Learning Navigation Structures

Approaches and Findings

Topics of our research group…

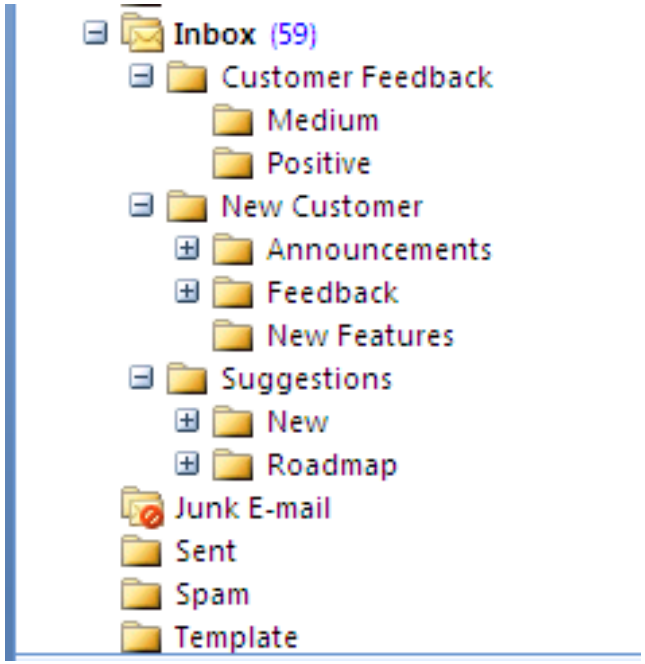
- «Information logistics»
  - meta data models
  - automatic generation of metadata
  - information extraction
  - Information retrieval

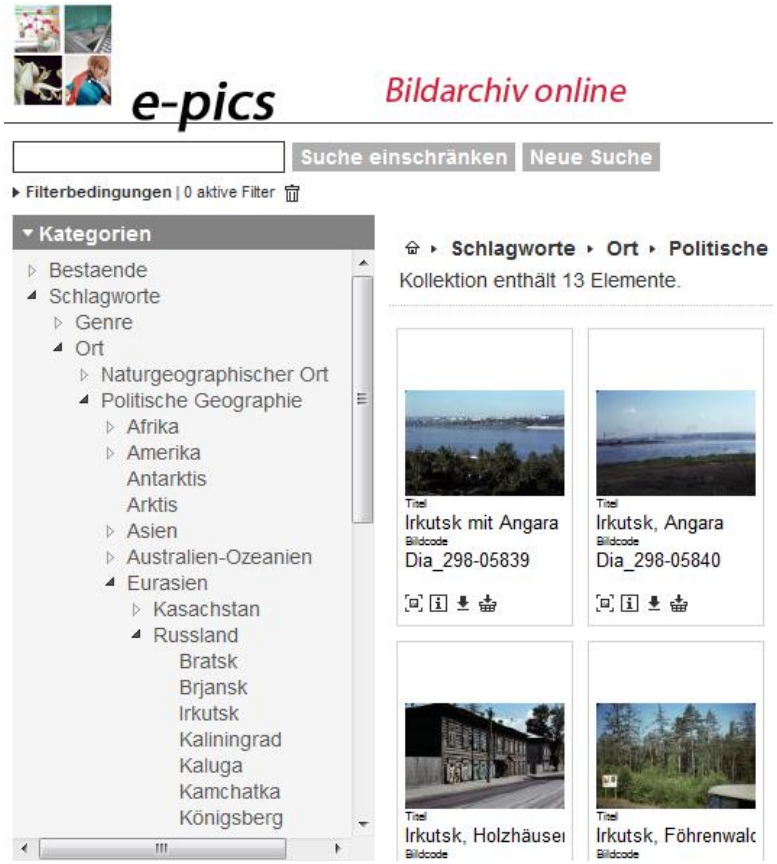
- Supporting knowledge-intensive processes
  - Re-use of experience and knowledge
  - Decision support (business intelligence)
  - Knowledge engineering

Introduction to the Problem

Hypothesis 1 (unproven)

