Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

The Lokahi Prototype: Toward the Automatic Extraction of Entity Relationship Models from Text

Prof. Dr. sc. inf. Michael Kaufmann

+41 41 757 68 48
m.kaufmann@hslu.ch

AAAI 2019 Spring Symposium on
Combining Machine Learning with Knowledge Engineering
March 25–27, 2019
@ Stanford University, Palo Alto, California, USA

FH Zentralschweiz

Data Intelligence Group

- Data intelligence is the ability to gain and apply knowledge and skills based on data.
- The Data Intelligence Group is researching systems and methods that can capture data, gain knowledge from data, and facilitate interaction with data.

Prof. Dr. Michael Kaufmann
Koordinator Forschung
Dozent

Dr. Alexander Denzler
Koordinator Team
Dozent

Dr. Luca Mazzola
Wissenschaftlicher
Mitarbeiter Senior

Dr. Ladan Pooyan-Weihs
Dozentin

Prof. Dr. Tim Weingärtner
Dozent

Andreas Waldis
Master-Assistent

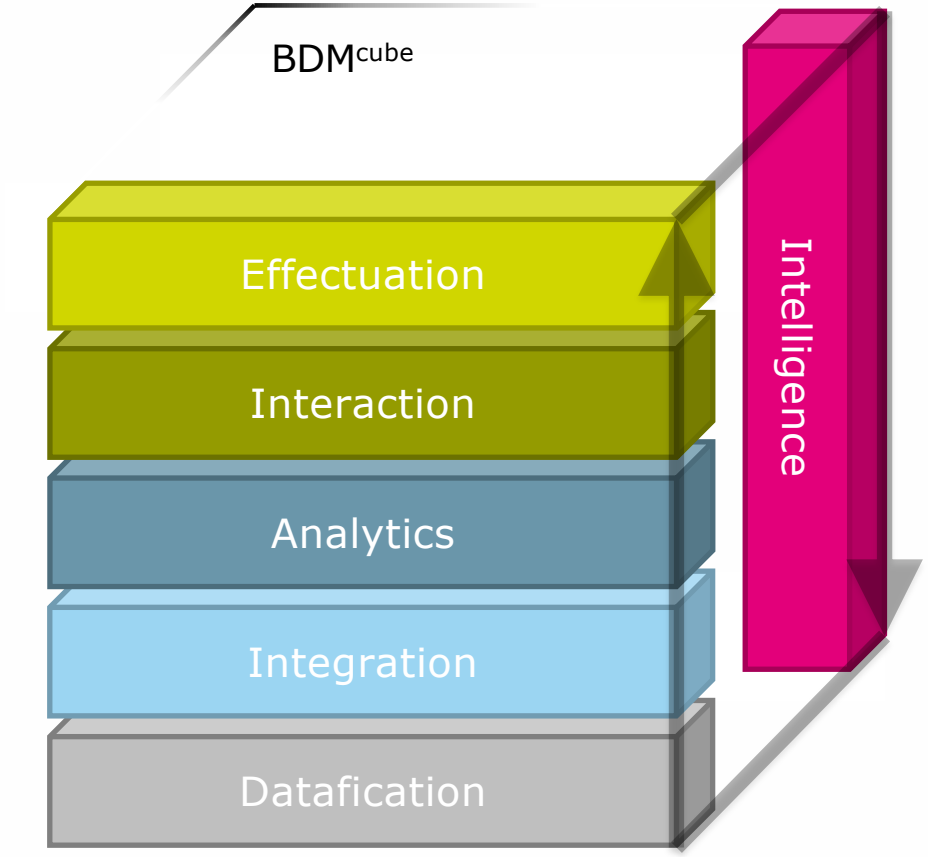
Patrick Siegfried
Wissenschaftlicher
Assistent

Florian Stalder
Master-Assistent

Christian Renold
Master-Assistent

Dr. Alevtina Dubovitskaya
Dozentin

Research Program



https://www.hslu.ch/de-ch/informatik/forschung/teams/data-intelligence-research/

Big Data Management Meta-Model (BDMcube)
Kaufmann, M., Eljasik-Swoboda, T., Nawroth, C., Berwind, K., Bornschlegl, M., Hemmje, M. (017). Modeling and Qualitative Evaluation of a Management Canvas for Big Data Applications. Accepted for publication as a regular paper at the 6th international Conference on Data Science, Technology and Applications DATA 2017, Madrid, 24 - 26 July 2017.

