Extracting and Analyzing Sequential Interaction-Patterns

Simon Walk

July 22, 2016

Motivation

Over the last decade, ontologies have become the mainstay in the biomedical domain.

- New and complementing areas of application
- Increased complexity & size

For example, ICD-11 consists of roughly $50,000$ classes.

- Highly specialized knowledge
- Many different areas of expertise

Ontologies have become **very hard to develop and maintain**.

Collaborative Ontology Engineering

Similar to Wikipedia, contributors engage remotely in developing ontologies.

- Many new and unexplored problems
- Layer of social interactions adds complexity

Administrators are in need of tools to better manage the complex collaborative engineering process.

Objective: Broaden our understanding of the dynamic social processes by analyzing edit patterns.

Outline

Interaction patterns in BioPortal

How can we explain edit patterns in collaborative ontology-engineering projects?

Do patterns & regularities exist in collaborative ontology-engineering projects?

How to identify regularities & patterns in collaborative ontology-engineering projects?

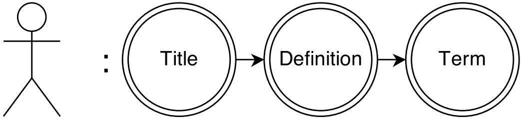
Datasets

Characteristics of the datasets used for the different collaborative ontology-engineering analyses.

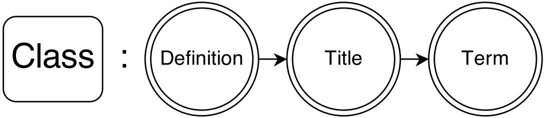
	ICD-11	ICTM	NCIt	BRO	OPL
# classes	48,771	1,506	102,865	528	393
# changes	439,229	67,522	294,471	2,507	1,993
# users	109	27	17	5	3
first change	18.11.2009	02.02.2011	01.06.2010	12.02.2010	09.06.2011
last change	29.08.2013	17.07.2013	19.08.2013	06.03.2010	23.09.2011
observation period (ca.)	4 years	2.5 years	3 years	1 month	3 months

(Sequential) Interaction-Sequences

- **User**-based sequences



- **Class**-based sequences



Identifying Interaction Patterns

Using Markov chains

- State space S, listing all possible states $s_1, s_2, ...s_n \in S$ with $|S| = n$.
- Transition matrix P with p_{ij} listing the probability to go from state s_i to s_j.

First-order Markov chain (Markovian property):

$$P(X_{t+1} = s_j | \underbrace{X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}}_{\text{all previous transitions}}) =$$

$$P(X_{t+1} = s_j | \underbrace{X_t = s_{i_t}}_{\text{current transition}}) = p_{ij}$$

Identifying Interaction-Patterns

Markov chain of order k means that k previous states contain (useful) predictive information about the next state.

$$P(X_{t+1} = s_j | \underbrace{X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}}_{\text{all previous transitions}}) =$$

$$P(X_{t+1} = s_j | \underbrace{X_{t-k+1} = s_{i_{t-k+1}}, ..., X_t = s_{i_t}}_{\text{k transitions}})$$

Overfitting and model complexity are problematic!

- Lower order models are nested in higher order models!
  - Solution: Model selection and prediction experiments
- Parameter increase: $\theta = |S|^k|S|$
  - Solution: Aggregated/abstract states

Process to Identify Interaction-Patterns

- Preprocessing
  - Mapping, Session Separation, State Selection, Path Extraction
- Model Fitting
- Model Selection
  - Akaike IC, Bayesian IC, Prediction Experiments
- Interpretation

[Walk et al., 2015b] Simon Walk, Philipp Singer, Markus Strohmaier, Denis Helic, Natalya Noy, Mark Musen: **How to apply Markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects.** Int. J. Hum.-Comput. Stud. 84: 51-66 (2015)

Outline

Do patterns & regularities exist in collaborative ontology-engineering projects?

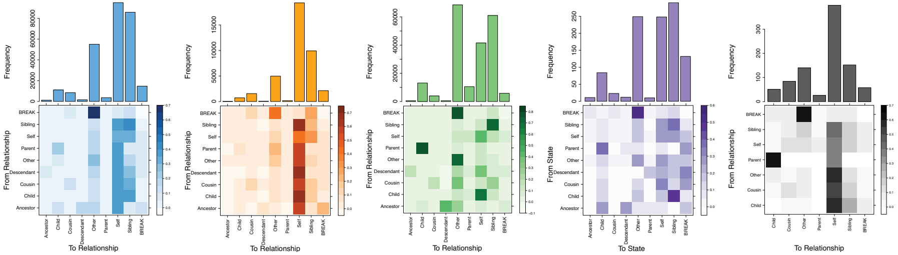
How to identify regularities & patterns in collaborative ontology-engineering projects?