- some people like to structure information hierarchically





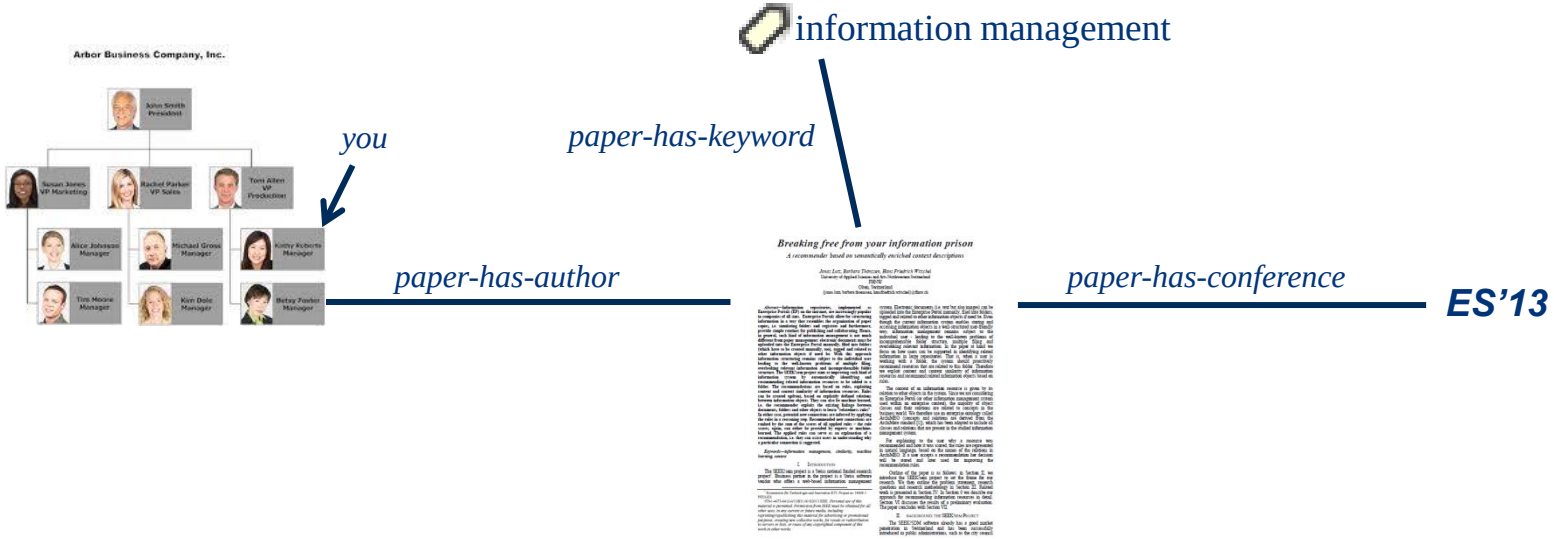
Hypothesis 2 (to be proven)

- *If you give me your email folders*
  → will I be able to make sense of them?
  → will I be able to find stuff?
  *→ or will there be a «**retrieval gap**» between your worldview and mine?*
- *Is it maybe easier (for me) to understand a hierarchy that was computed automatically?*

Hypothesis: yes, if the result is properly explained!

Starting point (1)

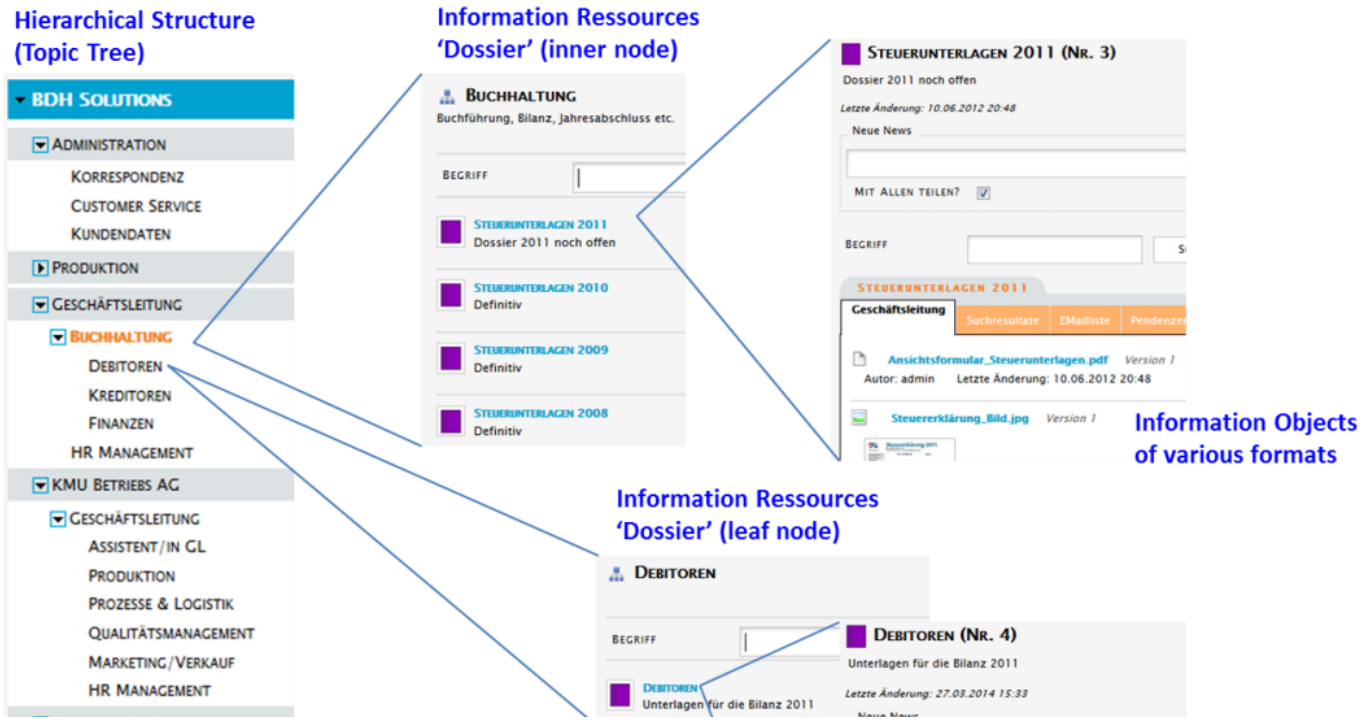
- We have resources (documents, pictures, …) with metadata
- we represent this as a graph…



