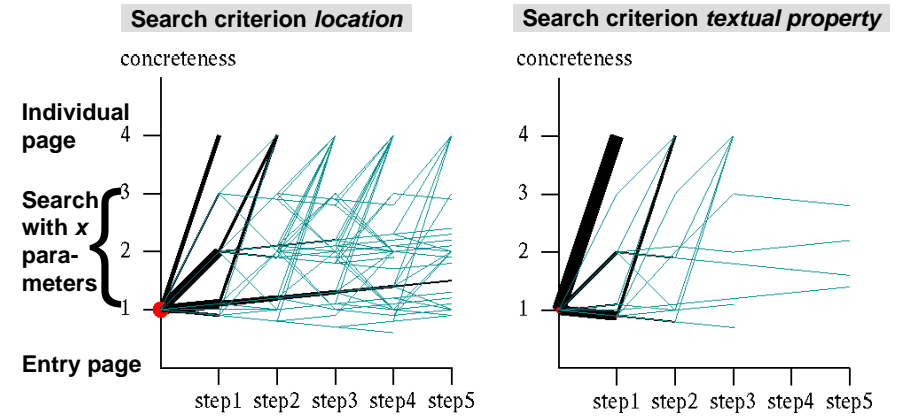
K	Goal found (2 steps)		Search continued		Search refined		New search	
	URL	no. of sequences	URL	no. of sequences	URL	no. of sequences	URL	no. of sequences
D	2.83% of 1907							
D1	29	12.66% of 158	GRITEL-48	30.38% of 158	GRITEL-13	8.23% of 158	11000-	25
			LALL		LALL		GRITEL	15.82% of 158
					GRITEL	5	GRITEL	4
					GRITEL	3.16% of 158	LALL	2.53% of 158
I	0.79% of 1907							
	I1	11	GRITEL	55	GRITEL	15	11000-	20
		6.50% of 167	LALL	32.93% of 167	LALL	8.98% of 167	GRITEL	17.37% of 167
	no next step:				GRITEL	5	GRITEL	3
					GRITEL	2.90% of 167	LALL	4.19% of 167
A0	goal found		GRITEL	152	GRITEL	36	11000-	30
	no next step:		LALL	26.23% of 579	LALL	6.22% of 579	GRITEL	12.63% of 579
	5.33% of 579							
	goal found		GRITEL	102	GRITEL	33	11000-	63
no next step:		LALL	17.62% of 579	LALL	5.33% of 579	GRITEL	10.94% of 579	
0.21	no-1300008		GRITEL	5	GRITEL	1	GRITEL	1
1.67% of 579			GRITEL	0.60% of 579	LALL	1.21% of 579		
				GRITEL	3	GRITEL	1	
				GRITEL	0.64% of 579		17.74% of 579	

[Berendt, *Data Mining and Knowledge Discovery*, 2002]

## Communication – Visual data mining

### Step 5 – Example

18

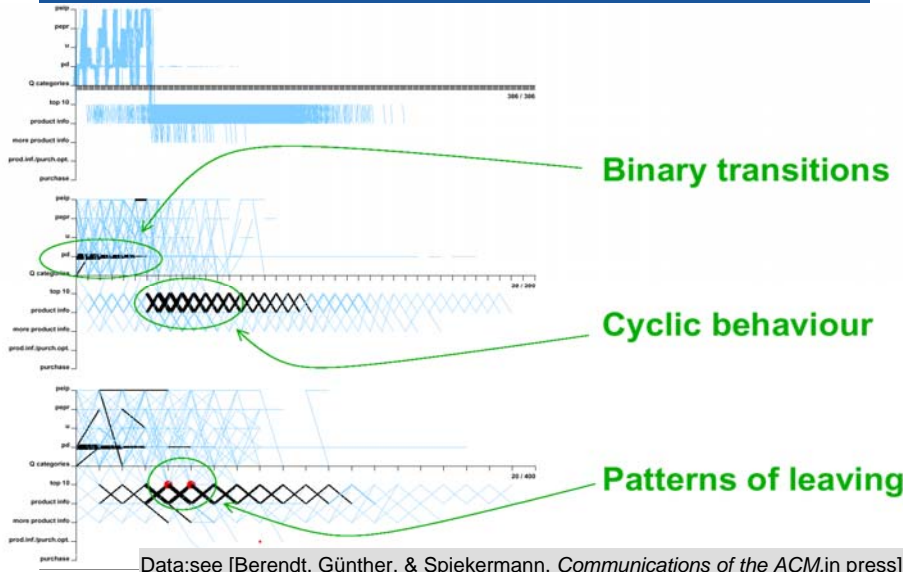


[Berendt, *Data Mining and Knowledge Discovery*, 2002], [Berendt, *Postproc. WebKDD 2001*]

