Research@MICS
Media Computer Science Passau

Prof. Dr. Michael Granitzer

work from

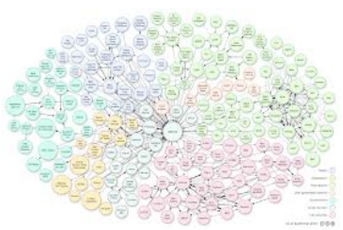
Dr. Christin Seifert (Habil cand.)

Stefan Zwicklbauer (PhD cand.)
Jörg Schlötterer (PhD cand.)
Johannes Jurgovsky (PhD cand.)
Sebastian Bayerl (PhD cand.)
Stefan John (PhD cand.)
Albin Petit (PhD cand.)

Lisa Wagner (MsC cand.)
Alexander Treml (MsC cand.)
Stefan Kunz (MsC cand.)

MICS in a Nutshell

The Data



Fields

- (Text) Data Mining
- Applied Machine Learning
- Information Retrieval/NLP
- Visual Analytics / HCI
- Semantic Web
- Social Networks

Projects

1. EEXCESS - Personalised, privacy preserving federated recommendations for cultural and scientific content (EU FP7)
2. CODE (finished) - Fact Extraction and Enrichment from Scientific Articles (EU FP7)
3. MICO - Media in Context – Cross-Media Recommendations and Semantic Representation (EU FP7)
4. BODA - Big and Open Data for SME's (Bayern)
5. mirKUL - Interaktive Multimedia Videos (BMBF)
6. Industrial Research Project – Credit Card Fraud Detection

Project 1: EEXCESS

FP 7 IP, 10 Partners, Scientific Coordinator, ongoing

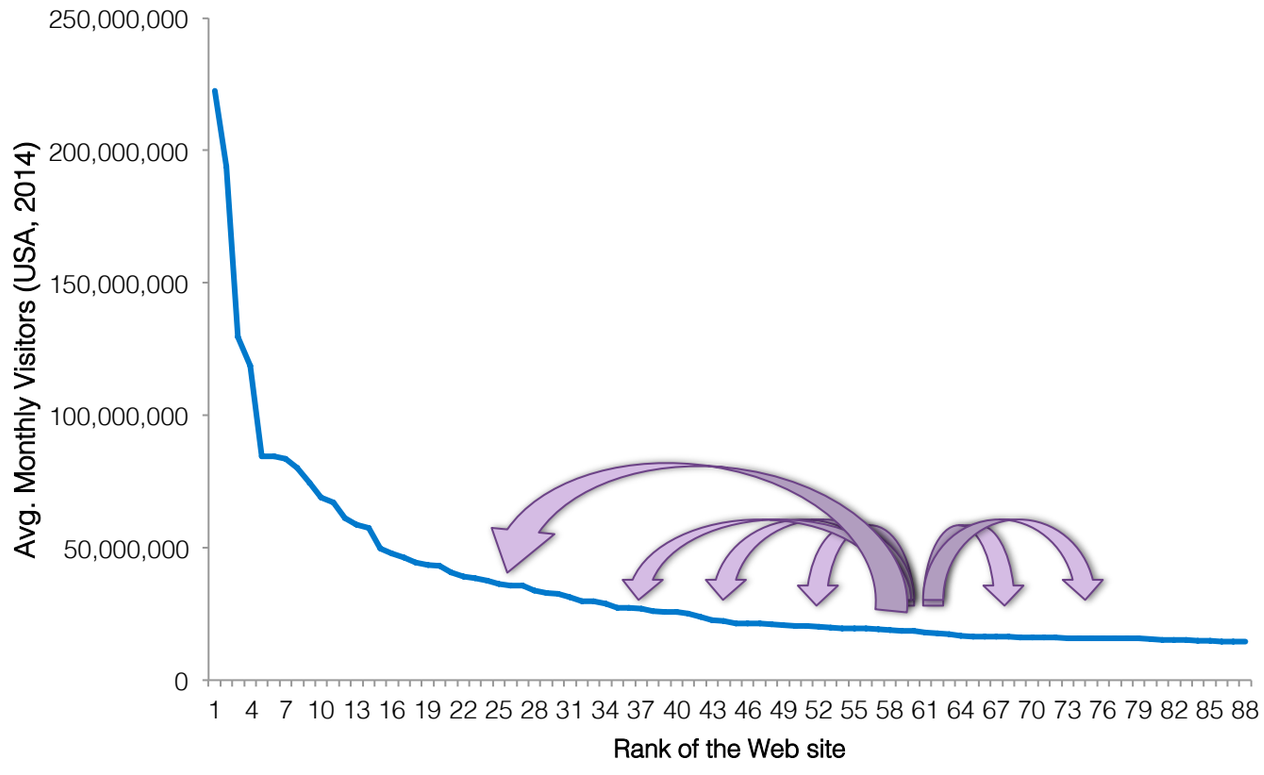
http://eexcess.eu/

EEXCESS - Goal

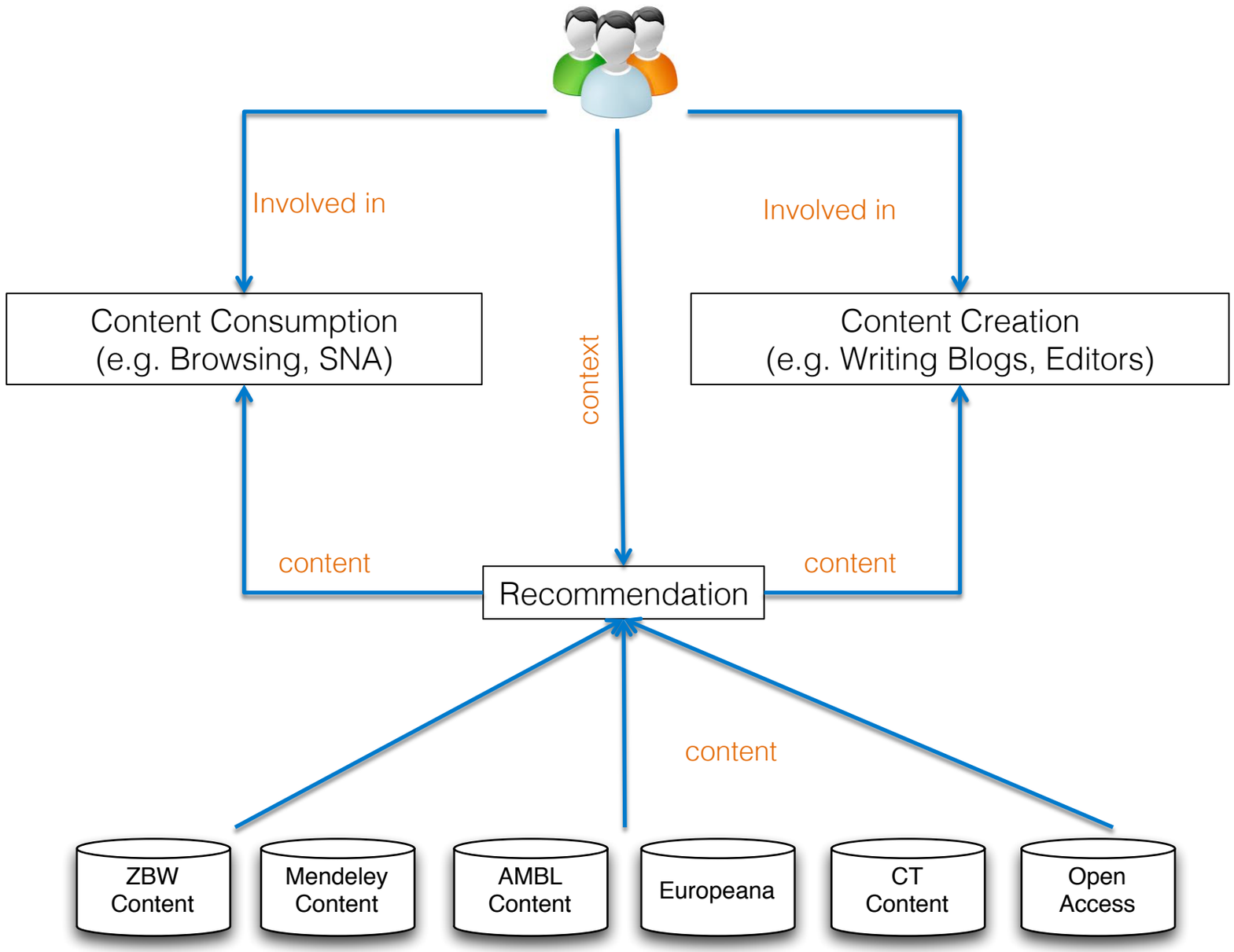
„Contextualised, privacy-preserving access to scientific and cultural long-tail content “

Inject cultural and scientific content into existing web channels

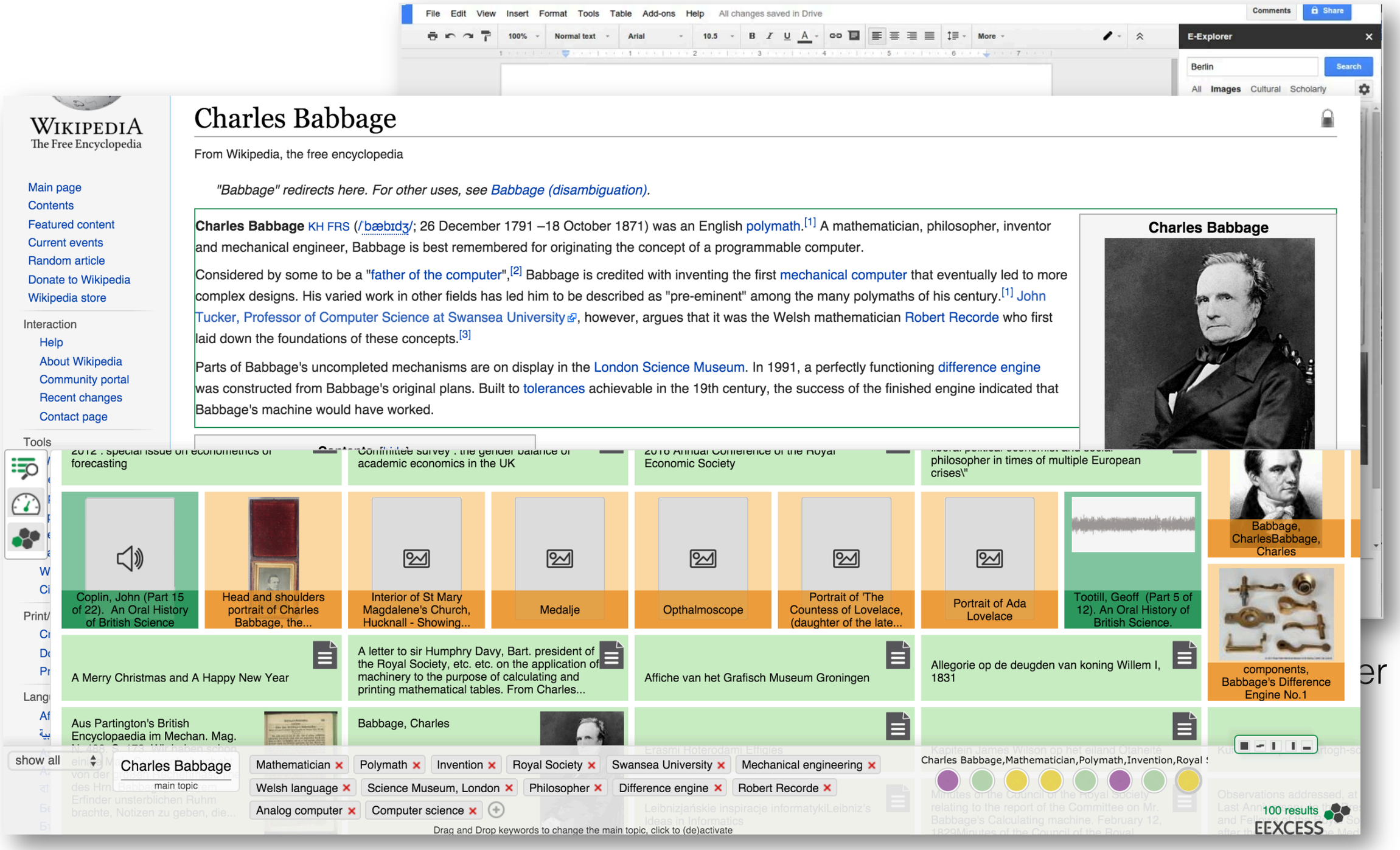
- Websites (Wikipedia, etc.)
- CMS/LMS
- Social media channels (Twitter, etc.)
- Content Consumption and Content Creation Processes