possible metadata: keywords, tags, topics, creator, creation date, type,…, conference/journal, publisher, editor,…

Starting point (2)

- Research project together with BDH Solutions, software SEEK!SDM:
  - Resources are organised manually into dossiers
  - Dossiers are organised manually into a topic tree



Goals for Text Mining

- Goal: make navigation structures…
  - …more easily understandable
  - … easier to maintain

- Therefore, we want to apply text and data mining to…
  - … create additional links in existing topic trees
  - … compute topic trees from scratch
  - … Provide explanations: justify additional links, label nodes of topic trees

Some technical details

Creating additional links: goals

- We would like to…
  - … **represent** the context of an information object (as a graph with typed vertices and edges)
  - … **discover** rules (=abstract paths) to recommend resources
  - … **learn** how good these rules are by analysing existing human-made assignments
  - …**explain** recommendations by means of our symbolic representation

Creating additional links: definitions

- A **dossier** is a collection of information resources in a target information management system.
- A **positive example** of similarity is a manual assignment, e.g. a pair (D,R) where R is a resource assigned to dossier D
- A **negative example** of similarity is any pair (D,R) where no such assignment exists
- An **abstract path** is a sequence of edge types. Example: *(creator, creator)*

Creating additional links: general procedure

Idea: for a dossier D, find resources that can be reached via many «promising» paths

1. Preparation: learn a score lift(p) for each abstract path p that tells how likely p connects positive examples (i.e. use manual assignments as training data)
2. For a given dossier D, find all resources R_i that can be reached from D via short paths, score R_i by the sum of scores of the corresponding abstract paths
3. Output: list of R_i, ranked by their score

Learning a topic tree from scratch (1)

- Goal: learn a complete topic tree that can be used for browsing
- We describe each resource by a list of vectors (one for each type of attribute, e.g. one for tags, one for keywords, one for authors etc.)
- We define a distance measure for these descriptions and use it to cluster resources hierarchically