## Communication – Visual data mining

### Step 6: Visual abstraction → new semantic patterns

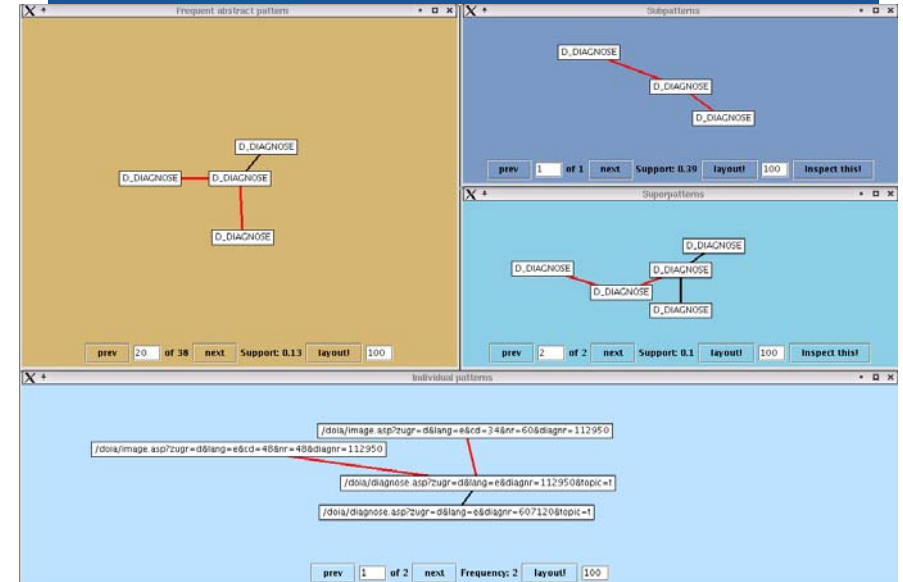
19



## More expressive patterns and visualisations

### Step 7: Graph mining and semantic detail-and-context

20



## What are problems with log analysis in different fields?

21

→ Here: limitations of log-mining style analyses

- **Samples**
  - Often only one site, one application domain, etc.
  - Self-selection of users
- **Observational field studies**
  - Only conjectures regarding goals, expectations, evaluations
  - Barely any control of relevant variables
  - Frequently confounded variables
- **Data**
  - Noise and errors resulting from
    - data collection procedures
    - preprocessing heuristics (but: they aren't as bad as you'd think!)
- **Time**
  - Momentary snapshots; co-evolution of Web and users

## How generalisable are the techniques/findings of log analysis on specific logs?

22

- **Techniques: rather trivially generalisable**
- **Findings:**
  - Good question! Good research issue!
  - A start?!: AOL vs. Microsoft search logs
    - Strohmaier et al., ADMI Workshop at WI 2008
    - But is there a real cross-validation?

23

## Research issues II General questions

## How can we evaluate approaches to log analysis? (What kind of benchmarks do we need, how do we generate them and what kind of evaluation campaign should be run?)

24

- **Answer depends on how logs are viewed:**
  1. logs as behavioural trails
  2. log analysis tools as software
  3. log analysis for applications
- **Ad (1): e.g., replicability – e.g. how stable are patterns across log parts (= sample across or along time), logs (= one site), different sites [but need good measures of concept drift!]**
- **Ad (2): e.g., understandability of patterns? (need analysts as experiment participants) actionability? (the latter is already a bit deployment too)**
- **Ad (3): Depends on the goal of the application; need application-dependent measures (e.g. personalisation; site re-design)**

Can we develop a meta-methodology that combines log analysis with other methods to provide a "truer" picture of the user - system - information interaction process?

25

- First observation:
- We need a combination of measurement methods
- But also
- A combination of questions / background theories

## Meta-methodology: One instance (Berendt & Kralisch, *Information Retrieval* 2009)

26

1. Cognition: language and information-seeking *behaviour* → log analysis
2. Economics: information and information flow on the Web (*data*) → Web statistics
3. Information systems and technology acceptance: *attitudes* → questionnaires

	Construct	Operationalised in Section...as...	
.1	Cognitive value of information	Assumed independent of language capabilities	
	Cognitive costs of (accessing) information	3.1–3.3	L1 vs. L2
		4	Proficiency in English
		4	The inverse of saved effort
2	Supply	3.1	Number of Web hosts
		3.2	Number of existing inlinks
	Demand	3.1	Size of language group
		3.2–3.3	Number of used inlinks
3	Ease of use	4	Saved effort
	Usefulness	4	(The inverse of) perceived amount of content supply on the Web
	Satisfaction	4	Satisfaction (has repercussions on cognitive value)

What are the future challenges/directions for the field of query log analysis/mining?