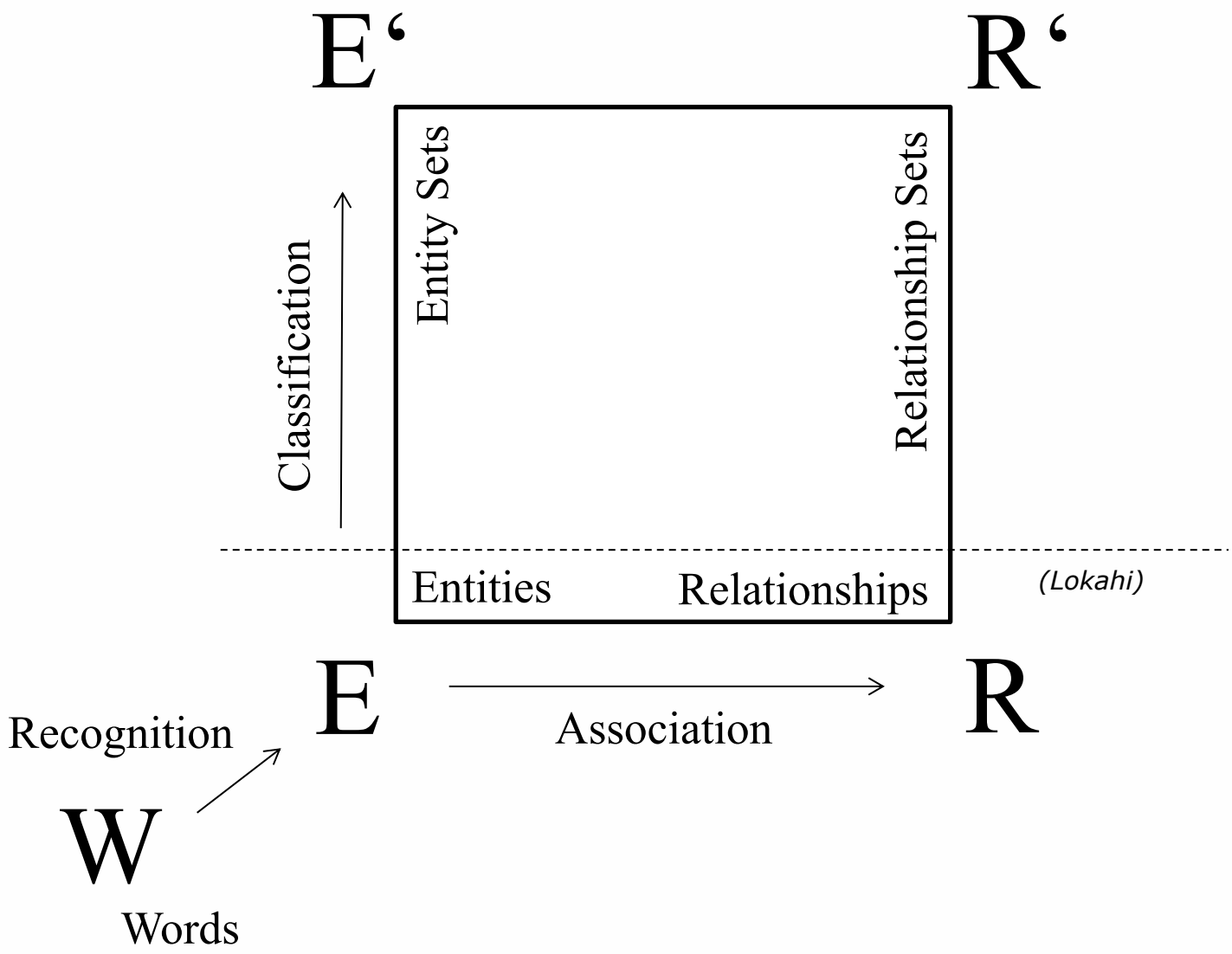
Motivation

- Data Explosion: The percentage of data we can make use of decreases (and eventually asymptotically approaches zero)
- Knowledge extraction techniques to provide overview
- Vision: automatic extraction of symbols structure from text
  - Entities
  - Relationships
- Possible applications:
  - automatic tagging of e-mails
  - automatic categorization in document management systems
  - explorative analysis of essential content in unstructured text data