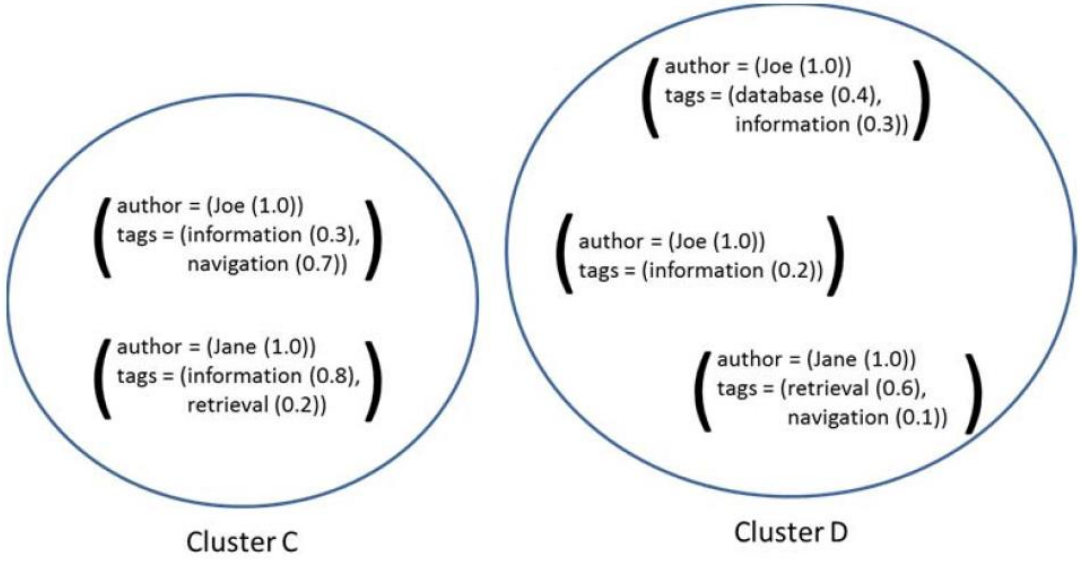


Figure 1: Two clusters with resource representations

Learning a topic tree from scratch (2)

- When we merge two clusters, we create a *similarity explanation* (summary of characteristics shared by both clusters)

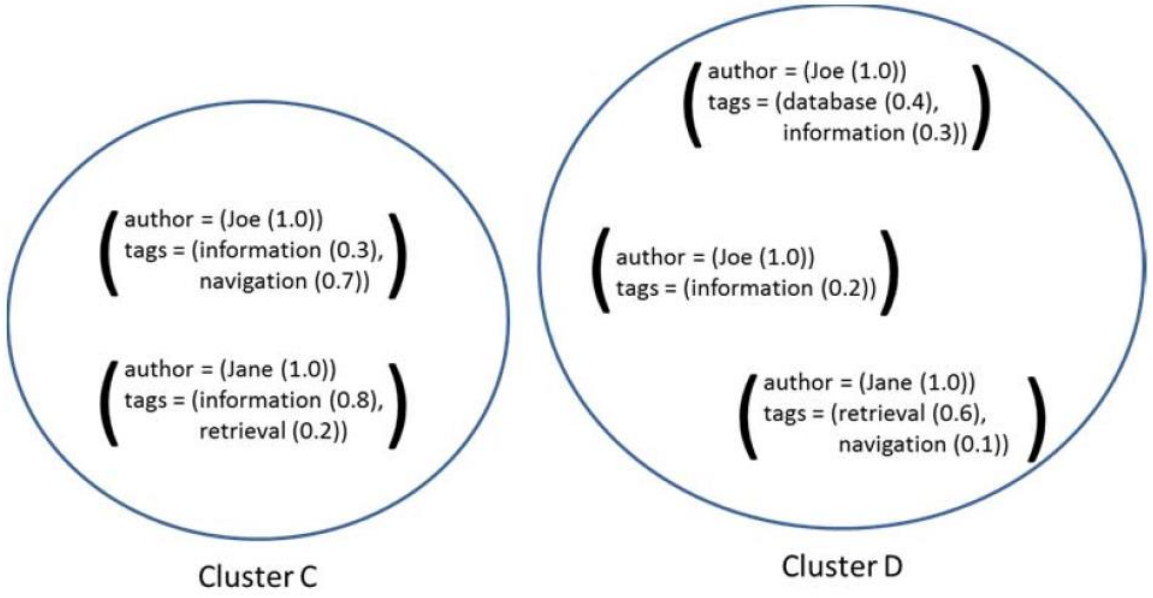


Figure 1: Two clusters with resource representations

Explanation:

author = (Joe (0.33), Jane (0.17))
tags = (information (0.67), retrieval (0.17),
navigation (0.17))

Learning a topic tree from scratch (3)

- Finally, we collapse the binary dendogram (clustering tree):
  - traverse the tree from top to bottom
  - for consecutive pairs of nodes on a path, compare their similarity explanations
  - If they were merged for «the same reason» collapse them into one node, otherwise create different nodes
- Label the final tree nodes using the similarity explanations

Results and discussion

Creating additional links: results

- Some qualitative results that we obtained:

Abstract path	Natural language description
dossier-has-resource -> resource-has-tag -> resource-has-tag	"the recommended resource shares a tag with a resource of this dossier"
dossier-has-tag -> resource-has-tag	"the recommended resource shares a tag with this dossier"
dossier-has-creator -> resource-has-creator	"the recommended resource was created by the person who also created this dossier"
dossier-has-topic -> dossier-has-topic -> dossier-has-creator -> resource-has-creator	The recommended resource was created by a person who also created a dossier with the same topic as this one.