EEXCESS Solution



EEXCESS – Our Solution



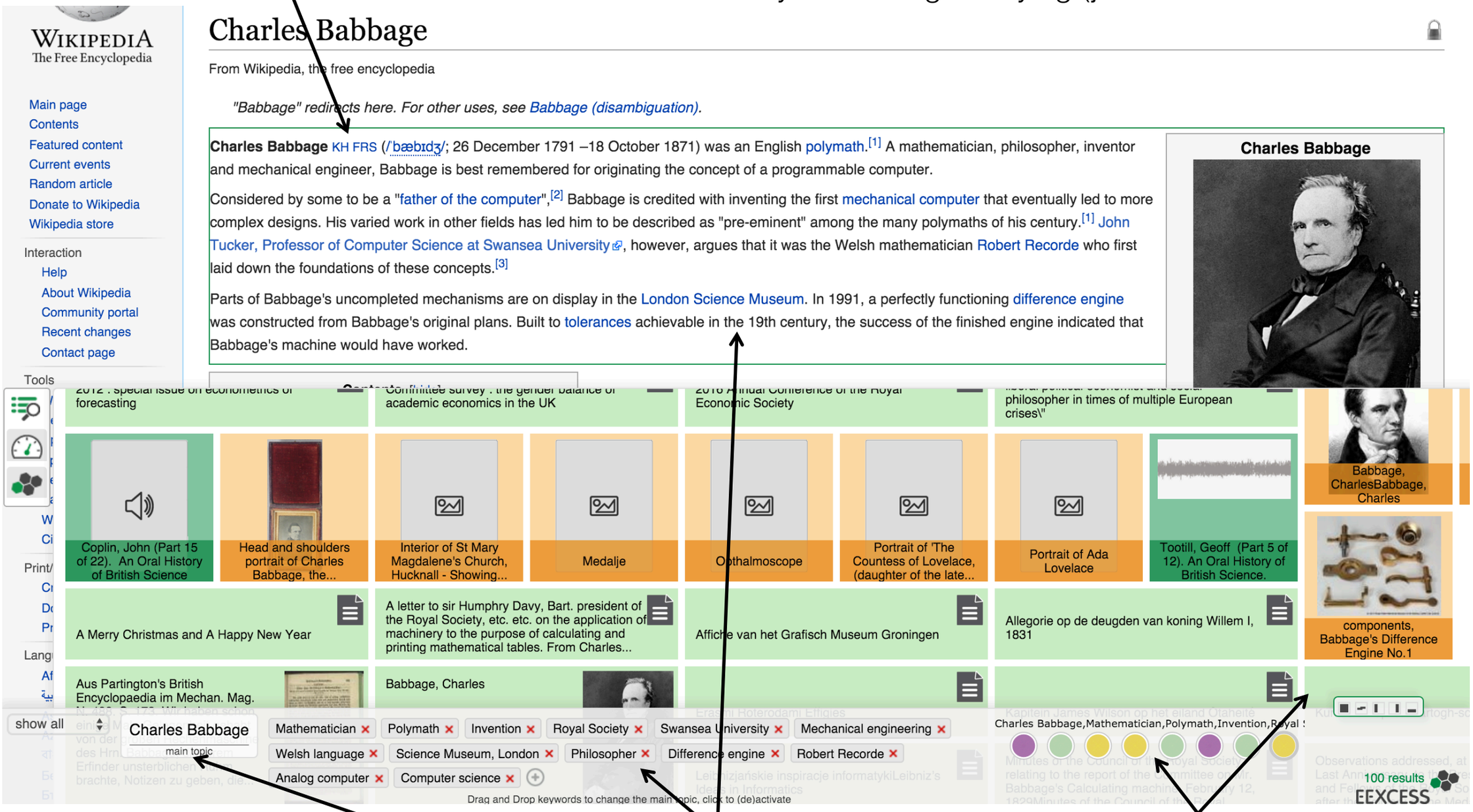
Install: go to Chrome Webstore -> Search for EEXCESS

EEXCESS – Our Research Questions

Paragraph Decomposition and Detection (heuristic)

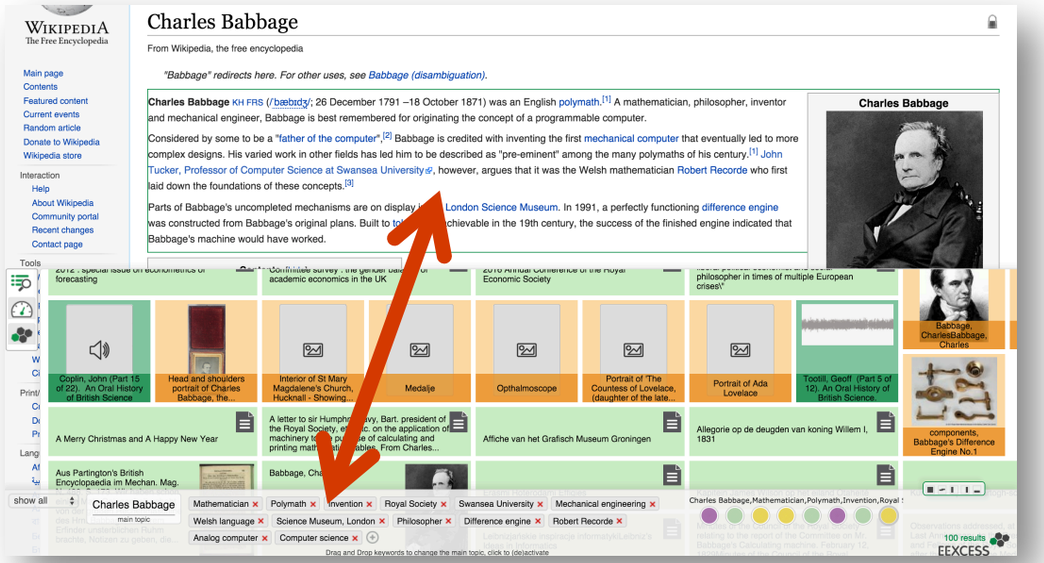
Privacy Preservation:

- Resource Efficient Text Mining on the Client
- Privacy Preserving Querying (joint PhD Student with INSA Lyon)



Paragraph Summarization + Query Generation

Visual Query Navigation and Search Result Visualisation (+ to ns of UI stuff)



Project 1: EEXCESS

Paragraph Summarization and Query Construction

Includes: Prediction Queries using CRF's; New Test Data Sets; A new state-of-the-art Entity Disambiguation approach

EEXCESS: Paragraph Summarization and Query Generation

Research Question
Can we predict manual queries from a given paragraph?

Experiment 1

- Given a text selection, train a linear chain CRF to annotate a word in the selection as query term/not query term.
- Evaluate on collected ground truth data
  - Browser-plugin on Wikipedia
  - Selection + manual queries + ratings + tasks
  - 2499 text selection query pairs

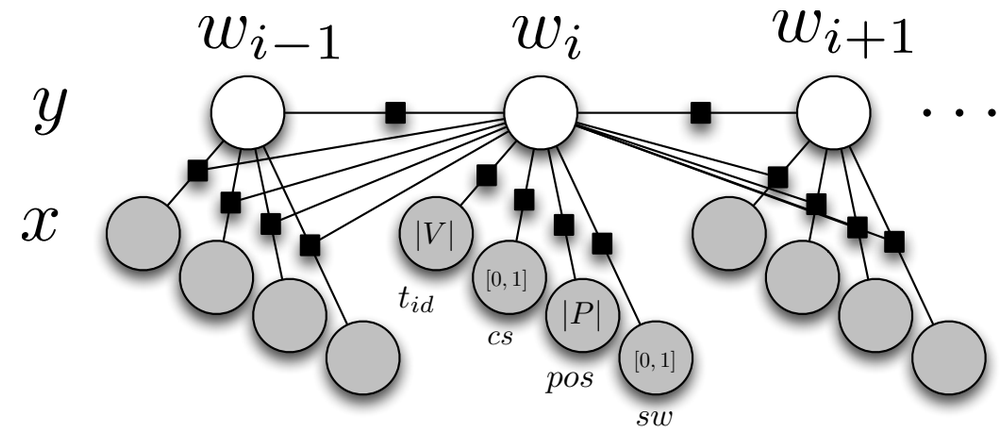
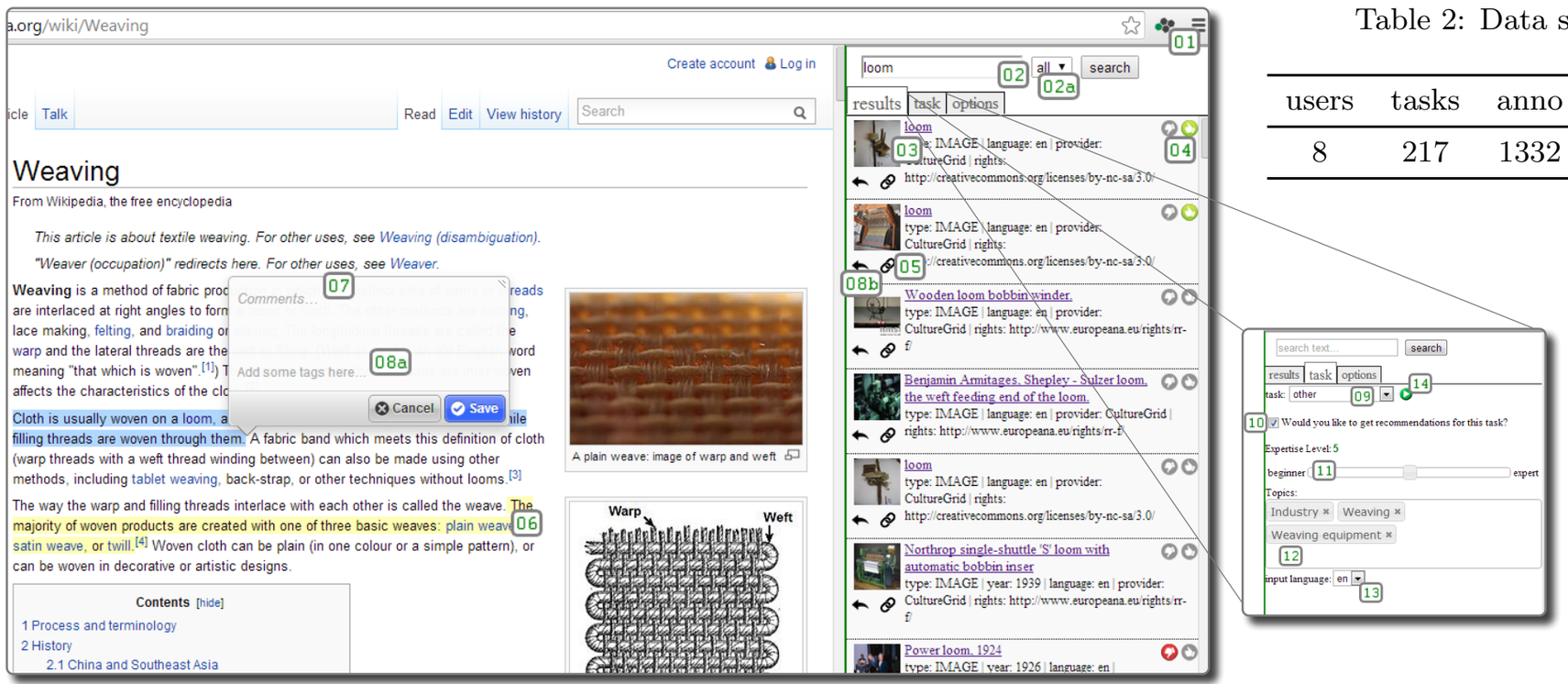


Table 2: Data set overview. Time in minutes.

users	tasks	anno	queries	views	ratings	time
8	217	1332	4562	3267	15043	10252



EEXCESS: Paragraph Summarization and Query Generation

Research Question
Can we predict manual queries from a given paragraph?

Experiment 1 – Results

In Accuracy (%) + Baseline (unbalanced dataset)

Splits:

- 10-folded
- Transfer model across
  - Users
  - Tasks

		feature set			trivial	
		i, c, t	i, t	c, t	rejector	acceptor
users	mean	76	77	75	51	49
	SD	15	15	18	35	35
tasks	mean	82	83	82	71	29
	SD	6	6	7	8	8
10-fold	mean	89	88	84	71	29
	SD	1	2	1	2	2

i - the identity of a term, i.e. the term itself
c - whether the term begins with upper- or lowercase
t - POS tag

Conclusion

- Model stable across tasks, not users
- Improvement over (useless) baseline
- Easy task comparable to noun phrase detection in sentences

Seifert, C.; Schlötterer, J. & Granitzer, M. Towards a Feature-Rich Data Set for Personalized Access to Long-Tail Content Proceedings ACM Symposium on Applied Computing, ACM, 2015; http://purl.org/eexcess/datasets/umlt

EEXCESS: Paragraph Summarization and Query Generation

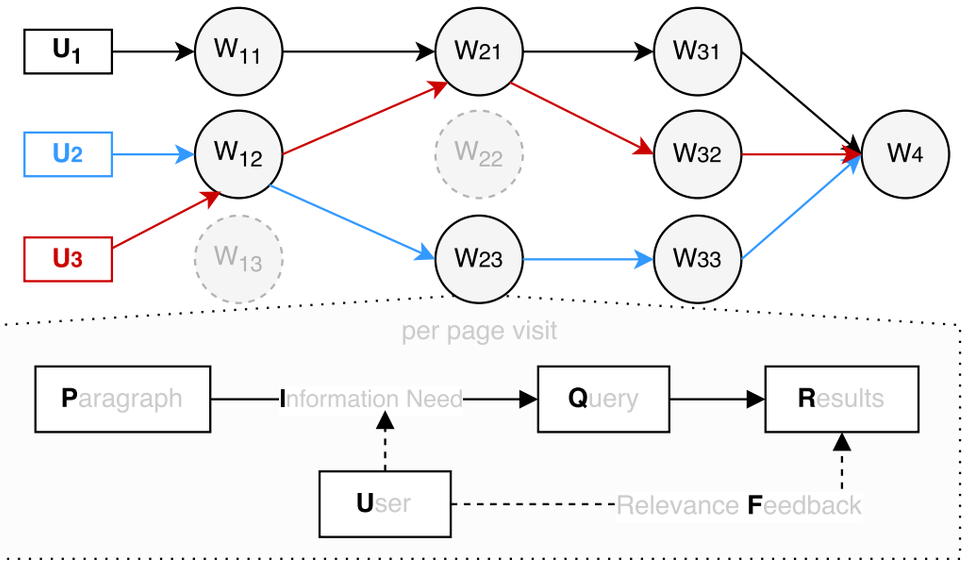
Research Question
Can we predict manual queries from a given paragraph?

Experiment 2 – Extend to Paragraphs (currently ongoing)

- Extend query to paragraph
- Challenges
  - Summarize paragraph through semantic features
  - One paragraph contains multiple, potentially correct search queries depending on the individual users preference
  - Models vary across users
- Gather manual test data according to Marchionini & White's Information Seeking Model on Wikipedia Pages

Preliminary results of linear Chain CRF

- 90% accuracy with 85% baseline (negative predictor)
- 0.64 precision, 0.44 recall
- Vocabulary has strong influence



#users	e	p	-	o	+	++	total
24	1	1	1086	377	610	123	2196
20	1	0	1112	416	479	124	2131
17	0	0	1076	383	532	97	2088
16	0	1	1062	321	608	122	2113
77			4336	1497	2229	466	8528

e - indicator of explanation feature
p - indicator of personalization feature

EEXCESS: Paragraph Summarization and Query Generation

Research Question
Can we utilize semantic resources (e.g. DBpedia) to improve query generation?