Sequences to Analyze for Patterns

- User Sequences
  - Who will change a class next?
  - Which type of change will a user perform next?
- Content-based Sequences
  - Which area of the user interface will a user use next?
  - Which property will a user change next?
- Structural Sequences
  - Which class is a user going to edit next?
  - Where is the next class located in the ontology?
  - Do users move along the ontological hierarchy when contributing to the projects?

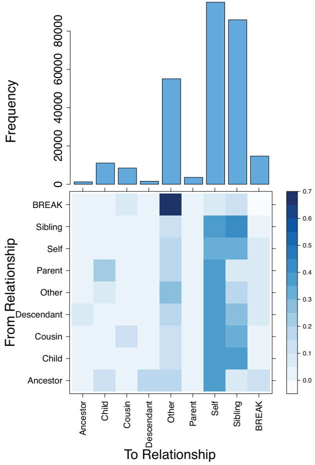
Do users move along the ontological hierarchy when contributing to the projects?

States: Self, Parent, Child, Sibling, Cousin, Ascendent, Descendent, Other

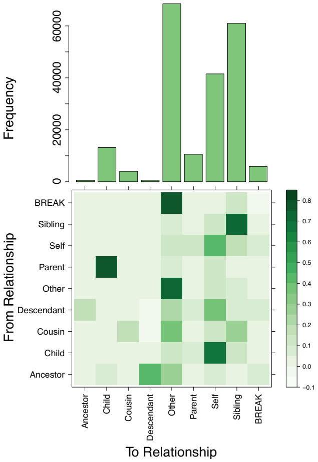


(a) ICD-11 (b) ICTM (c) NCIt (d) BRO (e) OPL

Do users move along the ontological hierarchy when contributing to the projects?



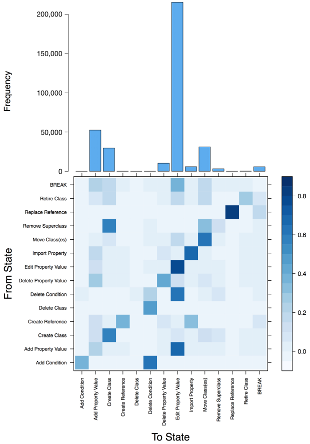
(a) ICD-11



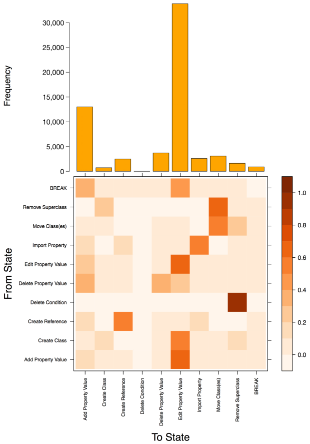
(b) NCIt

Understanding Editing Behaviors in Ontology-Development Projects!

States: Types of changes (aggregated) available in iCAT (Edit Property Value, Create Class, etc.).



(a) ICD-11



(b) ICTM

Modeling Sequential Interaction-Sequences

Summary of findings

- The edit behavior of users is influenced by the hierarchy of the ontology.
- Users edit the ontology top-down and breadth-first.
- Users work in micro-workflows.
- Roles of users can be identified.
- Users edit closely related classes.
- Users perform property-based workflows.

[Walk et al., 2014b] Simon Walk, Philipp Singer, Markus Strohmaier, Tania Tudorache, Mark Musen, Natalya Noy: **Discovering Beaten Paths in Collaborative Ontology-Engineering Projects using Markov Chains.** Journal of Biomedical Informatics 51: 254-271 (2014)

Predicting Aspects of Future Actions

k cross-fold prediction experiment

- k stratified splits
  - $k-1$ splits for the training set
  - 1 split for the test set
- Determine the rank of each transition in test set
  - Modified competition ranking
  - Natural occam's razor
- Calculate average rank over all transitions and splits
- Lowest average rank determines best performing Markov chain order
  - Best models: Average rank between 1.7 and 3
  - Worst models: Average rank between 2 and 6