27

- Relevant IMO
  - method/theory combinations
  - Interaction is more than navigation + querying / Social Media et al.
  - privacy preservation
    - (but: whose privacy? see Domingo-Ferrer, *Secure Data Management*, 2007)
  - time-series analysis: concept drift, co-evolution
- Certainly a priority, but I don't have anything to say about it
  - Web search advertising
  - What is the impact of services like Twitter on search? Can microblogging streams be seen as queries and thus analyzed in some way?
- Only in circumscribed settings (laboratory?!)
  - eye tracking
- Data preprocessing issues that are basically solved (or about which it is known to what extent they are solvable)
  - integration of multiple transaction logs
  - correlating transaction logs with user behaviour

28

Feasibility I:  
Within academia

## What approaches could be used to generate logs to share within the research community?

- User participation
- Consent
- Rewards/incentives!
  - Instant gratification?!
- Trade privacy for personalization benefits???

## How can we bring researchers from different disciplines closer together?

- Events like this one!
- More principled approaches for
  - Describing questions about behaviour AND
  - describing behavioural-research methods AND
  - exchanging these descriptions

**Feasibility II:**  
Transfer/collaborat. academia – industry

## Are there any application different from web search engines that generate logs that once analyzed might bring benefits to the user?

- Any kind of knowledge-accessing/-creating software
  - learning,
  - portals,
  - knowledge management
  - ...

## Where are areas for academic - industry collaboration? What are the niches & contrib.s that academia can make to log analysis?

33

- Methods expertise AND
  - Different questions
- vs.
- Requirements / different questions AND
  - data

## How can researchers get access to logs? (E.g. what will stop industry from sharing logs?)

34

[yes, this slide is on the polemical side, and there are other reasons too ...]



## How can researchers get access to logs?

35

- NDAs
- Better privacy-preserving methods
  - Attention: whose privacy? (not only) terminological confusion
- Look at other industry fields
  - ... Here, I am on thin ice ...
  - Medical research?
  - Ethics codes / committees?

## How can we effectively transfer research into industry? (Is there any application used by industry that could benefit from query log analysis?)

36

- The answer is a combination of 2 previous answers:
- Applications – any kind of knowledge-accessing/-creating software
  - learning,
  - portals,
  - knowledge management
  - ...
- Data need to be made available!

## How can we generate funding opportunities from grant agencies in log analysis?

37

- point out importance in basic research
- point out importance for supporting applications



What do you think?

## References with URLs (1)

39

[references highlighted in blue might be particularly interesting for participants of the TrebleCLEF workshop]

slide 5:

Berendt, B. & Spiliopoulou, M. (2000). Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9, 56-75. <http://vasarely.wiwi.hu-berlin.de/Home/berendt-spiliopoulou-vidbj00.pdf>

Berendt, B. & Brenstein, E. (2001). Visualizing Individual Differences in Web Navigation: STRATDYN, a Tool for Analyzing Navigation Patterns. *Behavior Research Methods, Instruments, & Computers*, 33, 243-257. [http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt\\_brenstein\\_2001.pdf](http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt_brenstein_2001.pdf)

Ritter, F., Berendt, B., Fischer, B., Richter, R., & Preim, B. (2002). Virtual 3D jigsaw puzzles: Studying the effect of exploring spatial relations with implicit guidance. In *Proceedings of Mensch und Computer 2002* (pp. 363-372), Hamburg, Germany, 2-5 September 2002. <http://warhol.wiwi.hu-berlin.de/~berendt/Papers/MC2002.pdf>

Teltzrow, M., Berendt, B., & Günther, O. (2003). Consumer behaviour at multi-channel retailers. In *Proceedings of the 4th IBM eBusiness Conference*, School of Management, University of Surrey, 9th December 2003. [http://warhol.wiwi.hu-berlin.de/~berendt/Papers/teltzrow\\_berendt\\_guenther\\_2003.pdf](http://warhol.wiwi.hu-berlin.de/~berendt/Papers/teltzrow_berendt_guenther_2003.pdf)

Teltzrow, M., & Berendt, B. (2003). Web-Usage-Based Success Metrics for Multi-Channel Businesses. In *Proceedings of the WebKDD 2003 Workshop - Webmining as a Premise to Effective and Intelligent Web Applications*. August 27th, 2003, Washington DC, USA. Held in conjunction with The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [http://warhol.wiwi.hu-berlin.de/~teltzrow/teltzrow\\_berendt\\_webkdd03.pdf](http://warhol.wiwi.hu-berlin.de/~teltzrow/teltzrow_berendt_webkdd03.pdf)