Current Status

- Work mainly done in the field of entity linking/entity disambiguation
- Evaluation in terms of query generation still pending

Now: Focus on (collective) Entity Disambiguation using Word2Vec based Semantic Embeddings

1. What is Word2Vec?
2. How to use Word2Vec for Entity Disambiguation?
3. Does it help? Yes ☺
   - Increase robustness of entity disambiguation approaches
   - New state of the art approach on most data sets
   - Can be made KnowledgeBase Agnostic, i.e. a general preprocessing step for text AND semantic resources (i.e. ontologies/thesauri)

Word2Vec: Neural Network based Language Model in a Nutshell

Goal: Estimate $P(w_t | w_{t-1})$ using a d-dimensional vector v_i per word w_i (aka Semantic Embeddings) using the softmax function:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^\top v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^\top v_{w_I}\right)}$$

Training:

- Predict the context of a word in a sliding word window
- Optimize Parameters using Stochastic Gradient Descent

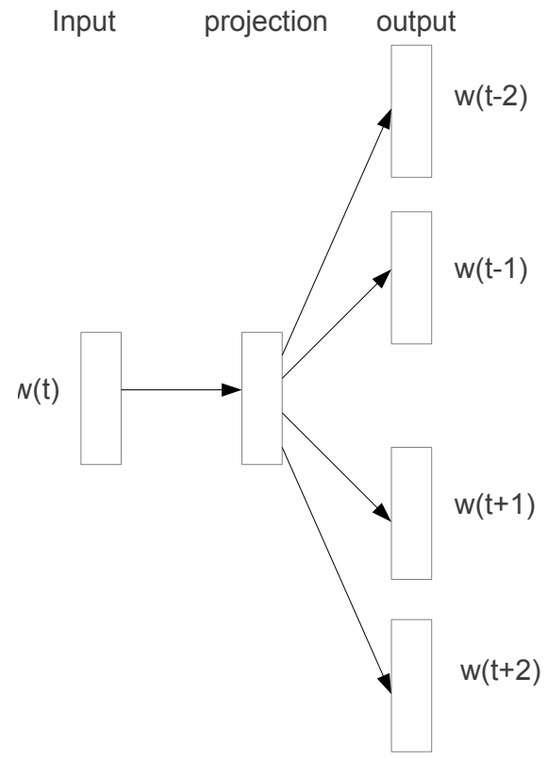
$$p(W \mid w_i) = \frac{\sigma(V_i \cdot H)}{\sum_{j=1..|w|} \sigma(V_j \cdot H)}$$

V as embedding matrix $|w| \times d$

H as hidden layer $d \times |w|$

σ as non-linear activation function (i.e. sigmoid)

V as vector of all Words



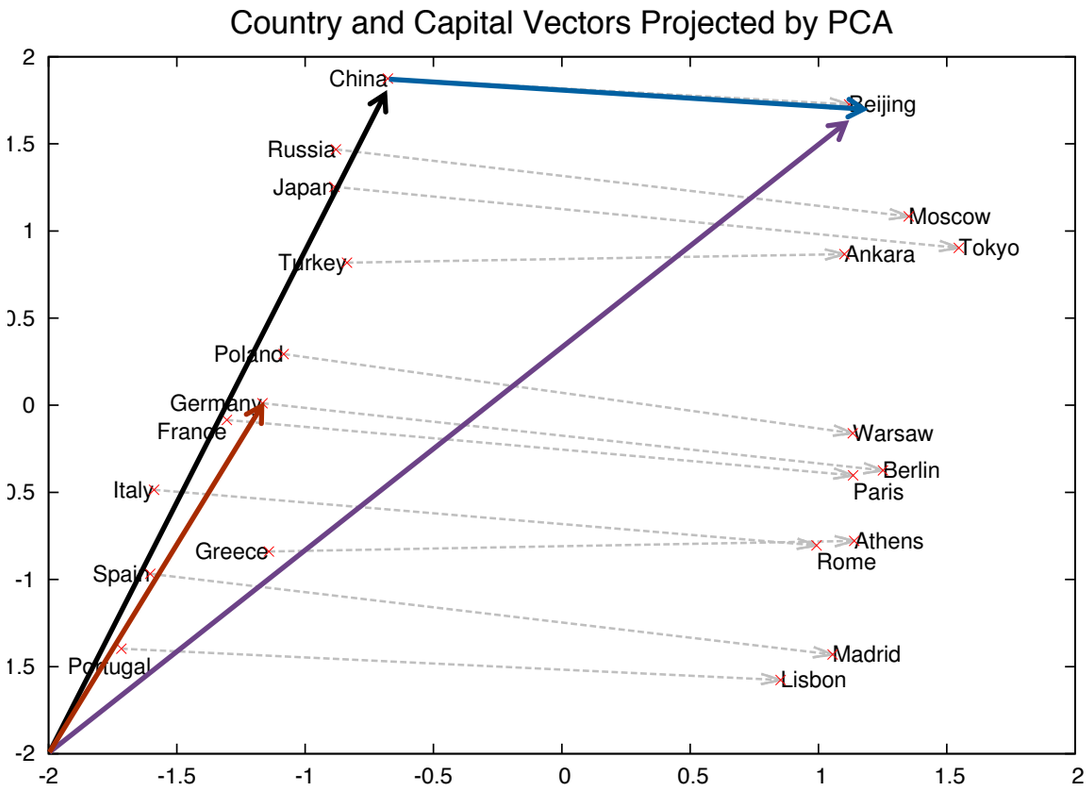
- Noise Contrastive Estimation for Speeding up Softmax

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]$$

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

Word2Vec: Neural Network based Language Model in a Nutshell

Results:



China - Beijing + Germany = near(Berlin)

~72% on the Word Analogy Reasoning Task

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

EEXCESS: Paragraph Summarization and Query Generation

Research Question

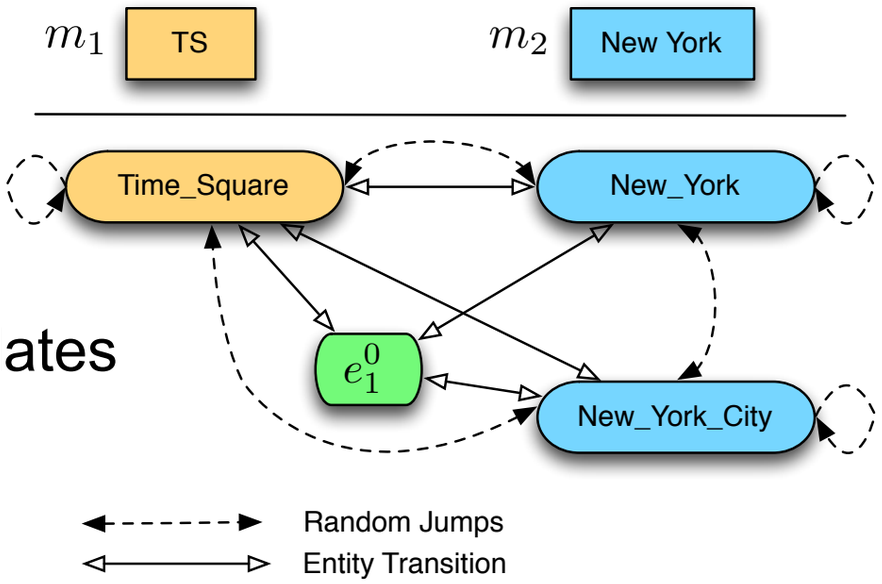
How can we accurately map surface forms to entities in a knowledge base?

A 29-year-old previously healthy woman, a doctor, was diagnosed with remitting relapsing multiple sclerosis after fulfilling McDonald's criteria for the diagnosis of definite multiple sclerosis.

Multiple_Sclerosis
Mc_Donalds (food chain)
Mc_Donalds (person)

Approach

1. Preprocessing: Create semantic embedding per entity
   - Word2Vec: word-level embedding
   - Doc2Vec: paragraph level embedding
2. Generate candidate entities E through index lookup
3. Create a directed, weighted entity graph from all candidates
   I. The weight is based on the mean of the
      - word-level embedding similarity, i.e. *word2vec(e_1,e_2)*
      - Paragraph-level embedding of entity with the surrounding context in the text of its surface form, i.e. *doc2vec(e_1,m_1)*
   II. The topic node e^0_1 summarizes already disambiguated entities
4. Solve page rank and take highest ranked entity per surface form (or abstain)



EEXCESS: Paragraph Summarization and Query Generation

Research Question

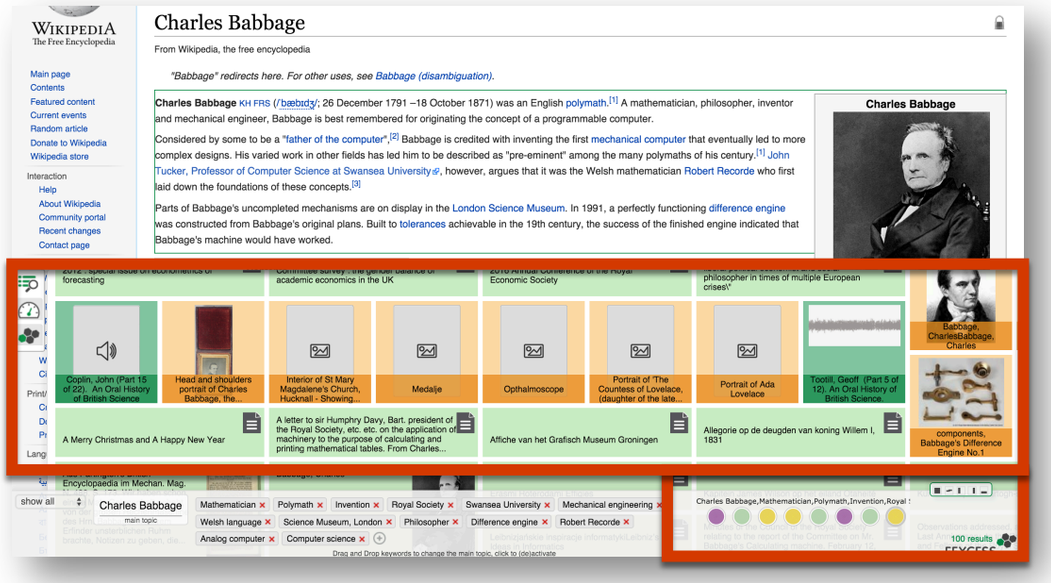
How can we accurately map surface forms to entities in a knowledge base?

Results (F1 Measure)

Data Set	Topic	#Doc.	#Ent.	Ent./Doc.	Annotation
ACE2004	news	57	253	4.44	voting
AIDA TestB	news/web	231	4458	19.40	voting
AQUAINT	news	50	727	14.50	voting
DBpedia Spot.	news	58	330	5.69	domain experts
MSNBC	news	20	658	32.90	domain experts
N3-Reuters	news	128	621	4.85	voting
IITB	news/web	103	11245	109.01	domain experts
Microposts	tweets	1165	1440	1.24	domain experts
N3-RSS 500	RSS-feeds	500	1000	2.00	domain experts

Data Set	DoSeR	DoSeR (-Doc2Vec)	PriorProb	Wikifier	Spotlight	AIDA	Babelfy	WAT
ACE2004	**0.907**	0.872	0.831	0.834	0.713	0.815	0.561	0.800
AIDA/CONLL-TestB	0.784	0.754	0.661	0.777	0.593	0.774	0.592	**0.843**
AQUAINT	0.842	0.842	0.820	**0.862**	0.713	0.532	0.652	0.768
DBpedia Spotlight	**0.810**	0.775	0.745	0.797	0.789	0.508	0.522	0.652
MSNBC	**0.911**	0.876	0.711	0.851	0.511	0.782	0.607	0.777
N3-Reuters	**0.850**	0.810	0.700	0.703	0.577	0.596	0.534	0.644
IITB	0.741	0.738	0.711	**0.766**	0.447	0.270	0.470	0.611
Microposts-2014 Test	**0.750**	0.704	0.630	0.586	0.453	0.453	0.473	0.595
N3 RSS-500	**0.751**	0.713	0.678	0.732	0.622	0.716	0.630	0.682
Average	**0.816**	0.787	0.726	0.768	0.602	0.605	0.560	0.708

→ Semantic Embeddings significantly improve the accuracy



Project 1: EEXCESS

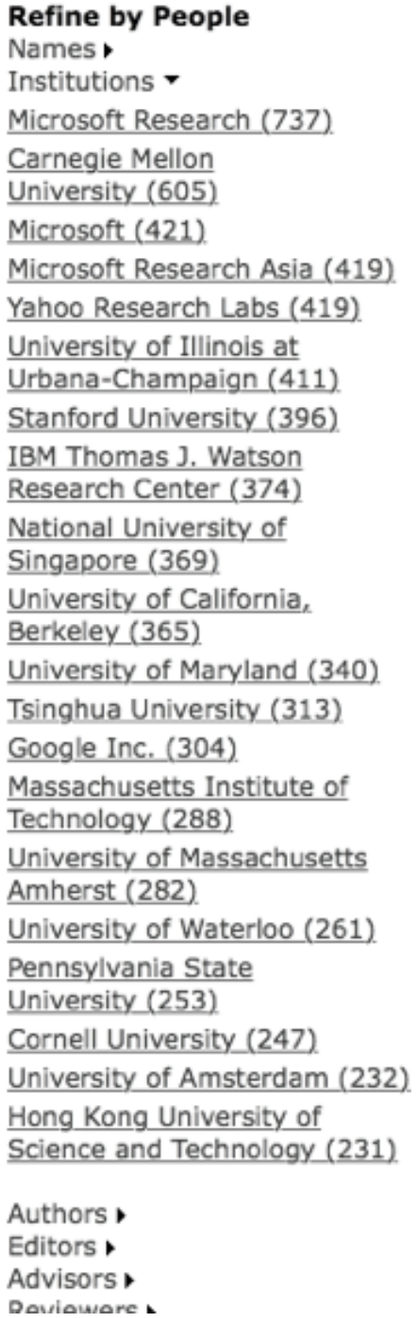
Visualising Search Results and Query Navigation Support