- Conclusion: this seems to make intuitive sense…

Automatically computed topic trees: experiment

- «Traditional» evaluations compare computed topic trees to manually created ones («gold standard»)
  - But: the computed tree might be better suited for navigation!

- Our approach: compare both trees by applying them in a user task
  - provide a manually created and a computed tree
  - give some randomly selected documents to test persons, ask them to locate these documents in both trees
  - record the time plus the number of backtracks needed to find resources

Automatically computed topic trees: results

Test Case	Tree	Testperson 1		Testperson 2	
		Time (sec.)	Back-tracks/ min.	Time (sec.)	Back-tracks/ min.
1	*Manual*	595*	2.9	270*	2.7
	SE-labels	202	1.8	173	0.3
2	*Manual*	65	0.9	191	3.8
	SE-labels	**45**	0	**20**	0
3	*Manual*	96	3.8	310*	7.5
	SE-labels	**21**	0	**58**	1.0
4	*Manual*	329	2.4	131	0.9
	SE-labels	**226**	1.1	456*	1.4

Table 1: Summary of time needed to locate test cases, and number of backtracks per minute. Cases that were given up are marked with *

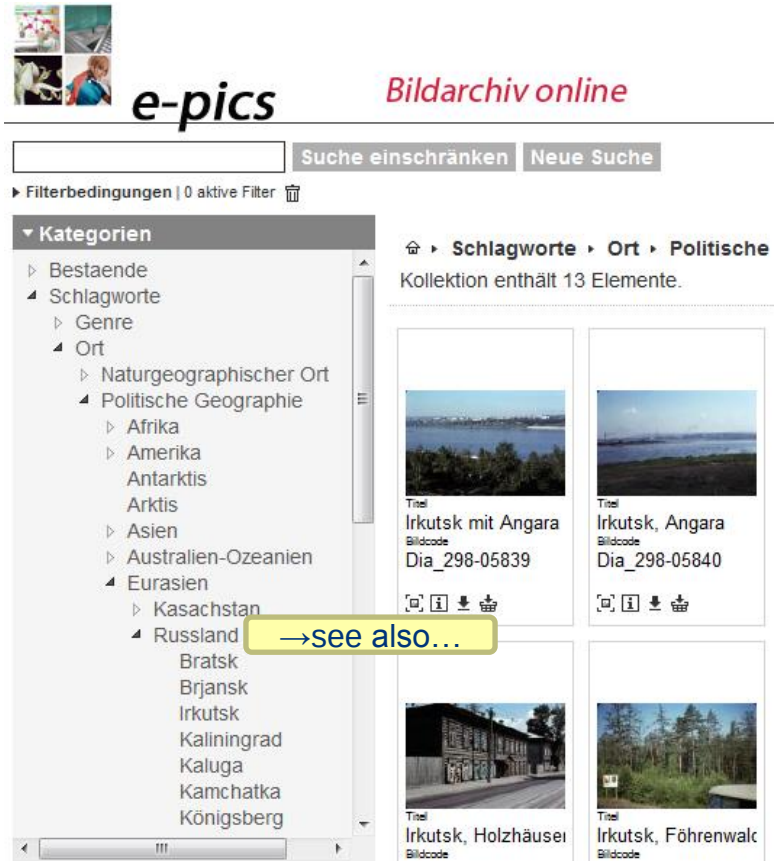
- test persons found the computed tree «cryptic» at first glance, but surprisingly easy to work with
- They took longer to find documents in the manually created tree because they were not familiar with the company's terminology and way of organising things

What might it mean for libraries?

- Manually created hierarchies of topics reflect a certain world view (e.g. that of librarians or domain experts) that may not be shared by the users
  - This can make it hard to find certain resources
  - Yes, automatic approaches may result in hierarchies that are easier to navigate
  - **But: does the manually created hierarchy not in itself contain valuable knowledge (about how to structure a topic) that is learned when browsing it?**
  - **What is more important?**
    → creation of additional links might be a good compromise…

Additional links in a library catalogue?

- In a library catalogue, a manually created hierarchy provides orientation about a topic
- To mitigate the «retrieval gap» between the librarians' and the users' worldview, we could add links to similar resources
- Similarity could be based on
  - content
  - metadata
  - usage patterns



¡Thank you!

References

- J. Lutz, B. Thönssen, and H. F. Witschel. Breaking free from your information prison - A recommender based on semantically enriched context descriptions. In *Proceedings of the First International Conference on Enterprise Systems*, 2013.
- H. F. Witschel, B. Thönssen and J. Lutz. Where Did I(t) put it? A holistic solution to the automatic construction of topic trees for navigation. In *Proceedings of KMIS'14*, 2014.