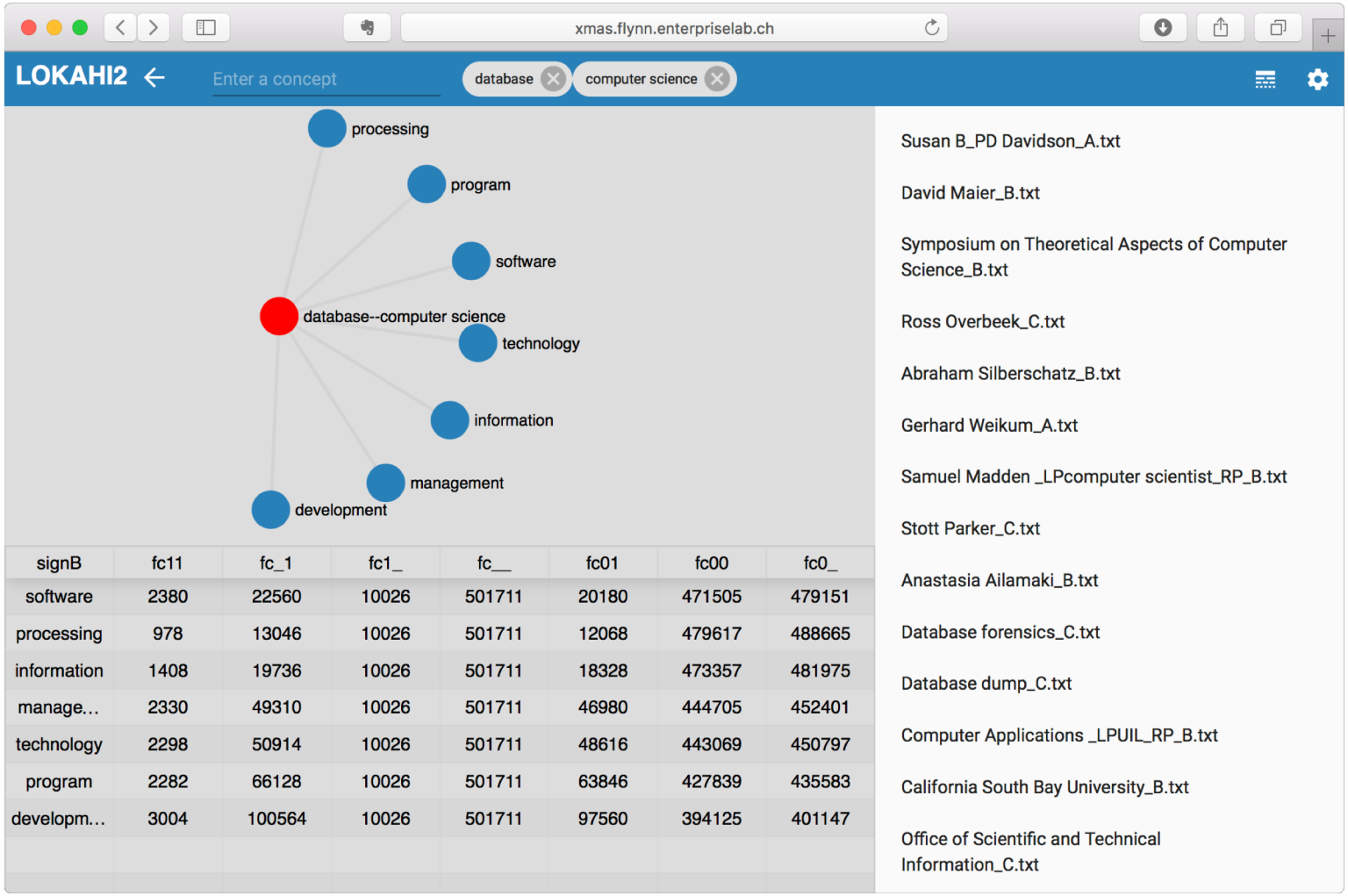
Representing Knowledge as Networks

- Semantic Networks (Quillian, 1967)
- Conceptual Graphs (Sowa, 1976)
- Entity-relationship models (Chen, 1976)
- Concept Maps (Novak & Gowin, 1984)
- Topic Maps (Rath & Pepper, 1999)
- Semantic Web (RDF)
- Data-driven knowledge networks (Maedche & Staab, 2001, Alani et al., 2003)