Predicting Aspects of Future Actions

Best performing Markov chain orders

	ICD-11	ICTM	NCIt	BRO	OPL
Predict Users for Classes	1	1	1	1	2
Predict Change Types for Users	3	2	-	1	1
Predict Change Types for Classes	4	3	-	2	2
Predict Properties for Users	1	1	-	1	0
Predict Properties for Classes	1	1	-	3	2

[Walk et al., 2014a] Simon Walk, Philipp Singer, Markus Strohmaier: **Sequential Action Patterns in Collaborative Ontology-Engineering Projects: A Case-Study in the Biomedical Domain.** CIKM 2014: 1349-1358

Outline

How can we explain edit patterns in collaborative ontology-engineering projects?

Do patterns & regularities exist in collaborative ontology-engineering projects?

How to identify regularities & patterns in collaborative ontology-engineering projects?

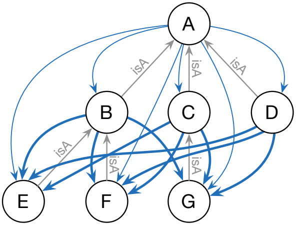
Formulating & Comparing hypotheses

- Hypotheses are *potential explanations* as opposed to actual empirical transitional observations.
- Can be expressed as *hypothesis matrix* Q where
  - q_{ij} represents the belief in the transition between states s_i and s_j
  - and $\sum_j q_{ij} = 1$ for each row i of Q.

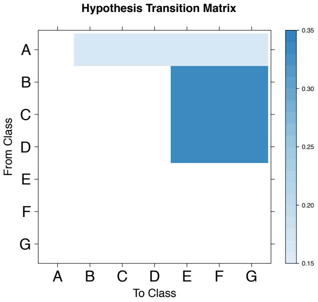
Example: Top-down hypothesis

Classes deeper in the hierarchy than the previously edited class are more likely to be changed next.

$$q_{ij} = \begin{cases} 1, & \text{if } depth_i < depth_j, \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

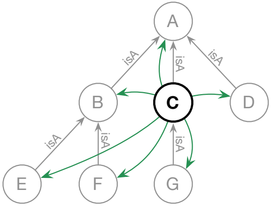


(c) Top-Down Example



(d) Hypothesis Matrix Q

Hypotheses



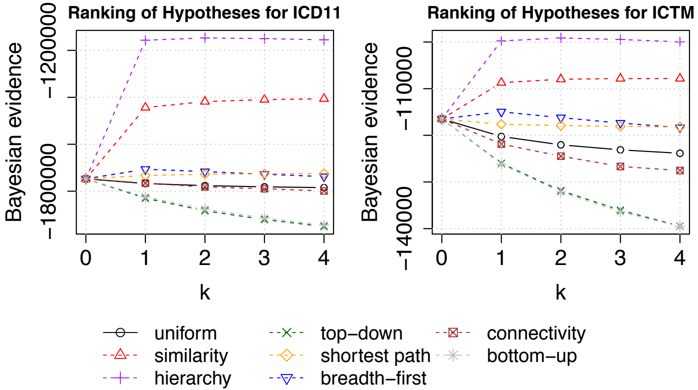
Uniform Hierarchy Similarity

HypTrails

A framework to study the relative plausibility of hypotheses (about the production of human edit sequences).

- Sequences modeled as first-order Markov chain.
- Uses Bayesian inference (marginal likelihood) for comparing different hypotheses.
  - The marginal likelihood $P(D|H)$ describes the probability of data D given hypothesis H.
  - Higher evidences indicate higher plausibility.
- Factor k, describing the strength of our belief in a hypothesis.
- Produces a ranked list of hypotheses (Bayesian evidences).

HypTrails Results



[Walk et al., 2015a] Simon Walk Philipp Singer, Lisette Espin Noboa, Tania Tudorache, Mark Musen, Markus Strohmaier: **Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects.** International Semantic Web Conference (1) 2015: 551-568

Outline

Interaction patterns in BioPortal

How can we explain edit patterns in collaborative ontology-engineering projects?

Do patterns & regularities exist in collaborative ontology-engineering projects?

How to identify regularities & patterns in collaborative ontology-engineering projects?