Includes: Two new visualisation: FacetScape and Query Crumbs

EEXCESS: Visualising Search Results

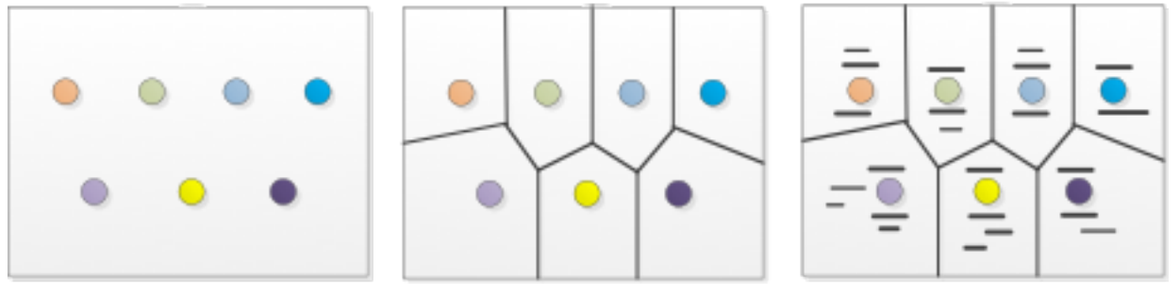
Research Question

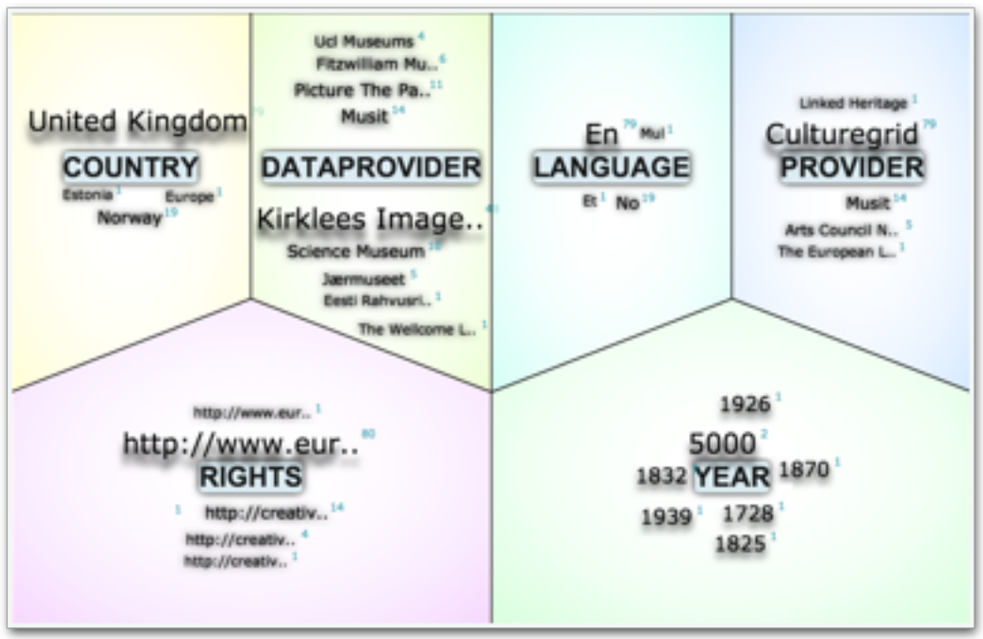
Can we provide a better overview over facets of a query?

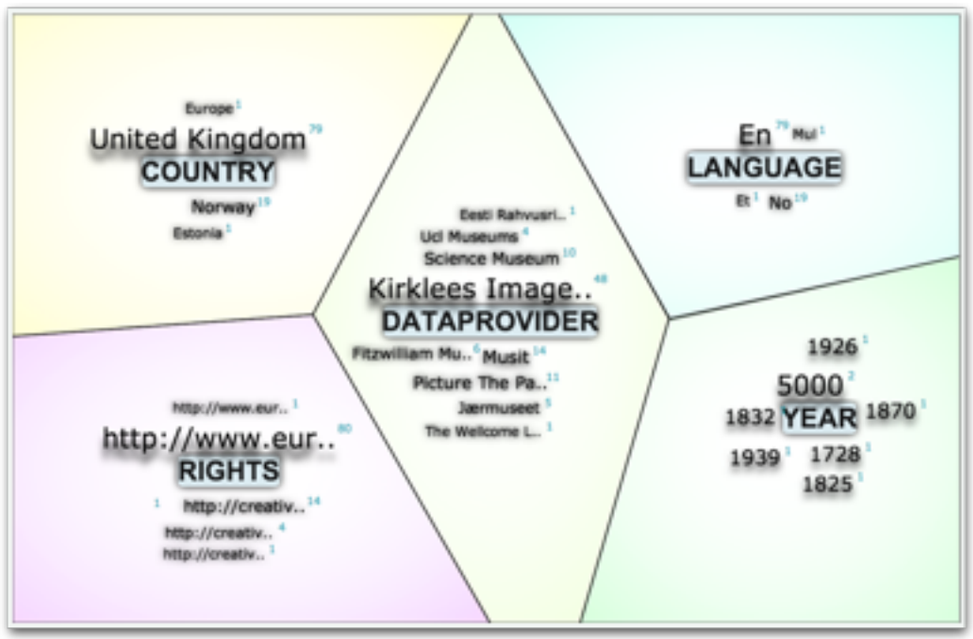


Approach

1. AW Power Voronoi
2. Tag Layout
3. Interactions (Zoom, Remove, Filter)
4. Comparative user eval





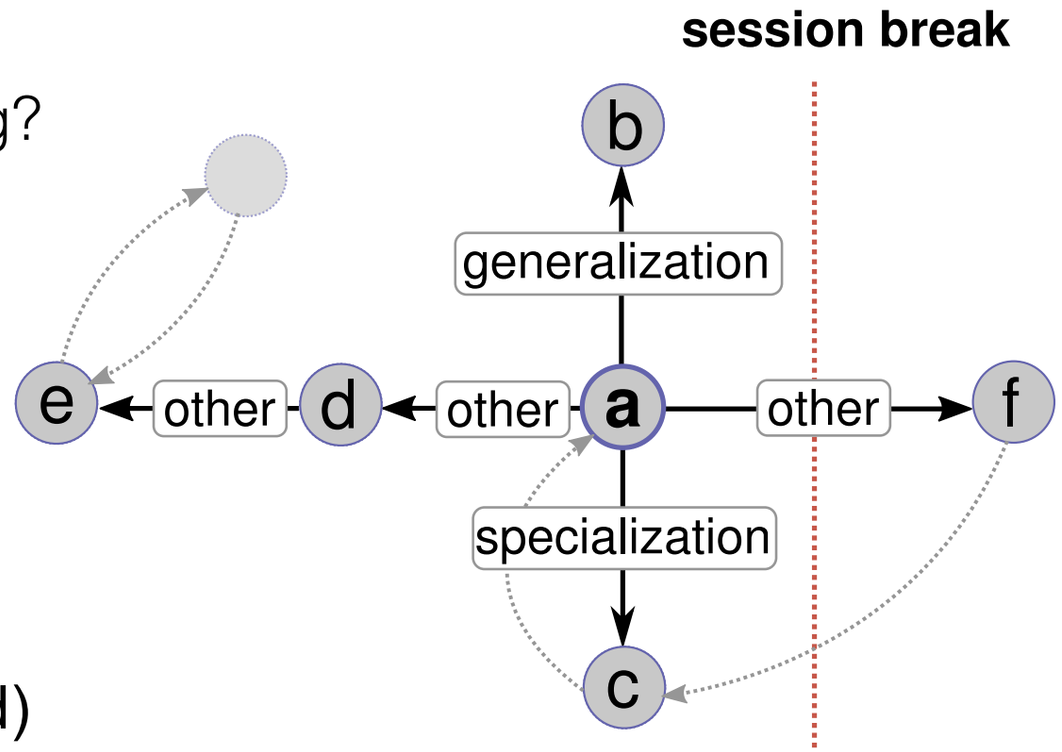


Christin Seifert, Johannes Jurgovsky, and Michael Granitzer. “FacetScape: A Visualization for Exploring the Search Space”. In: *Proc. International Conference on Information Visualization (IV)*. BEST PAPER AWARD

EEXCESS: Support Query Navigation

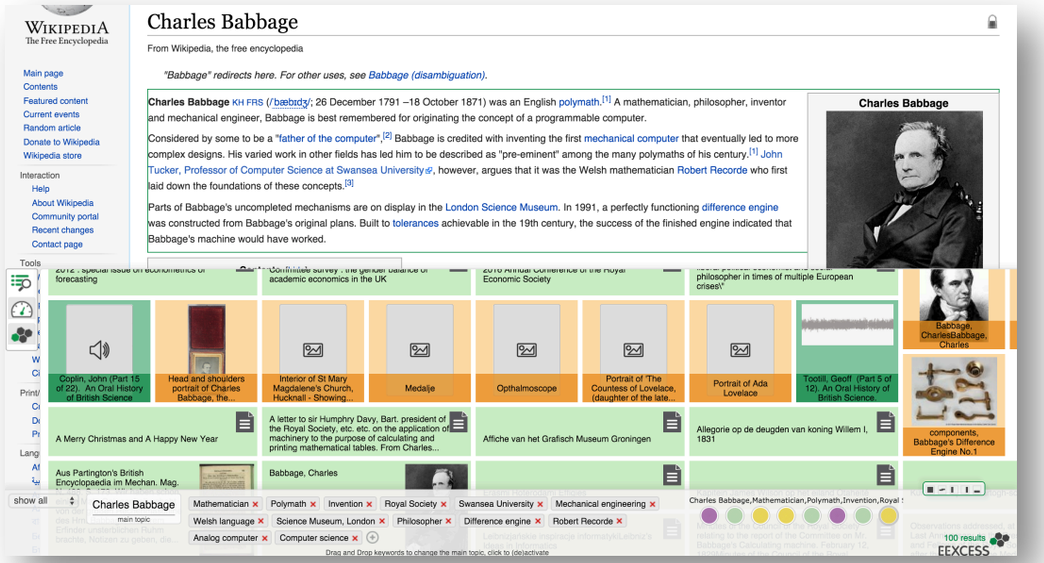
Research Question

How can we support re-querying and backtracking?



Approach

1. Developed query history model
2. Bread-crumb like mini-visualisation
   - Query similarity (binary, percentage, detailed)
3. Formative user evaluation
   - Understandable without explanation
   - Usable without explanation
   - Uptake (60% voluntarily choose to use the vis)



Project 1: EEXCESS

Resource Efficient Text Mining on Web-Clients

Includes: Analyse the reduction of word embeddings
Excludes: software development

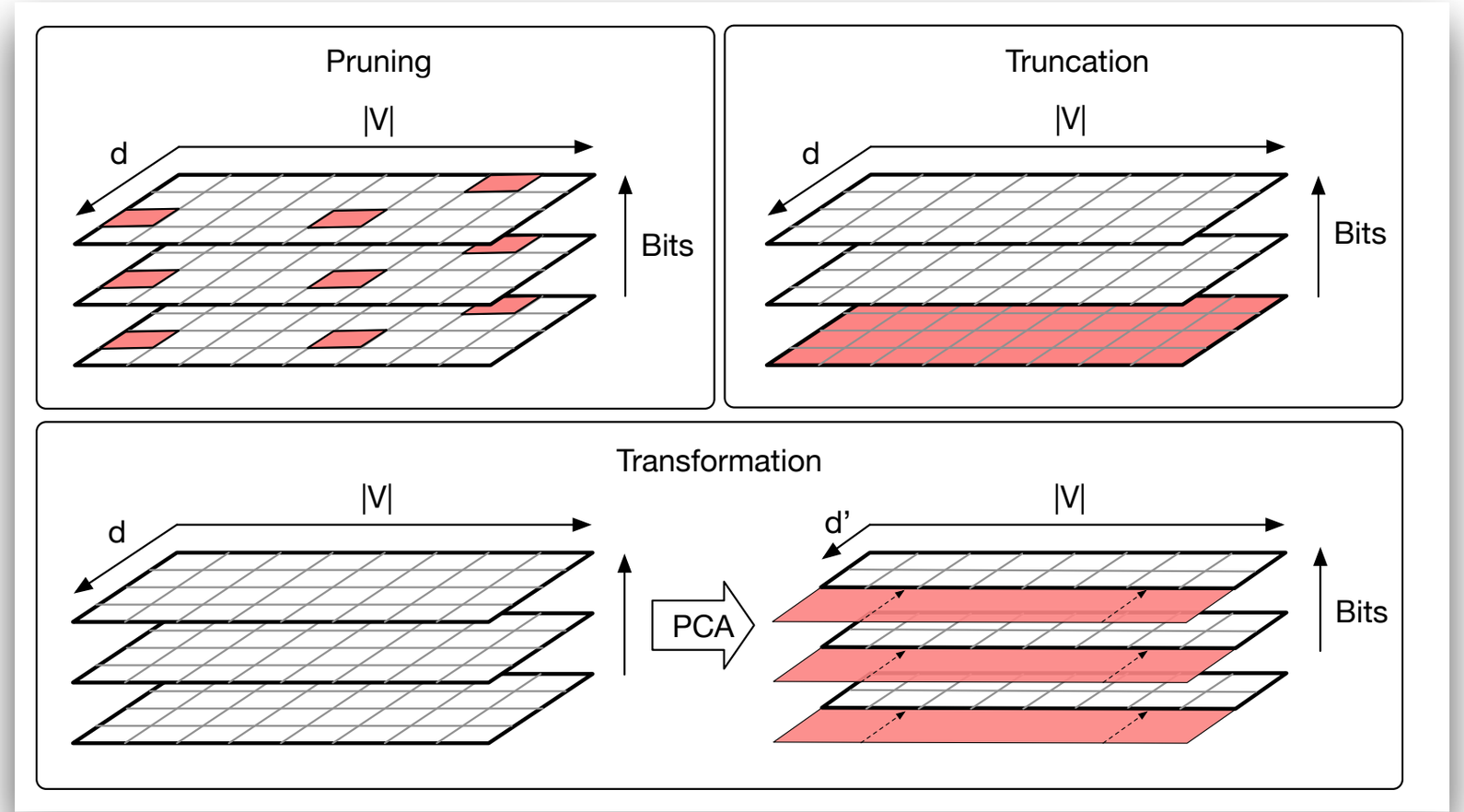
EEXCESS: Resource Efficient Text Mining on Web-Clients