Berendt, B., Günther, O., & Spiekermann, S. (2005). Privacy in E-Commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, 48(4), 101-106. <http://warhol.wiwi.hu-berlin.de/~berendt/Papers/p101-berendt.pdf>

Kralisch, A., Eisend, M., & Berendt, B. (2005). Impact of Culture on Website Navigation Behaviour. In *Proceedings of 11th International Conference on Human-Computer Interaction*, Las Vegas, NE, 22-27 July 2005. [http://www.cs.kuleuven.be/~berendt/Papers/kralisch\\_berendt\\_eisend\\_2005.pdf](http://www.cs.kuleuven.be/~berendt/Papers/kralisch_berendt_eisend_2005.pdf)

Berendt, B. & Kralisch, A. (2009). A user-centric approach to identifying best deployment strategies for language tools: The impact of content and access language on Web user behaviour and attitudes. *Journal of Information Retrieval*, 12 (3), 380-399. [http://www.cs.kuleuven.be/~berendt/Papers/berendt\\_kralisch\\_2009.pdf](http://www.cs.kuleuven.be/~berendt/Papers/berendt_kralisch_2009.pdf)

## References with URLs (2)

40

[references highlighted in blue might be particularly interesting for participants of the TrebleCLEF workshop]

Slide 10

Educational portals / search / \* Requested (search) services

Berendt & Spiliopoulou (2000), s.a.  
Oberle, D., Berendt, B., Hotho, A., & Gonzalez, J. (2003). Conceptual user tracking. In E. Menasalvas Ruiz, J. Segovia, & P.S. Szczepaniak (Eds.), *Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003, Proceedings* (pp. 155-164). Berlin: Springer, LNCS 2663. <http://warhol.wiwi.hu-berlin.de/~berendt/Papers/AWIC03.pdf>

Medical information site / \* Search service: modality; \* ICD-9 classification of diseases

Berendt, B. (2006). Using and learning semantics in frequent subgraph mining. In Nasraoui, O., Zaiane, O., Spiliopoulou, M., Mobasher, B., Yu, P., & Masand, B. (Eds.), *WebKDD05 - Selected revised papers*, (pp. 18-38). Springer LNCS 4198. [http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt\\_webkdd05\\_book\\_to\\_appear.pdf](http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt_webkdd05_book_to_appear.pdf)

E-commerce / \* Stage in buying process

Teltzrow et al. (2003), s.a.

Slide 11

Language / L1 vs. L2  
Berendt & Kralisch (2009), s.a.

Culture / Hofstede's country-based Cultural Dimensions

Kralisch et al. (2005), s.a.  
Berendt, B. & Kralisch, A. (2007). From World-Wide-Web Mining to Worldwide Webmining: Understanding People's Diversity for Effective Knowledge Discovery. In B. Berendt, A. Hotho, D. Mladenovic, & G. Semeraro (Eds.), *From Web to Social Web: Discovering and Deploying User and Content Profiles* (pp. 102-121). LNAI 4737. Berlin etc.: Springer. [http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt\\_kralisch\\_2007.pdf](http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt_kralisch_2007.pdf)

„Relation to site“ / \* Customer (or just visitor); \* Method of first contact w. site; \* Possible offline customer?

Teltzrow & Berendt (2003), s.a.  
„Relation to site“ / \* Type of shopped-for goods  
Berendt et al. (2005), s.a.

## References with URLs (3) [references highlighted in blue might be particularly interesting for participants of the TrebleCLEF workshop]

### Slide 12

all of the cited papers

### Slide 13

Berendt (2006), s.a.

Berendt, B. (2002). Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*, 6, 37-59. [http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt\\_2002.pdf](http://warhol.wiwi.hu-berlin.de/~berendt/Papers/berendt_2002.pdf)

### Slides 14, 15

Oberle et al. (2003), s.a.

### Slides 16-18

Berendt & Spiliopoulou (2000), s.a.

Berendt (2002), s.a.

### Slide 19

Berendt et al. (2005), s.a.

Berendt, B. (2002). Detail and context in Web usage mining: coarsening and visualizing sequences. In R. Kohavi, B.M. Masand, M. Spiliopoulou, & J. Srivastava (Eds.), *WEBKDD 2001 - Mining Web Log Data Across All Customer Touch Points* (pp. 1-24). Berlin etc.: Springer, LNAI 2356. <http://warhol.wiwi.hu-berlin.de/~berendt/Papers/paper2.pdf>

### Slide 20

Berendt (2006), s.a.