Knowledge Extraction Framework

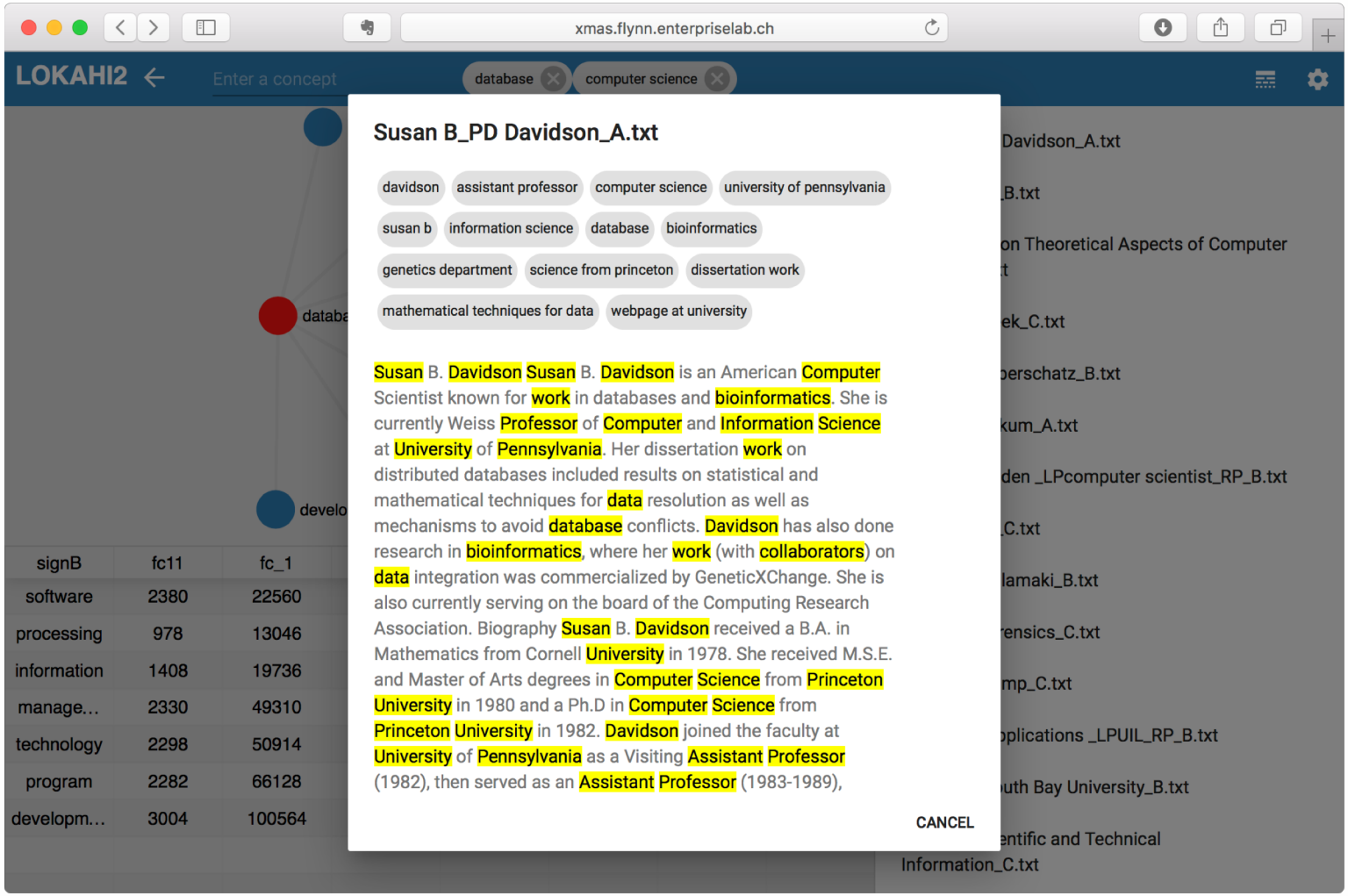


The Lokahi Prototype: Search Interface



signB	fc11	fc_1	fc1_	fc__	fc01	fc00	fc0_
software	2380	22560	10026	501711	20180	471505	479151
processing	978	13046	10026	501711	12068	479617	488665
information	1408	19736	10026	501711	18328	473357	481975
manage…	2330	49310	10026	501711	46980	444705	452401
technology	2298	50914	10026	501711	48616	443069	450797
program	2282	66128	10026	501711	63846	427839	435583
developm…	3004	100564	10026	501711	97560	394125	401147

The Lokahi Prototype: Automatic Tagging



Extraction of Entities with TF*IDF

- Core Idea: Reversing the search engine.
  - Instead of computing relevant documents for terms
  - => compute relevant terms for documents
  - => Keyphrases are, with high probability, named entities

- S(t,d) = TF(t,d) * IDF(t)

- IDF(t) = 1 + log(n / (DF(t) + 1))