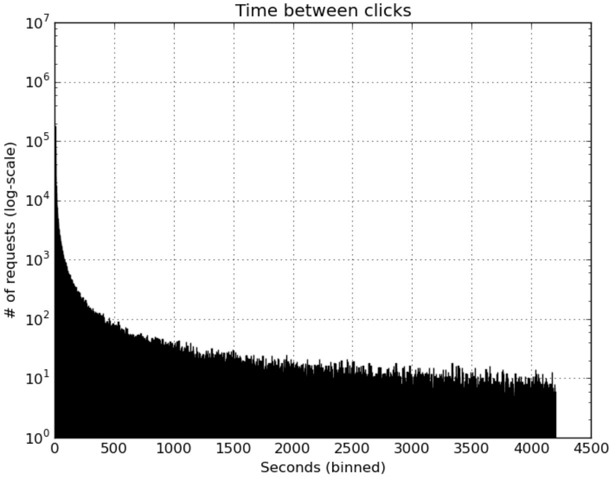
BioPortal (Jan – Apr 2016)

Apply Markov chains and conduct analyses on Request Logs of BioPortal

Feature	**Value**
Requests before filtering	$\sim$50 M
Requests after filtering	$\sim$16.2 M
Click-requests (interactions)	$\sim$2.1 M
Distinct IPs	$160,325$
Number of Ontologies	$1,652$ (IDs + Names) $\sim$730 IDs $\sim$500 unresolvable names

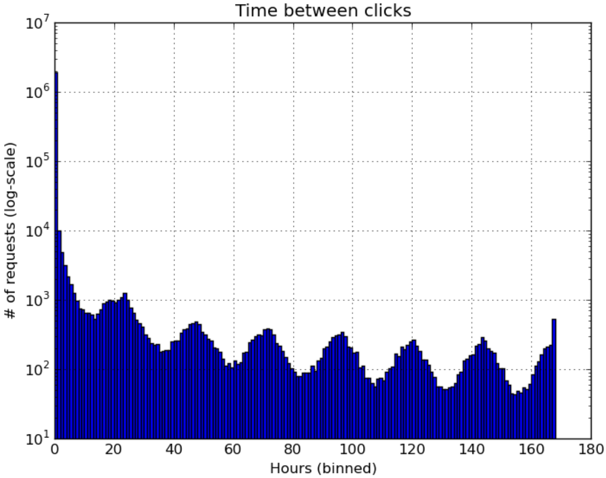
Seconds between click-requests (cut-off at $4,200$ seconds)

Ordered by IP, considered only users with > 1 request.



Hours between click-requests (cut-off at 168 hours)

Ordered by IP, considered only users with > 1 request.



Click-Sessions

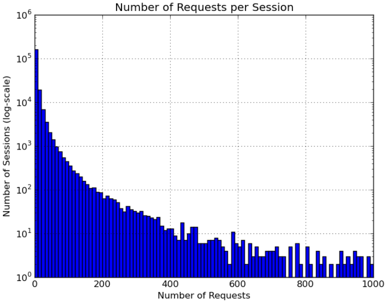
Definition:
A sequence of clicks, where each click is performed **within 1,800 seconds** (30 minutes) of the previous click. However, sessions can be longer than 30 minutes!

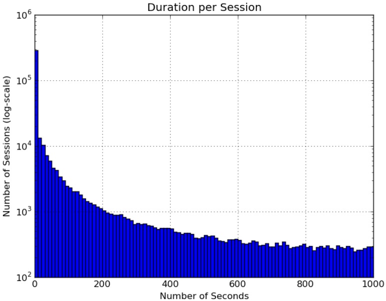
Problem:
Due to the dynamic nature of BioPortal (AJAX, caching, ...), Referrer is often "wrong" (or missing), making it impossible to trace sessions, tabbed browsing, back-clicks, etc.!