Research Question

Given the success of word-embeddings (i.e. word2vec), can we reduce their size in order to use them in low-memory environments (i.e. Java Script client)

Approach

- Analyse the effect of different pruning strategies on word embeddings



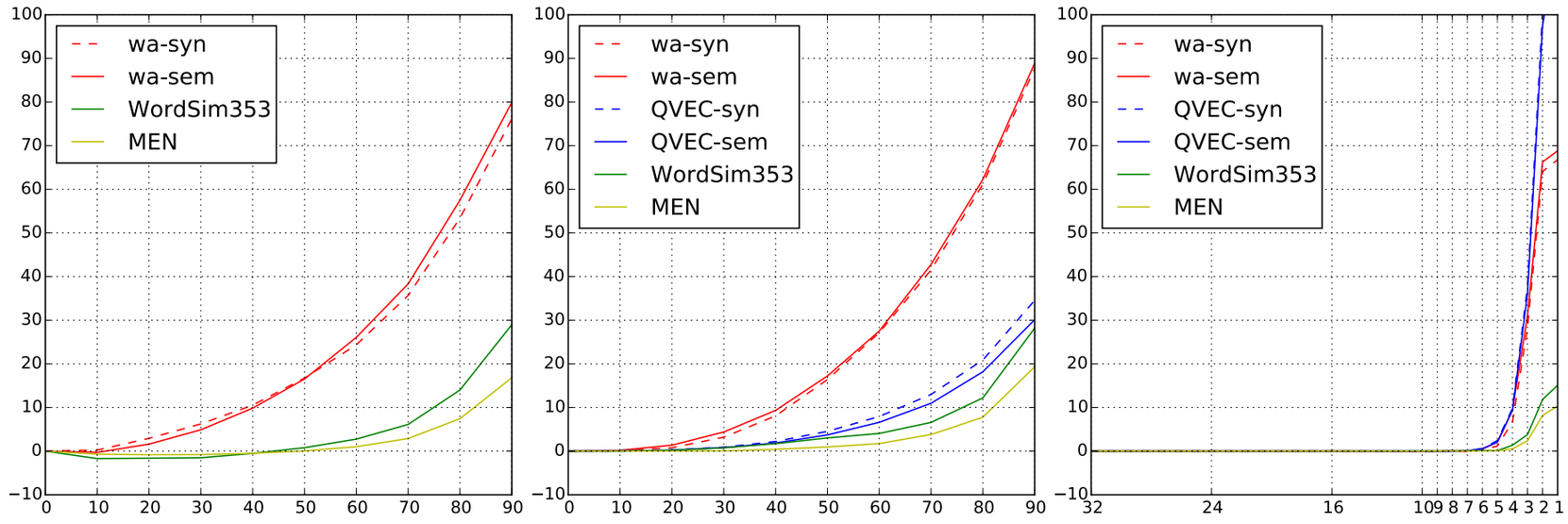
Jurgovsky, J.; Granitzer, M. & Seifert, C. Evaluating Memory Efficiency and Robustness of Word Embedding Proc. of European Conference on Information Retrieval (ECIR), 2016

EEXCESS: Resource Efficient Text Mining on Web-Clients

Research Question

Given the success of word-embeddings (i.e. word2vec), can we reduce their size in order to use them in low-memory environments (i.e. Java Script client)

Results



(a) PCA (b) Pruning (c) Bit-Truncation

Fig. 2: Mean relative loss of embeddings after (a) PCA: percentage of removed dimensions, (b) Pruning: percentage of removed parameters and (c) Bit-Truncation: remaining Bits. Scores on the QVEC datasets are not shown for PCA since they are not comparable across different word vector sizes.

Datasets/Tasks:

- wa, and QVEC for word analogy task
- WordSim353 and MEN for word similarity

Averaged over dimensionalities are 50, 100, 150, 300, 500

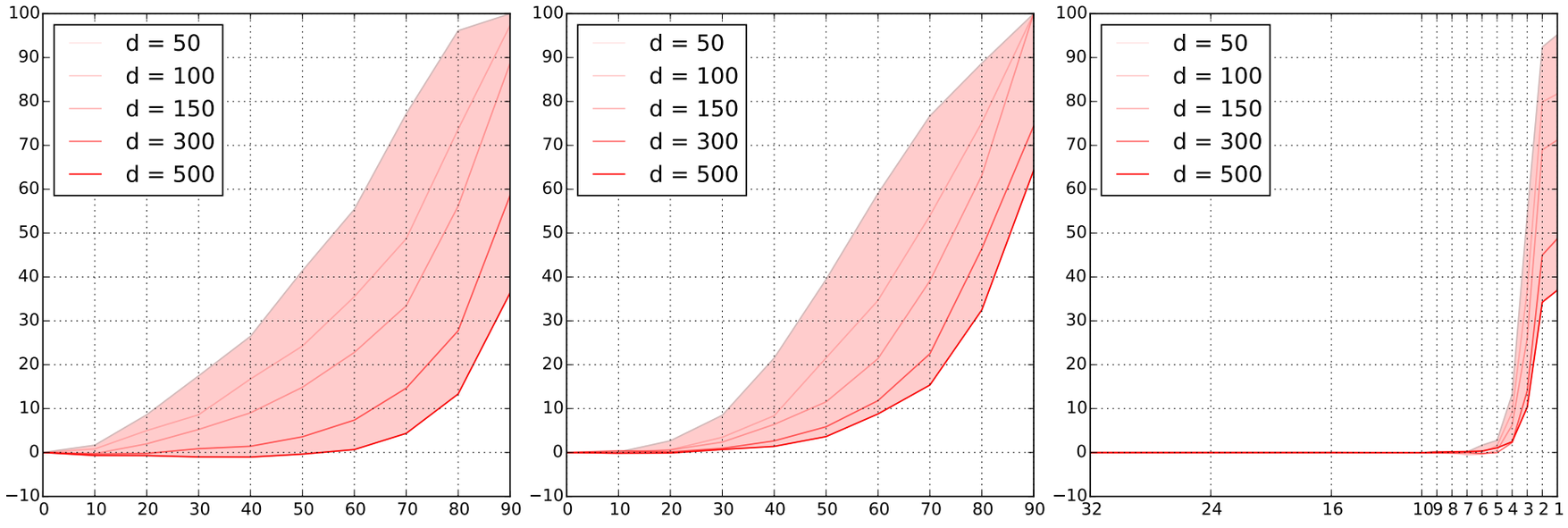
Jurgovsky, J.; Granitzer, M. & Seifert, C. Evaluating Memory Efficiency and Robustness of Word Embedding Proc. of European Conference on Information Retrieval (ECIR), 2016

EEXCESS: Resource Efficient Text Mining on Web-Clients

Research Question

Given the success of word-embeddings (i.e. word2vec), can we reduce their size in order to use them in low-memory environments (i.e. Java Script client)

Results for different dimensionalities



(a) PCA (b) Pruning (c) Truncation

Fig. 3: Relative loss of embeddings on the syntactic word analogy dataset (wa-syn) after PCA (a), Pruning (b) and Bit-Truncation (c).

Jurgovsky, J.; Granitzer, M. & Seifert, C. Evaluating Memory Efficiency and Robustness of Word Embedding Proc. of European Conference on Information Retrieval (ECIR), 2016

EEXCESS: Resource Efficient Text Mining on Web-Clients

Research Question
Given the success of word-embeddings (i.e. word2vec), can we reduce their size in order to use them in low-memory environments (i.e. Java Script client)

Same idea, but different Learning Approach and Task

- Train Recursive Autoencoder and Convolutional Neural Networks for Sentiment Classification
  - Note: Embeddings are trained implicitly
  - Logistic Regression Layer as Decision Layer
- Prune the embeddings by removing weights with low absolute value
- Observe changes in accuracy

→ Accuracy drop after
  → 90% for CNN
  → 80% for RAE

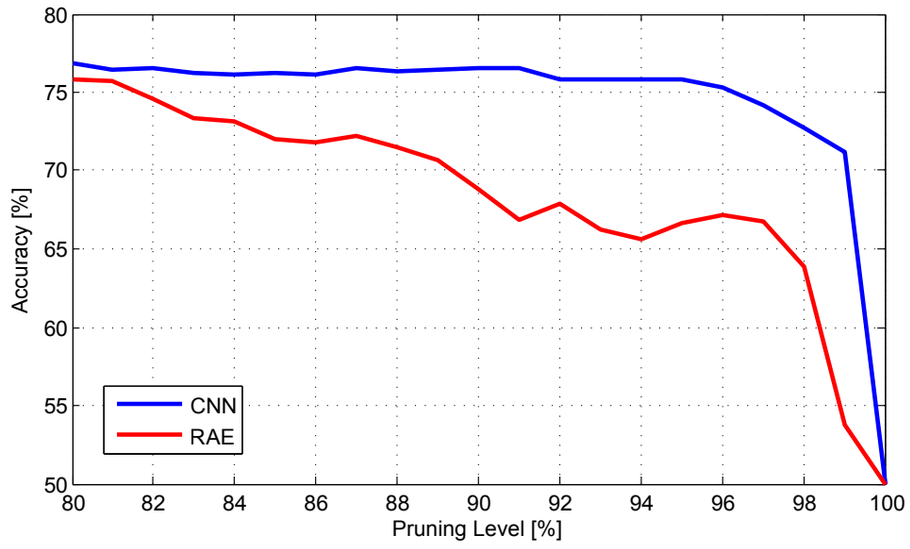
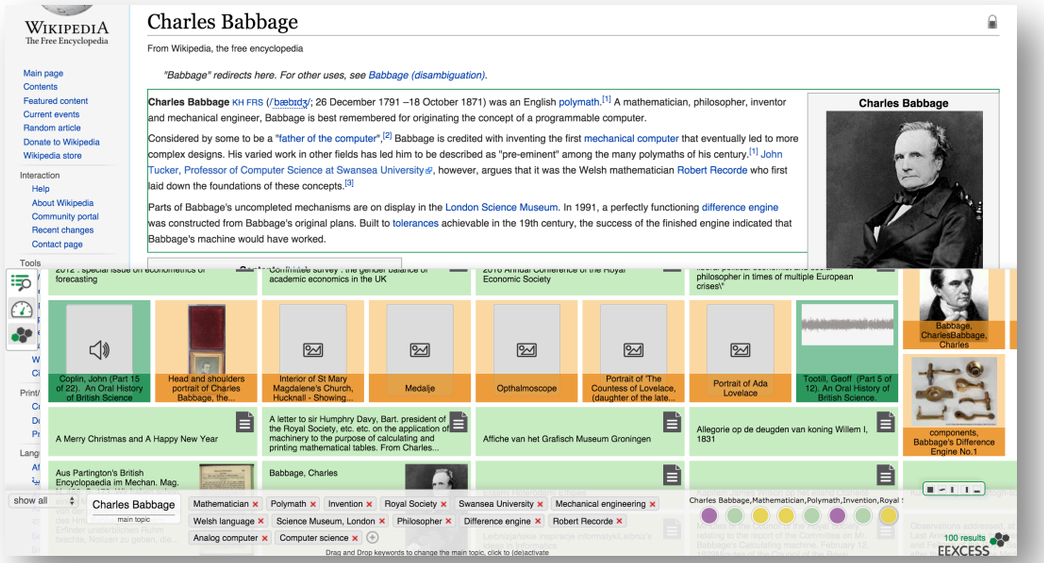


Fig. 1. Binary sentiment polarity classification accuracy of Logistic Regression. Underlying sentence representations were extracted with pruned versions of CNN or RAE at different pruning levels.

J. Jurgovsky and M. Granitzer, “Comparing Recursive Autoencoder and Convolutional Network for Phrase-Level Sentiment Polarity Classification,” in Natural Language Processing and Information Systems, C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais, Eds. Springer International Publishing, 2015, pp. 160–166.