- S'(t,d) = ($TF(t,d)^2$ + IDF(t)) / |d| <= Experimentally better results

- Question: What about n-grams? => Mazzola et al. 2018, 2019 (to appear)
  - Extraction of n-grams by document-based avg. PMI and ratio of Std.Dev. of word probability

$$pmi(w_1, w_2) = \frac{P(w_1, w_2)}{P(w_1) * P(w_2)}$$

$$avgPMI(\boldsymbol{w}) = \frac{1}{(n-1)} * \sum_{i=1}^{n-1} pmi(w_i, w_{i+1})$$

$$invsd(\boldsymbol{w}) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} |P(w_i) - \bar{P}(\boldsymbol{w})|}$$

$$\bar{P}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} P(w_i)$$

$$score(\boldsymbol{w}) = \begin{cases} log(TF(\boldsymbol{w})) * si(\boldsymbol{w}) & \boldsymbol{w} \text{ is a proper word} \\ log(TF(\boldsymbol{w})) * (avgPMI(\boldsymbol{w}) + invsd(\boldsymbol{w})) & \boldsymbol{w} \text{ is a proper n–gram} \end{cases}$$

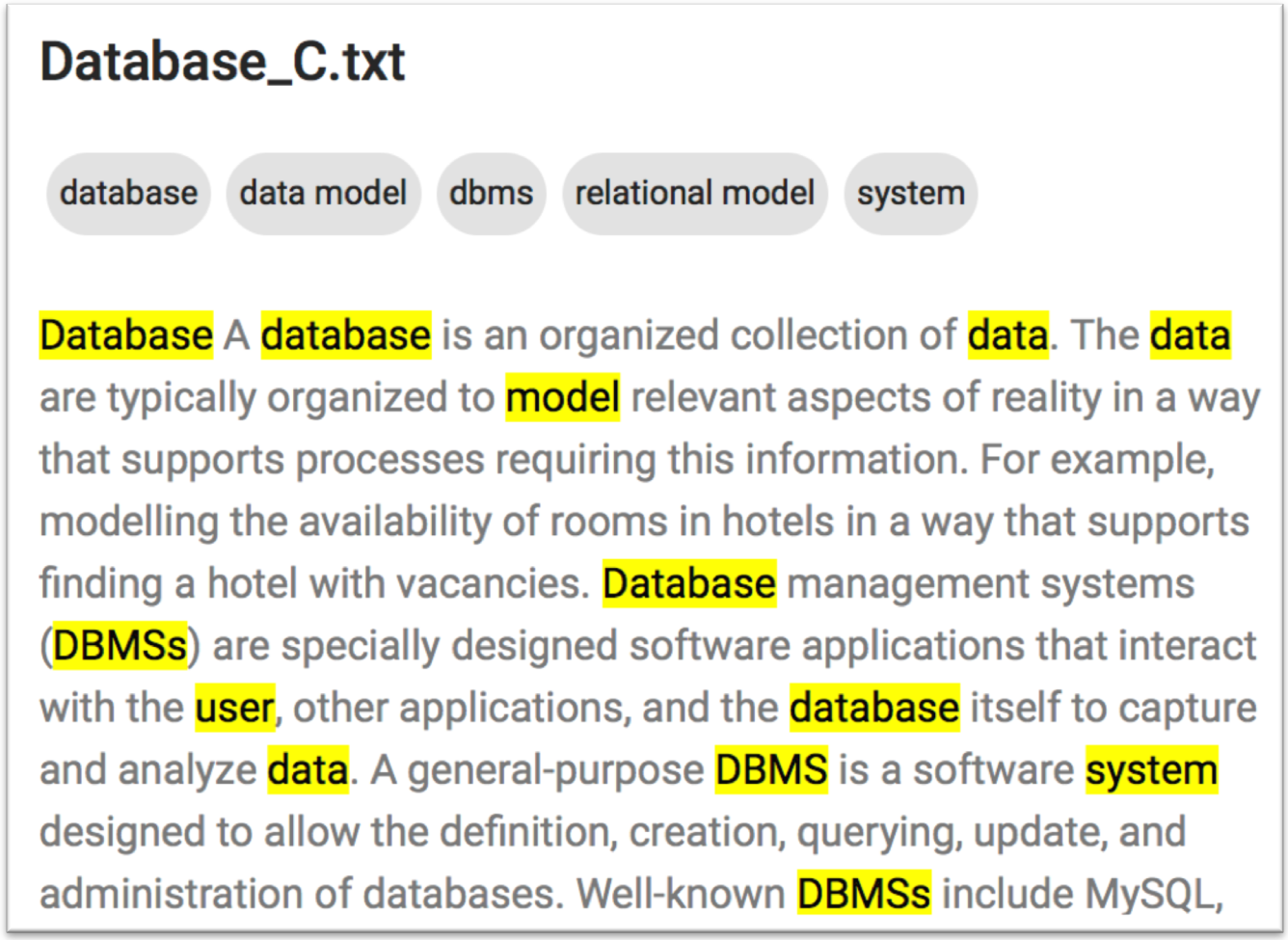
Extraction of Keyphrases for „Computer Science“