Click-Sessions in BioPortal (Jan – Apr 2016)

Feature	**Value**
Click-requests	$\sim$2.1 M
Sessions	198,610
1-click sessions	64,492
$\geq$10-click sessions	38,535
Min/Max clicks per session	1 / 3,678
Average/Median/Mode clicks per session	10.5 / 3 / 1
Min/Max duration	0 / 6h
Average/Median/Mode duration	175.6s / 2s / 1s

Requests & Duration per Click-Session





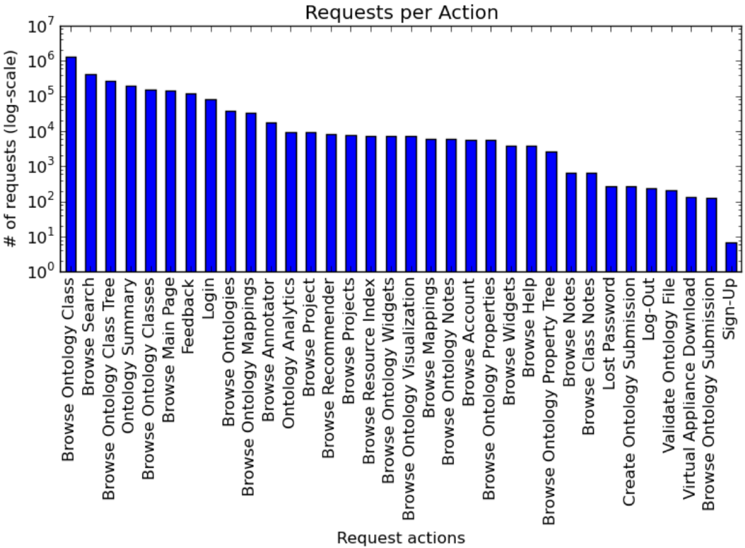
Example Click-Session

Timestamp	Request
2016-03-14 09:07:46	/
2016-03-14 09:07:48	/login?redirect=http%3A%2F%2Fbioportal.bioontology.org%2F
2016-03-14 09:07:50	/login
2016-03-14 09:08:04	/
2016-03-14 09:08:22	/ontologies/MCCV
2016-03-14 09:09:34	/ontologies/MCCV/submissions/new
2016-03-14 09:08:58	/ontologies/MCCV/submissions
2016-03-14 09:07:59	/ontologies/success/MCCV
2016-03-14 09:10:14	/ontologies/MCCV

Example Click-Session

Click-Action (General)	Click-Action (Specific)
Browse Main Page	Browse Ontology Class
Browse Ontologies	Browse Ontology Class Tree
Browse Search	Ontology Summary
Browse Help	Browse Ontology Classes
Browse Mappings	Browse Ontology Mappings
Browse Recommender	Ontology Analytics
Browse Annotator	Browse Ontology Widgets
Browse Resource Index	Browse Ontology Visualization
Browse Projects	Browse Ontology Notes
Browse Notes	Browse Ontology Properties
Login	Browse Widgets
Log-Out	Browse Ontology Property Tree
Sign-Up	Browse Class Notes
Lost Password	Create Ontology Submission
Browse Account	Validate Ontology File
Feedback	Virtual Appliance Download
	Browse Ontology Submission

Frequency of Click-Actions



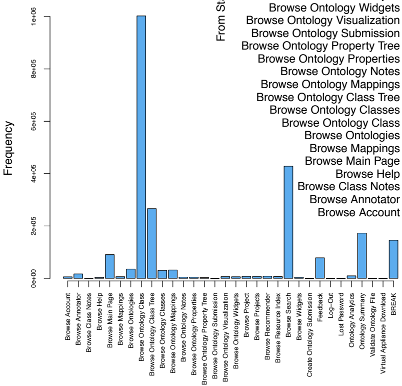
Example Click-Session

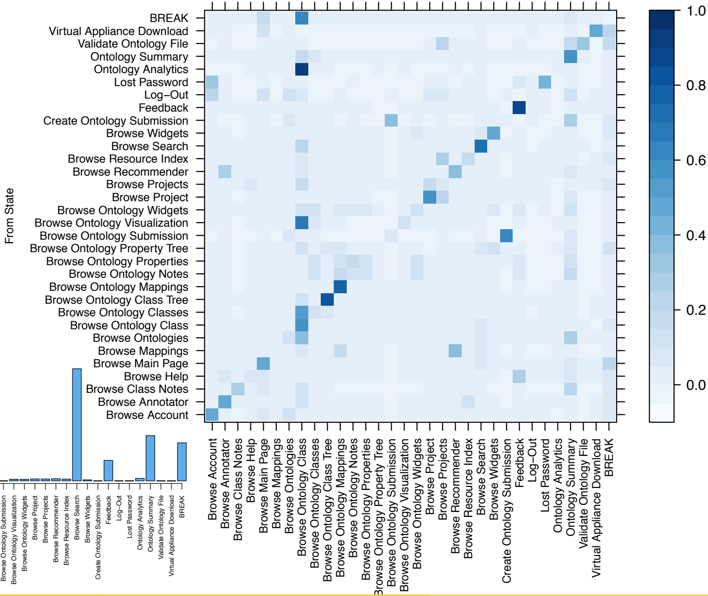
Timestamp	Click-Action	Request
2016-03-14 09:07:46	Browse Main Page	/
2016-03-14 09:07:48	Login	/login?redirect=http%3A%2F%2Fbioportal.bioontology.org%2F
2016-03-14 09:07:50	Login	/login
2016-03-14 09:08:04	Browse Main Page	/
2016-03-14 09:08:22	Ontology Summary	/ontologies/MCCV
2016-03-14 09:09:34	Create Ontology Submission	/ontologies/MCCV/submissions/new
2016-03-14 09:08:58	Browse Ontology Submission	/ontologies/MCCV/submissions
2016-03-14 09:07:59	Create Ontology Submission	/ontologies/success/MCCV
2016-03-14 09:10:14	Ontology Summary	/ontologies/MCCV