Project 1: EEXCESS

Privacy Preserving Querying

Only a brief overview. Joint work with INSA Lyon (and most parts have been developed there)

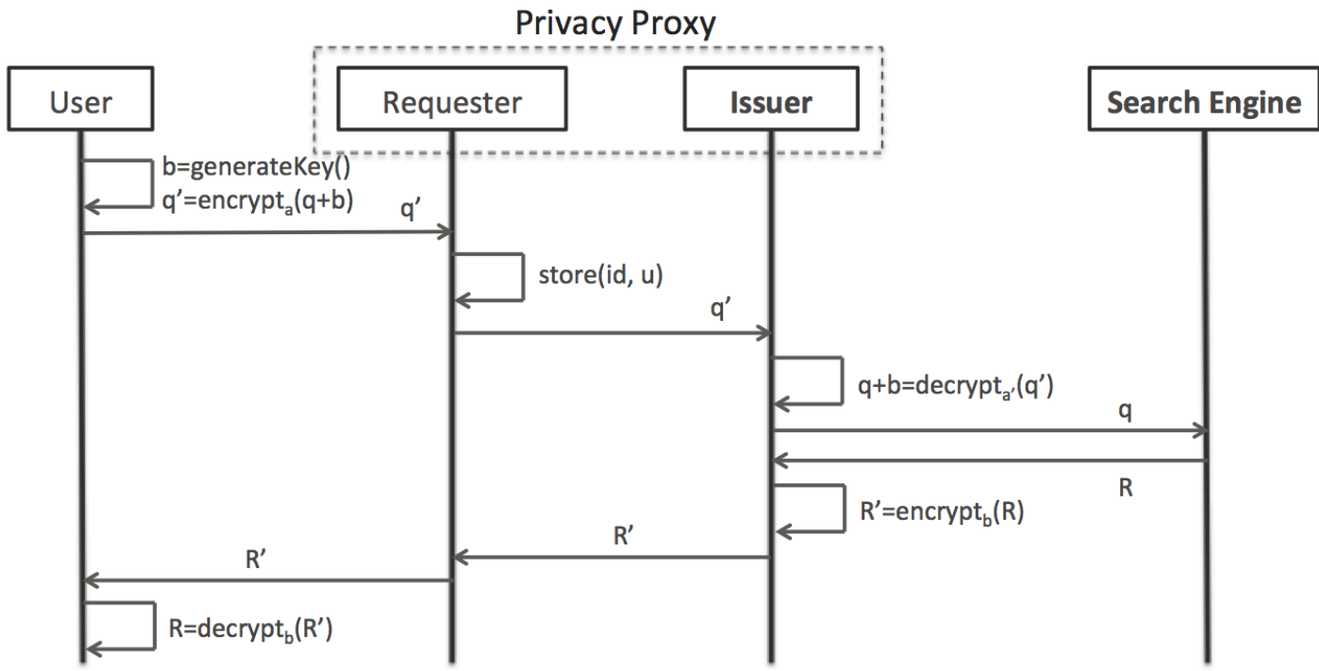
EEXCESS: Privacy Preserving Querying

Research Question
How can we avoid that a search engine profiles their users?

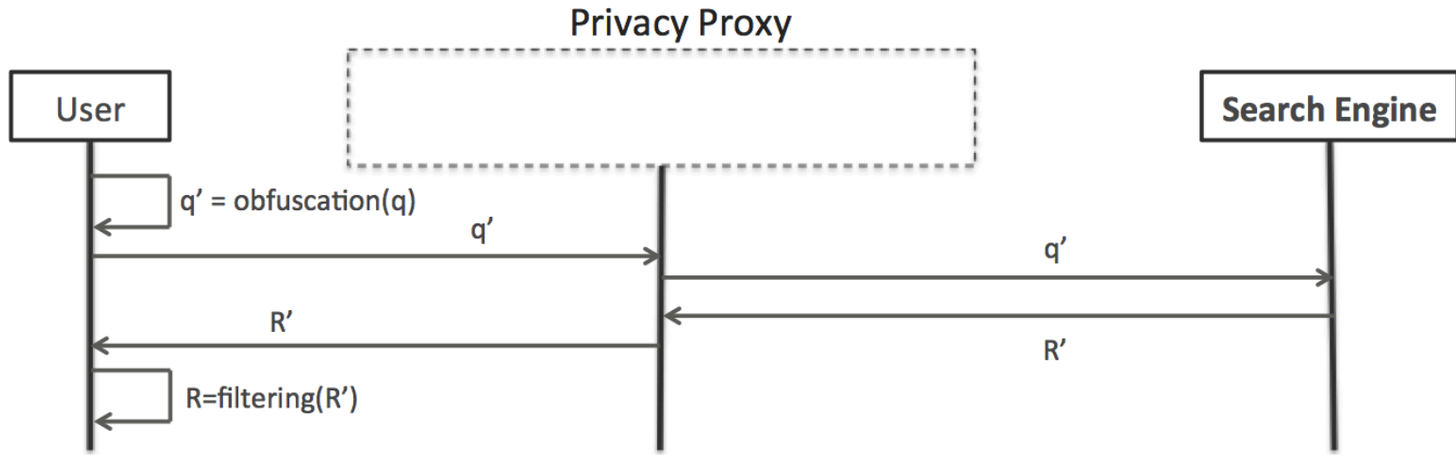
Approach

Develop a new protocol based realized through an proxy server consisting of two components

1. Unlinkability: Avoid linking the query itself to the user id
   - Encrypt query + key to encrypt results
   - User-id and encrypted query processed by different services



2. Indistinguishability: Hide the true query from the search engine
   - Assume Boolean Queries
   - Generate fake queries
   - OR fake queries with real query
   - Filter OR results based on title



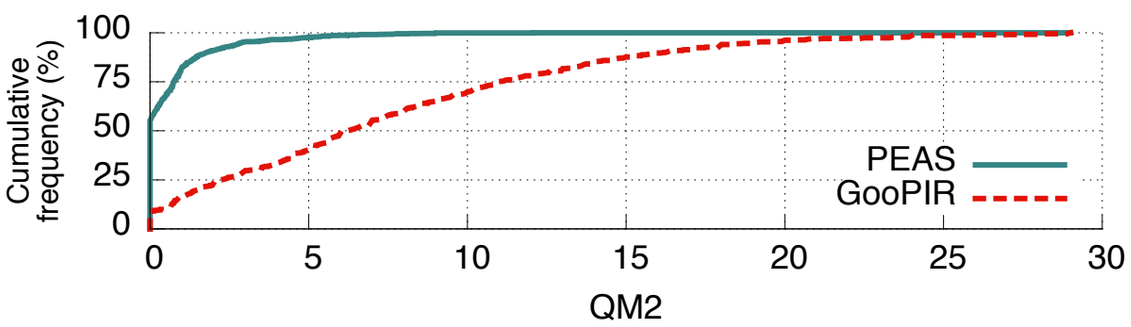
Petit, A.; Cerqueus, T.; Ben Mokhtar, S.; Brunie, L.; Kosch, H., "PEAS: Private, Efficient and Accurate Web Search," in *Trustcom/BigDataSE/ISPA, 2015 IEEE* , vol.1, no., pp.571-580, 20-22 Aug. 2015

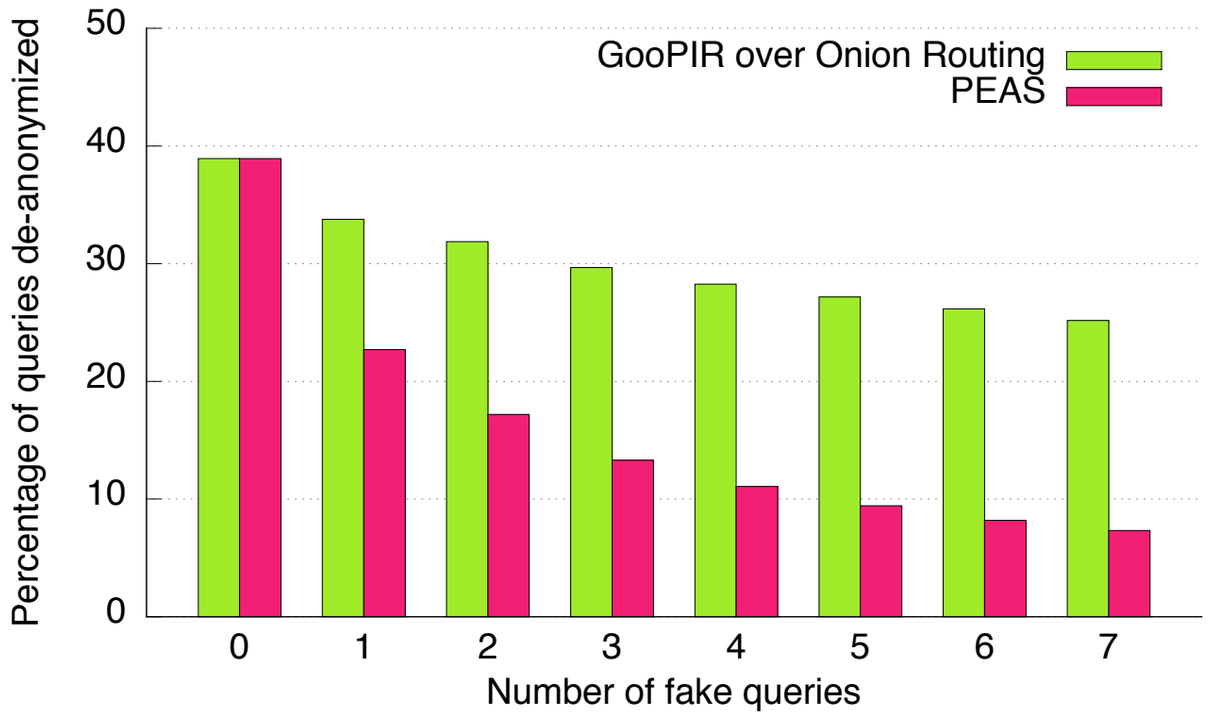
EEXCESS: Privacy Preserving Querying

Research Question
How can we avoid that a search engine profiles their users?

Results

- 2 Attacks based resp. on machine learning and term similarity
- 3 concurrent methods of privacy-preserving Web search (GooPIR,TrackMeNot,TOR)
- Dataset: 343,548 queries from 300 active users - AOL search logs
- Metrics: percentage of queries de-anonymised vs. accuracy of results
  - Accuracy measured in # of rank differences between original and reconstructed results





EEXCESS – Follow Ups

Zero-effort Querying for Digital Libraries

- In-depth writing style analysis of paragraphs
- Learning boolean queries

Representational Learning

- Learning feature representations instead of feature engineering
- Study semantic embeddings on different kinds of data
  - Networks
  - Time Series Data

Deep Learning for Media Analysis

- How powerful are deep learning methods?
- Efficiency?
- Applicability with small number of datasets?

Project 2: CODE

FP 7 IP, 4Partners, Scientific Coordinator, finished

CODE - Goal

Improve access to facts from scientific literature

- Extract facts from PDFs
- Link those facts to the Linked Open Data Cloud
- Provide decentralised, usable search interfaces
- Provide visual analysis tools for linked data
- Crowd-based quality control

Why?

- “Improvements that don’t add up”
  Armstrong et. al. 2009
- “Why most research results are false”
  Ioannidis, 2005

Extracting Facts from Research Publications

Journal of Experimental & Clinical Cancer Research

BioMed Central

Research

Open Access

Nestin and CD133: valuable stem cell-specific markers for determining clinical outcome of glioma patients

Mingyu Zhang, Tao Song, Liang Yang, Ruokun Chen, Lei Wu, Zhuanyi Yang and Jiasheng Fang*

Address: The Neurosurgery Department of Xiangya Hospital of Central South University, Changsha, Hunan, 410078, PR China

Email: Mingyu Zhang - zhangmy@hotmail.com; Tao Song - songtao97117@hotmail.com; Liang Yang - yang@163.com; Ruokun Chen - chenrk@sina.com; Lei Wu - wul@yahoo.cn; Zhuanyi Yang - shen.wai@hotmail.com; Jiasheng Fang* - songtao97117@hotmail.com

* Corresponding author

Published: 24 December 2008

Journal of Experimental & Clinical Cancer Research 2008, **27**:85 doi:10.1186/1756-9966-27-85

Received: 21 October 2008

Accepted: 24 December 2008

This article is available from: http://www.jeccr.com/content/27/1/85

Abstract

Aim: Gliomas represent the most frequent neoplasm of the central nervous system. Unfortunately, surgical cure of it is practically impossible and their clinical course is primarily determined by the biological behaviors of the tumor cells. The aim of this study was to investigate the correlation of the stem cell markers Nestin and CD133 expression with the grading of gliomas, and to evaluate their prognostic value.

Methods: The tissue samples consisted of 56 low- (WHO grade II), 69 high- (WHO grade III, IV) grade gliomas, and 10 normal brain tissues. The expression levels of Nestin and CD133 proteins were detected using SABC immunohistochemical analysis. Then, the correlation of the two markers' expression with gliomas' grading of patients and their prognostic value were determined.

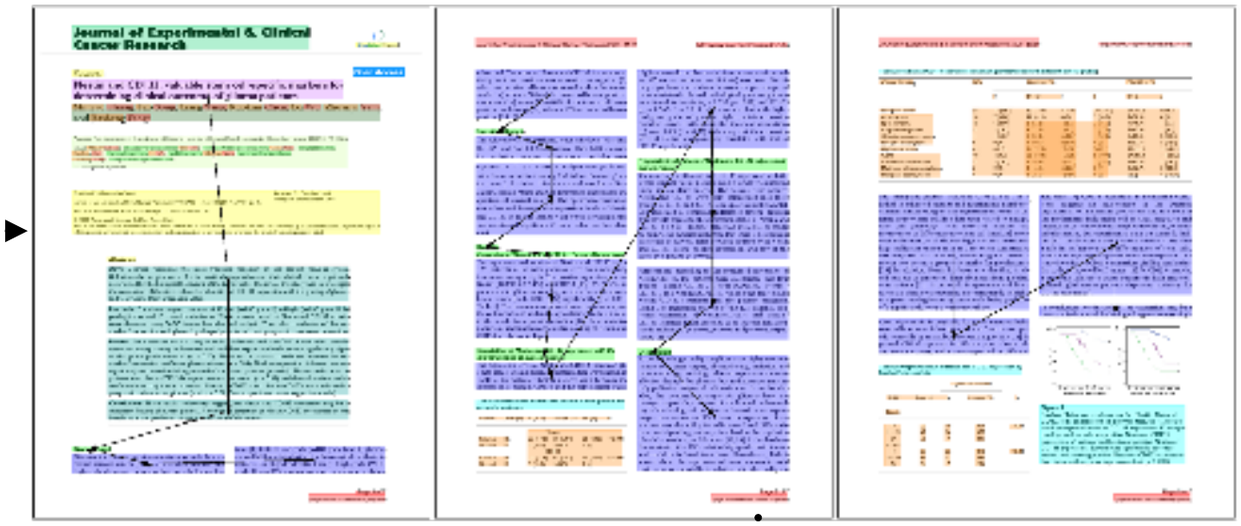
Results: Immunohistochemical analysis with anti-Nestin and anti-CD133 antibodies revealed dense and spotty staining in the tumor cells and their expression levels became significantly higher as the glioma grade advanced ($p < 0.05$). There was a positive correlation between the two markers' expression in different gliomas tissues ($rs = 0.89$). The low expression of the two markers significantly correlated with long survival of the glioma patients ($p < 0.05$). The survival rate of the patients with Nestin+/CD133+ expression was the lowest ($p < 0.01$), and the multivariate analysis confirmed that conjoined expression of Nestin+/CD133+ and Nestin-/CD133- were independent prognostic indicators of gliomas (both $p < 0.01$, Cox proportional hazard regression model).