Computer science_B.txt

computer science | computing | computation | software engineering | theory | academic discipline | study | system | data | field | design | programming language | application | theoretical

Computer science Computer science is the scientific and practical approach to computation and its applications. It is the systematic study of the feasibility, structure, expression, and mechanization of the methodical procedures (or algorithms) that underlie the acquisition, representation, processing, storage, communication of, and access to information, whether such information is encoded as bits in a computer memory or transcribed in genes and protein structures in a biological cell. A computer scientist specializes in the theory of computation and the design of computational systems. Its subfields can be divided into a variety of theoretical and practical disciplines. Some fields, such as computational

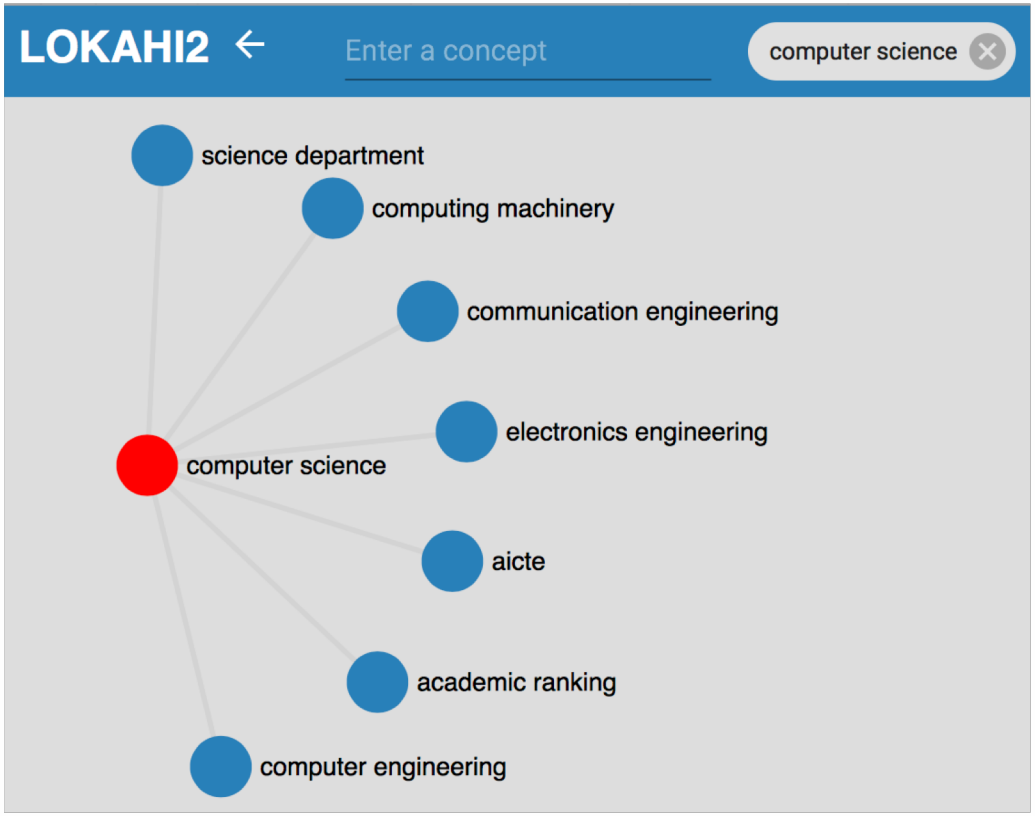
Extraction of Keyphrases for „Database“

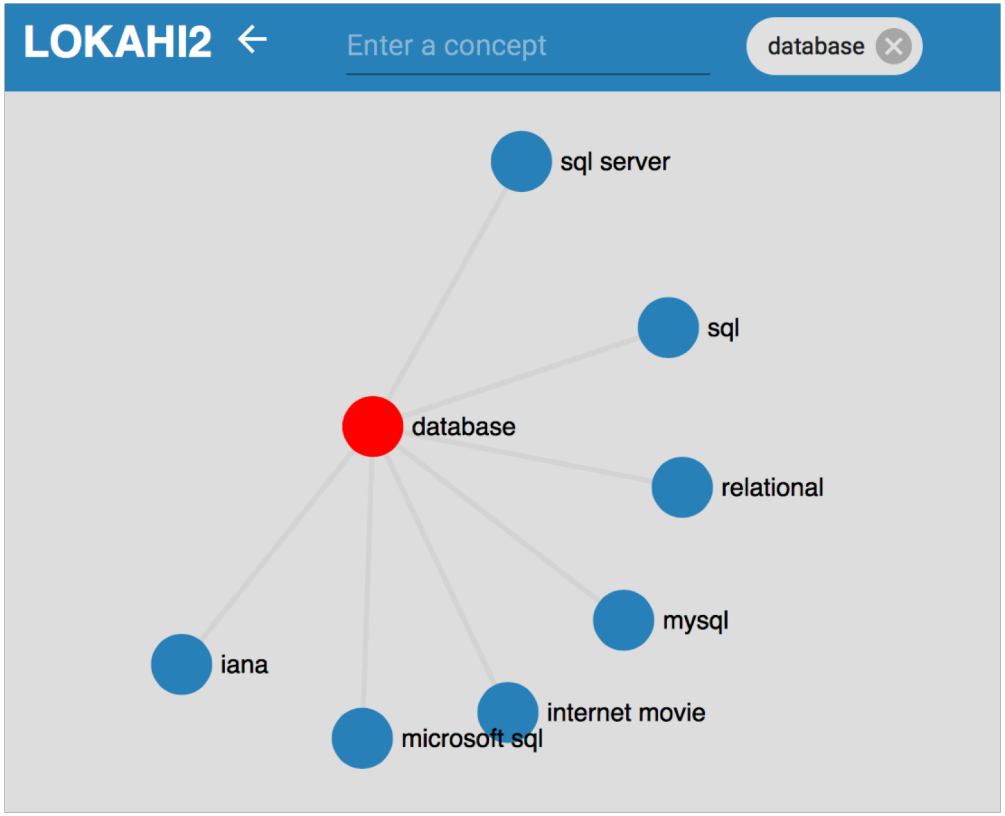


Extraction of Relationships by Co-Occurrence

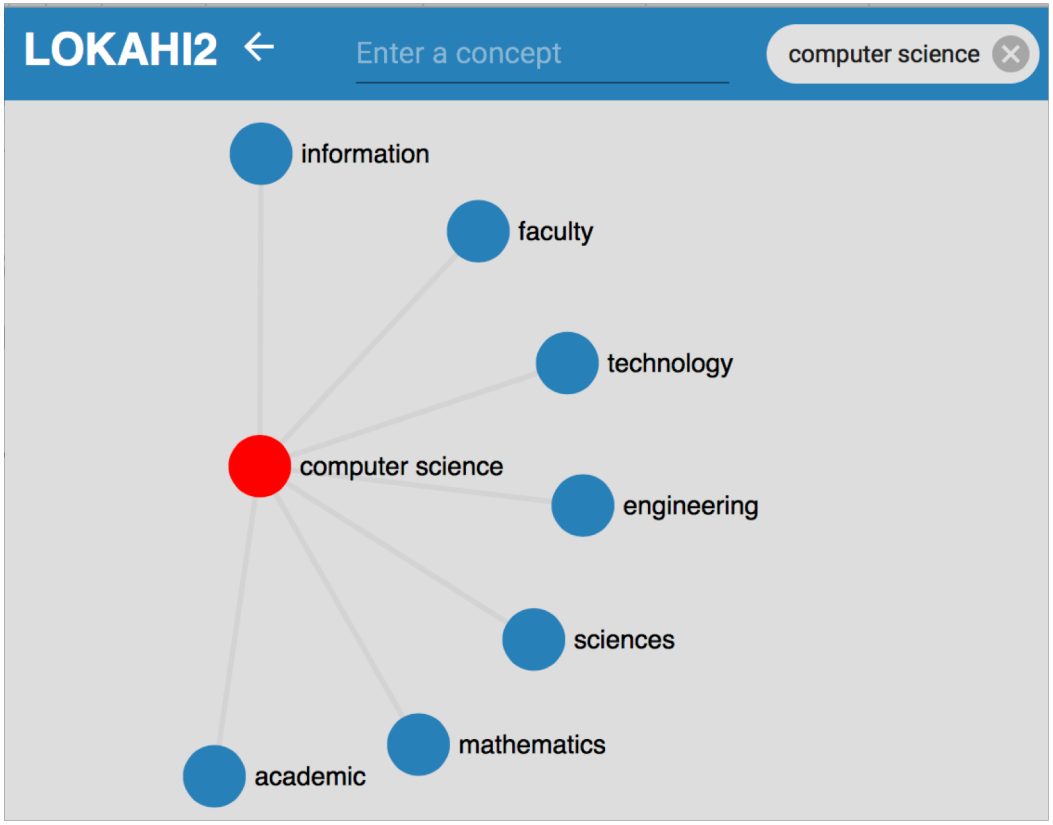
- Core Idea: Basket analysis / association mining / frequent itemsets.
  - “Basket” => Documents
  - “Items” => Keyphrases
  - Probability p(A): 1 / Number of Documents containing phrase A
- Measures based on corpus-level distribution:
- PMI(a,b) = log(p(a,b) / (p(a) * p(b)))
- LR(a,b) = p(a | b) / p(a | not b)
- It is evident that likelihood-based approaches based on corpus-level co-occurrence can extract semantic relationships.
- However, the relationship sets / classes / labels cannot be extracted this way => further research