Interaction Sequence:

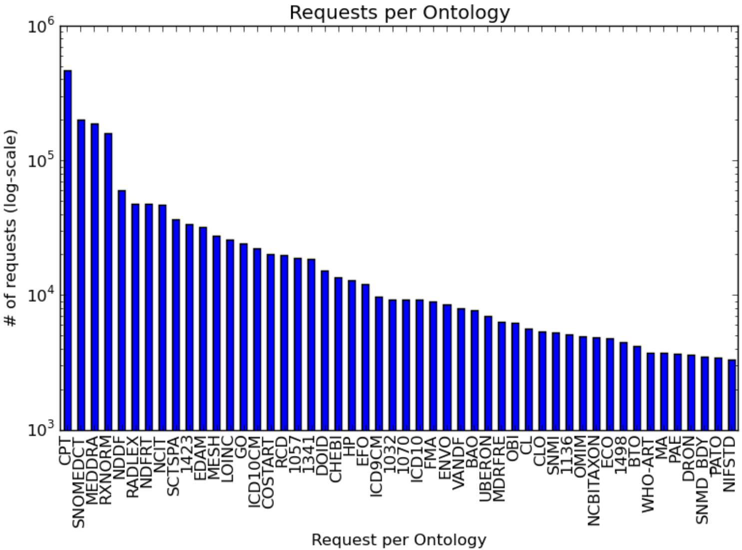
Browse Main Page → Login → Ontology Summary → Create Ontology Submission → Browse Ontology Submission → Ontology Summary

BioPortal Click-Transitions (first-order)

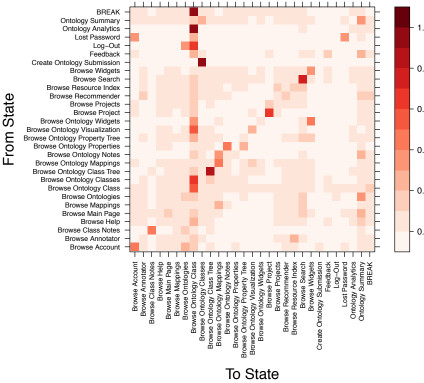




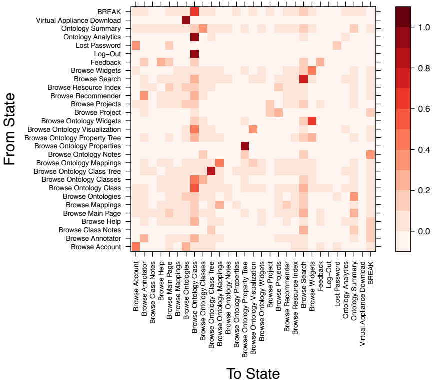
BioPortal Click-Requests per Ontology (Top 50)