Conclusion: These results collectively suggest that Nestin and CD133 expression may be an important feature of human gliomas. A combined detection of Nestin/CD133 co-expression may benefit us in the prediction of aggressive nature of this tumor.

Background

Gliomas, the most frequent primary tumors in the human central nervous system (CNS), are classified into low- and high-grade glioma according to their morphological features [1]. Cells of low-grade (WHO grade I and II) gliomas are well differentiated with clear histological similarity to astrocyte or oligodendrocyte lineage. High-grade (WHO grade III and IV) gliomas are more anaplastic, with fea-

Clustering

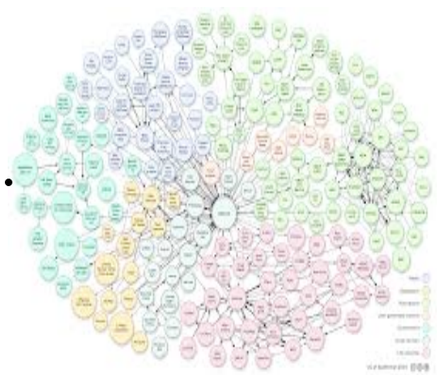


Fact Extraction

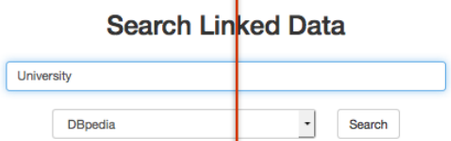
Experiment/Group	$Prec_1$	Rec_1	$Prec_2$	Rec_2
Paired	**0.77**	**0.50**	**0.75**	**0.69**
ACM	0.77	0.53	0.78	0.78
arXiv	0.89	0.58	0.90	0.61
BMC	0.74	0.33	0.26	0.22
IEEE	0.75	0.55	0.87	0.81
Nature	0.86	0.33	0.45	0.40
Physical Reviews	0.89	0.42	0.88	0.53
CRF	**0.50**	**0.21**	**0.73**	**0.73**
ACM	0.43	0.14	0.73	0.66
arXiv	0.53	0.27	0.75	0.77
IEEE	0.52	0.24	0.78	0.79
Nature	0.49	0.09	0.56	0.42
Mixed	**0.81**	**0.62**	**0.94**	**0.91**
ACM	0.84	0.69	0.88	0.84
arXiv	0.84	0.63	0.94	0.92
BMC	0.94	0.91	0.96	0.95
IEEE	0.75	0.54	0.94	0.91
Nature	0.91	0.67	0.94	0.92
Physical Reviews	0.85	0.59	0.89	0.86
all groups	0.81	0.63	0.93	0.91

Type Inference

Disambiguation



Search



Our Work in the Project

Column Type Inference

Given:

- Table Column: header as type + data in cells (I_1 … I_i)
- Knowledge Base: Set of Entities (E) + Typehierarchie(T)

Goal: Link Columns to Entity

Approach:

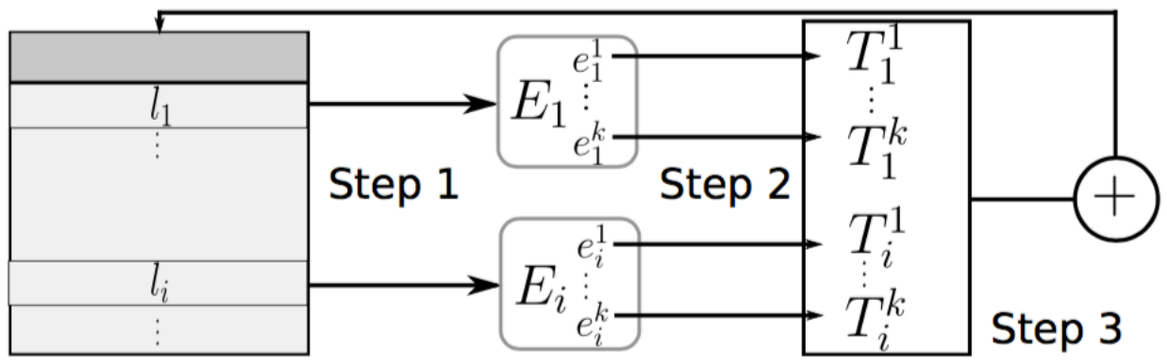


Fig. 1. Annotation process. 1) Cell labels are disambiguated to entity candidates. 2) Types of entity candidates are determined. 3) Type information is aggregated to determine the header type candidates.

S. Zwicklbauer, C. Einsiedler, M. Granitzer, and C. Seifert, “Towards Disambiguating Web Tables,” in Proceedings of the International Semantic Web Conference 2013 (ISWC 2013), 2013, p. (Poster).

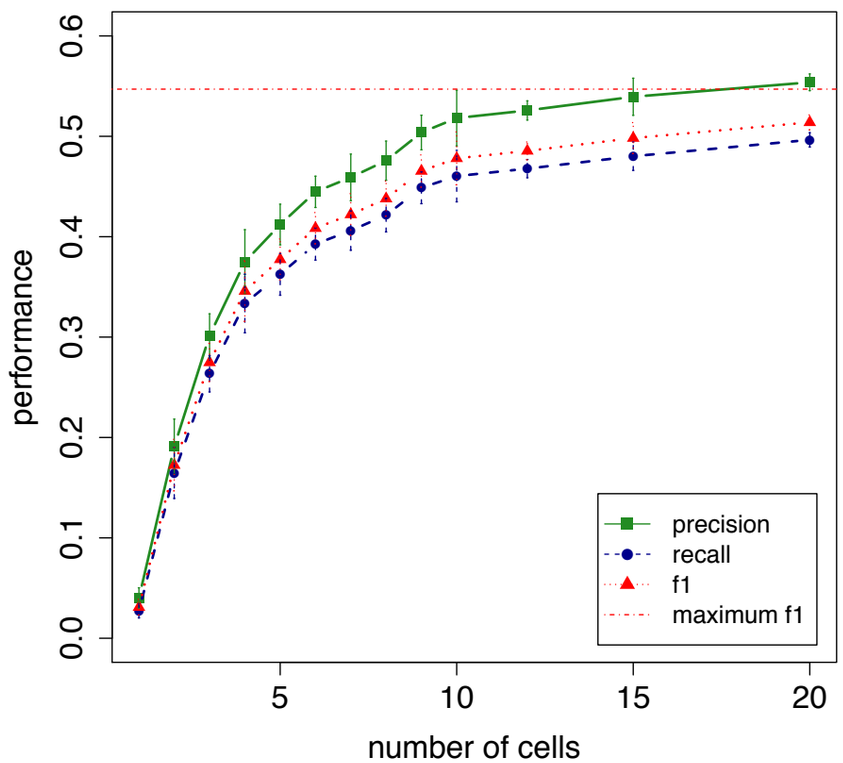
Column Type Inference

Results:

- DBPedia as Knowledge Base (rdf:type, dc:terms)
- 50 Tables from Wikipedia (according to prior work of Limaye et al.)
  - 132 columns, 10-232 rows.
  - Manual annotation of columns (2.5 annotations per column header)
    - 169 rdf:type
    - 169 dc:terms

Table 1. Performance for different cell annotation methods and type vocabularies. Reporting macro-averaged precision π, recall ρ, F_1.

Vocabulary	π^M	ρ^M	F_1^M
Rdf-Type	0.24	0.22	0.23
DublinCore	0.59	0.51	0.55
Rdf-Type + DublinCore	0.64	0.27	0.38



S. Zwicklbauer, C. Einsiedler, M. Granitzer, and C. Seifert, “Towards Disambiguating Web Tables,” in Proceedings of the International Semantic Web Conference 2013 (ISWC 2013), 2013, p. (Poster).

Discovering and Merging RDF Data Cubes

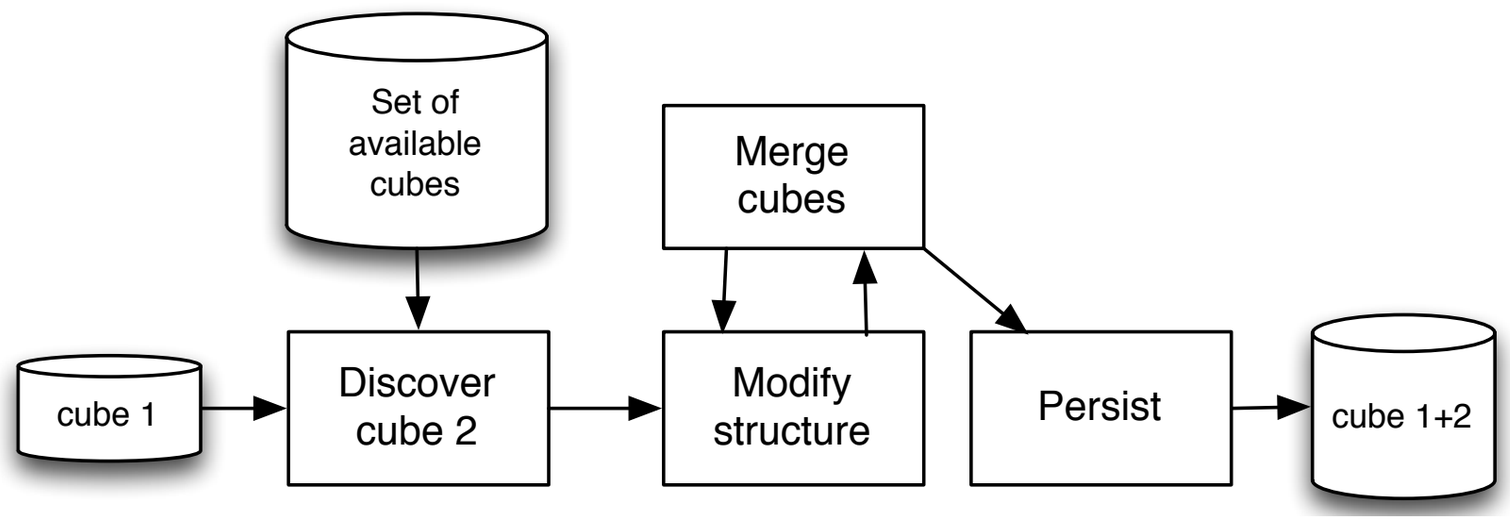
RDF Data Cube

- Semantic Web Format for representing statistical data (i.e. OLAP Cubes)
- Dimensions, Measures and Attributes

Given: RDF Data Cubes residing in decentralised repositories

Goal: Discover similar data cubes that can be merged

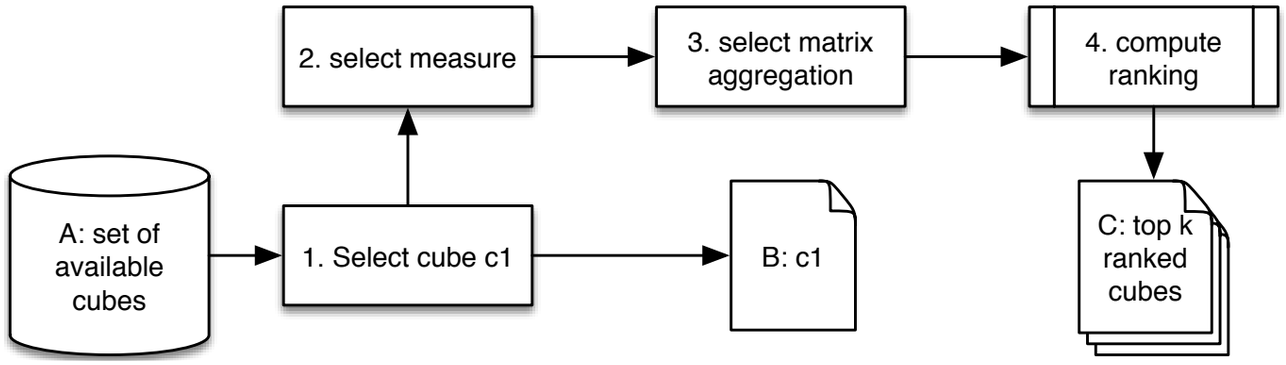
- E.g. Cube1: Census Uni Passau + Cube 2: Census Uni Weimar



S. Bayerl and M. Granitzer, “bacon: Linked Data Integration based on the RDF Data Cube Vocabulary,” in Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics - WIMS ’15, 2015

Discovering and Merging RDF Data Cubes

Approach



2. Measures to compare cube components

- TFIDF, Word2Vec, Label Similarity, Graph Distance, Concept Similarity

3. Estimate Mergability Value

- Average, Maximum
- Bi-partite Graph Matching for finding optimal pairs

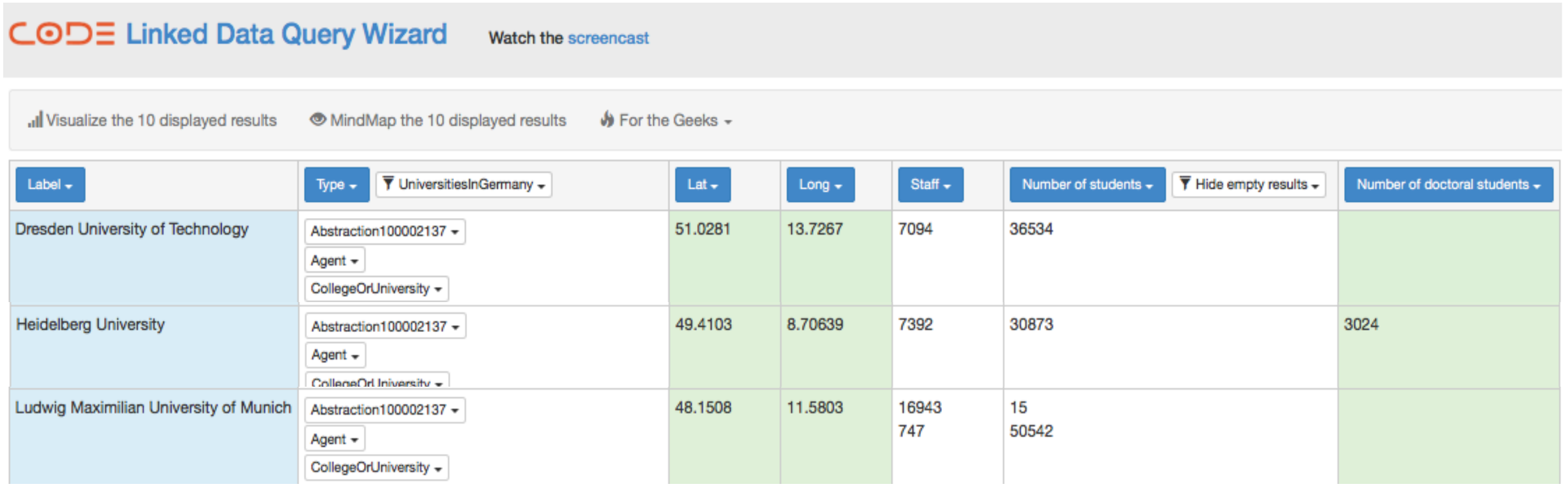
C1 \ C2	D1	D3	D4	M2
D1	1	0.3	0.1	0
D2	0.5	0.2	0.9	0.1
M1	0.1	0.3	0.1	0.8