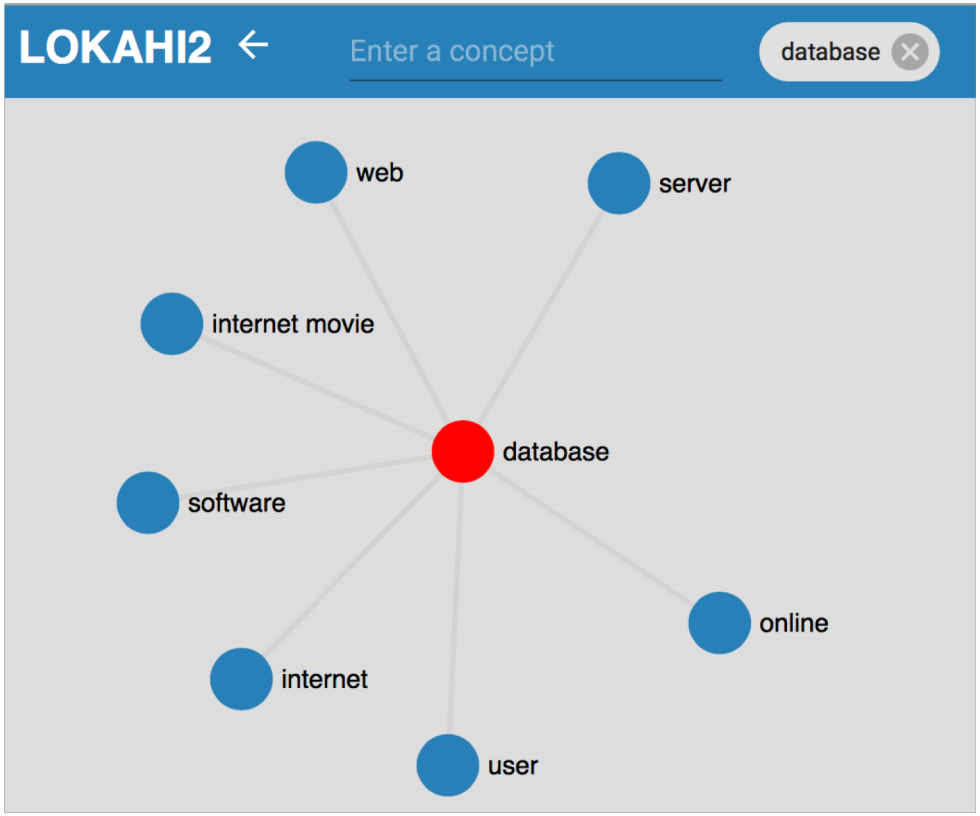
Extraction of Relationships based on Likelihood Ratios





Extraction of Relationships based on Pointwise Mutual Information





Qualitative Evaluation: Expert Feedback

- Goal: extraction of „meaningful" and „useful" entities and relationships
- Presentation and discussion of results with tho industry partners
  - FIVE Informatik AG, Bern: Information Technology service provider
  - Netcetera AG, Zürich: Information Technology service provider
- Key Findings:
- 1. Quality of extracted knowledge is, at this point, not suitable for commercial application
- 2. Market for enterprise search is small; document management systems could be a better field of applicattion
- 3. The network as metaphor or meta-structure is interesting but unusual
- 4. Hybrid Approach is needed, where humans can edit the resulting knowledge network

Conclusions and Further Research

Boxology ☺

- Research with the Lokahi system could demonstrate that purely statistical computation can extract semantically meaningful entities and relationships

- Neither KBS nor ANN but statistical approach
  - Shi & Griffith 2009: In the human brain, symbols are grounded in statistical sampling and can be modelled with Bayesian approaches
  - => Griffith 2011: Rethinking language: statistical instead of rule based
  - Can handle ambiguous or contradictory observations compared to POS

- Potential for semantic and explorative analysis of unstructured data (SMM)

- However: formidable effort needed to fulfill the vision of automatic extraction of complete entity relationship models

Further Research
  - 1. Assessing more measures for relevance ranking
  - 2. Evaluating more methods to combine single terms to n-gram
  - 3. automatically classify entities and relationships to classes / types
  - 4. Incorporate human knowledge


THANK
YOU