Mergability Value: → 0.9

Preliminary Results

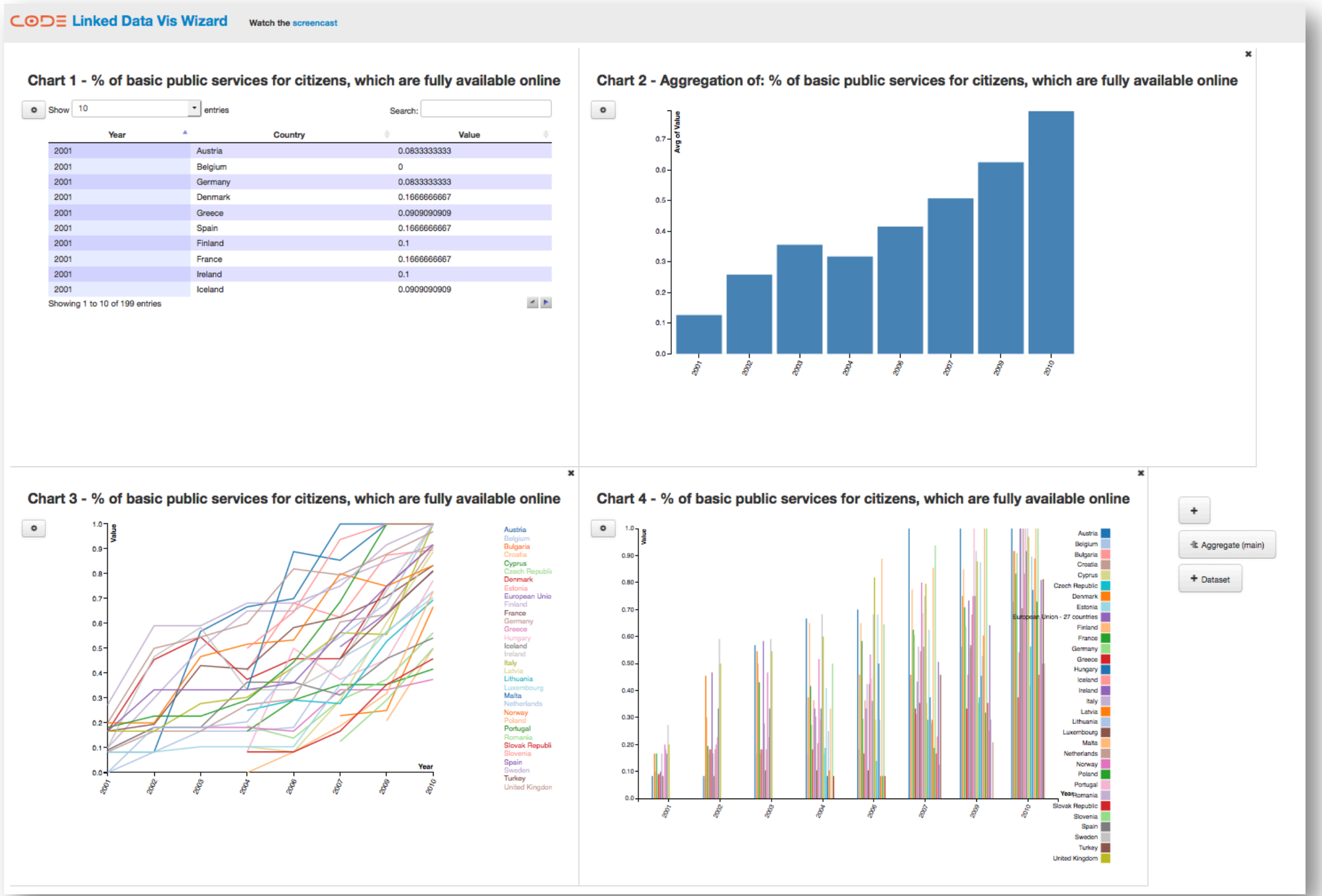
- 3 Cubes out of 69 manually annotated cubes
- Best solution: Word2Vec + Concept Similarity with Bi-partite Graph Matching
- Large Variance

Linked Data Query Wizard



Problem: Data Quality in the LOD

Web-based Visualisation



CODE - Reflection

- **Extraction Quality:** Some additional steps in extracting facts, but lacking quality yields to low usability
- **Data Quality:** Small parts of the LOD are OK, the rest has extremely low quality. I doubt future uptake.
- **Search for Research Facts:** Still interesting topic, but low uptake due to low LOD quality
- **Crowd based Quality control:** Missing motivation in our scenario. Could not identify a concrete use-case + missing ambitions from our partner to go beyond standard business

CODE – Possible Follow Ups

Fact Extraction from Scientific Literature

- Still unsolved. Needs stronger recognition methods (e.g. Deep Nets)
- Impact dimishing over the next years
  - Newer publications > older publications
  - Research data (+ more semantics) to be published with articles in future

Integrating (Linked) Open Data into Data Mining Processes

- E.g. Weather data in predicting customer behavior
- More focus on machine learning algorithms
  - Augment the data set
  - Optimize models:
    - Specialised Kernels
    - Topology selection in Neural Networks

Other Projects Overview

Credit Card Fraud Detection

Feature Engineering

- Primary Attributes: Date, Time, Amount, Currency….
- Derived Features to aggregate user behaviour: Avg. Amount, Transactions/Day …

Machine Learning Algorithm

- RF>NNet, SVM, LR
- Data set selection as crucial property
- Non-stationary models, unbalanced data sets, incremental learning

Questions

- Can we integrate new features from Linked (Open) Data sources automatically to improve CCRD performance?
- How does Deep Learning compare in terms of accuracy to current ML Algorithms?
- Can Deep Neural Networks work in near real time settings for training and testing?
- How can users/engineers understand decisions taken and their impact?

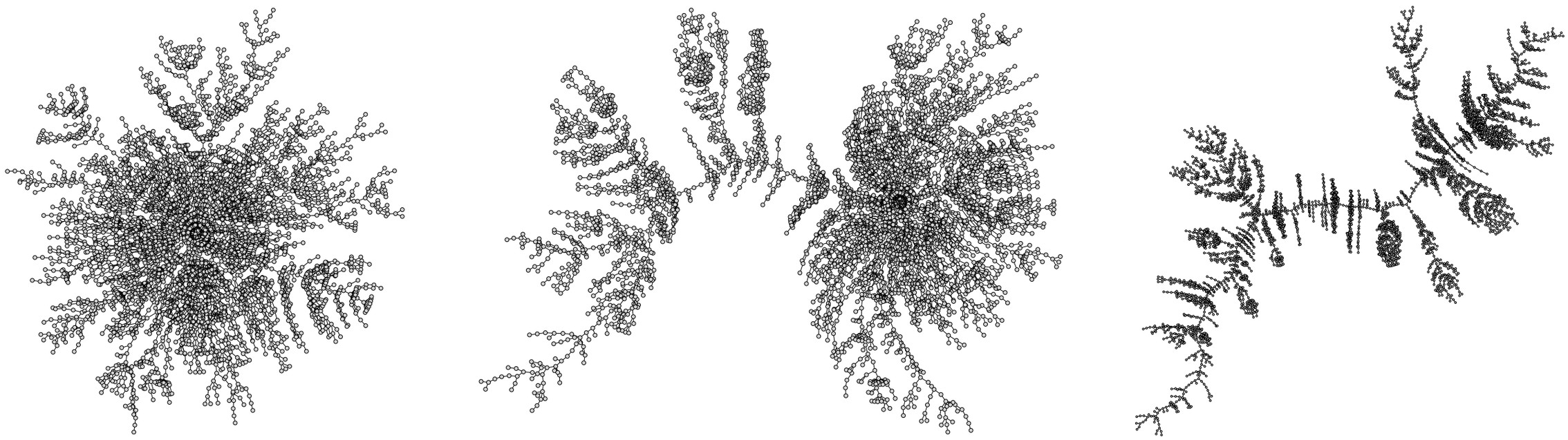
Analyse Navigation Behaviour in Information Networks

Research Goal

- Agents navigate in an Information Network (i.e. Wikipedia) based on hierarchical background knowledge (in terms of nodes in the network)
- Find the optimal hierarchy of nodes using genetic algorithms

Approach

- Develop recombination mechanisms for trees and according fitness functions
- Study the best hierarchies

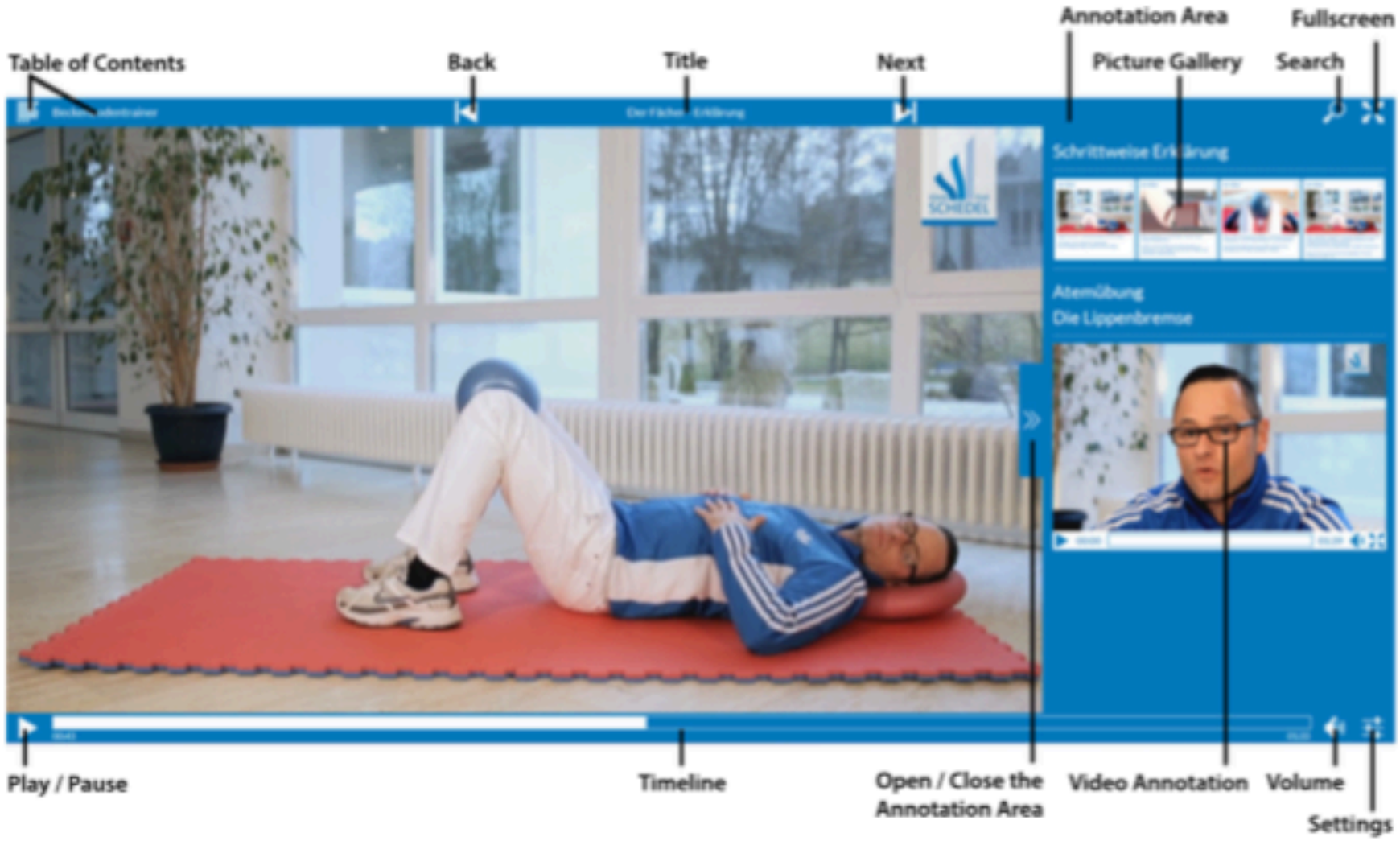


(a) The good (b) The bad (c) The ugly

Fig. 7. Three Wikipedia hierarchies of different quality. Subfigure (a) shows the best hierarchy created by our experiments, with a global stretch of 8.54. In subfigure (b) the global stretch is slightly lower (11.97). A subtree splitting off from the navigational core can be observed. The last hierarchy has a global stretch value of 22.35. A very linear structure is already clearly noticeable.

Hypervideos for Knowledge Transfer

Knowledge transfer for manual tasks using hypervideos and 2nd Screens (e.g. mobile phones)



Thanks for your attention. Questions?