BioPortal Click-Transitions per Ontology

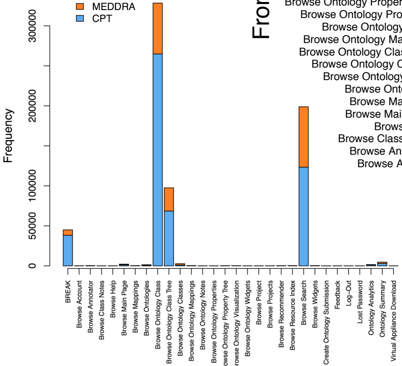


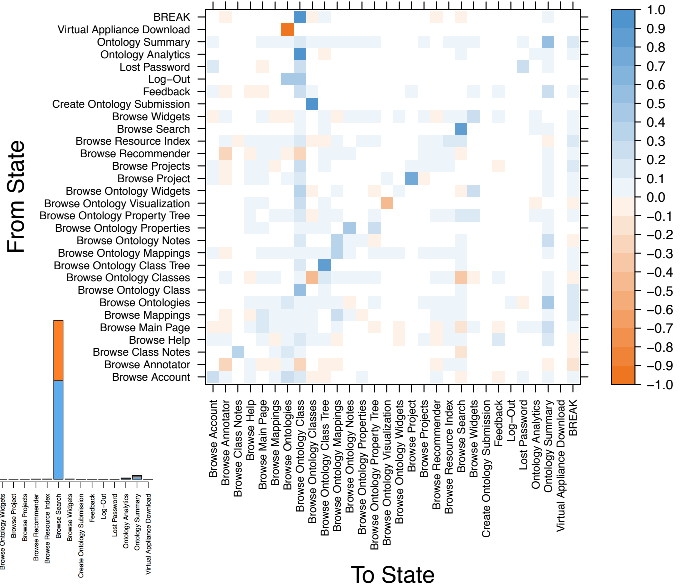
(a) CPT



(b) MEDDRA

Absolute Click-Transitions of CPT & MEDDRA





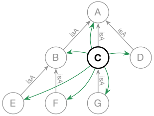
Next Steps

- Interpret and further analyze differences between ontology click-transitions on BioPortal.
- Compare usage of the REST API and the UI.
- Cluster users according to their click-action sequences (similarities)!
  - Calculate stationary distribution vectors and use these to determine distances for clustering.
  - Compare Browsing behaviors between different clusters!
- Compare editing behaviors before and after specific events (e.g., ICD-11 iCAT editing vs. Public ICD-11 Beta Draft) for different datasets!
- Use HypTrails to analyze which collaborative ontology-engineering methodologies people (most likely) follow, when developing an ontology “in the wild”.

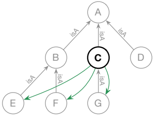
Questions?

Thanks!

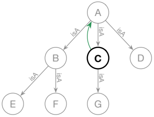
Hypotheses



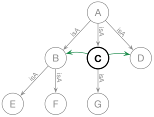
Uniform



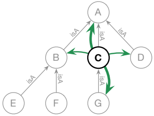
Top-down



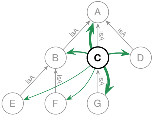
Bottom-up



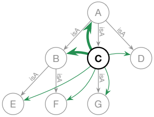
Breadth-first



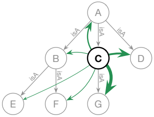
Hierarchy



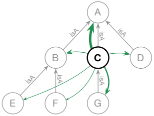
Shortest path



Connectivity



Similarity



Empirical

HypTrails

Distribute chips to elicit Dirichlet prior

$$\beta = \overbrace{m^2}^{\text{uniform prior}} + \underbrace{k * m^2}_{\text{informative prior}} \tag{2}$$

Process to α_{ij}:

- Initial uniform distribution (m^2)
- Informative distribution ($Q = \frac{Q}{||Q||_1} * \beta$)
  - Normalize Q over ℓ_1-norm
  - Multiply with remaining β
- Remaining informative distribution
  - Rank and distribute according to $Q = Q - \lfloor Q \rfloor$

HypTrails - Eliciting Prior Example

$$\beta = 3^2 + k * 3^2, k = 1$$

$$Q = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} \qquad Q = \begin{pmatrix} 0 & 0.33 & 0.66 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\frac{Q}{||Q||_1} = \begin{pmatrix} 0 & 0.165 & 0.33 \\ 0 & 0 & 0 \\ 0.5 & 0 & 0 \end{pmatrix} \; Q\beta = \begin{pmatrix} 0 & \lfloor 1.485 \rfloor & \lfloor 2.97 \rfloor \\ 0 & 0 & 0 \\ \lfloor 4.5 \rfloor & 0 & 0 \end{pmatrix}$$

$$\beta = 9 - 7$$

Model selection

Likelihood ratio test

$$ {}_k\eta_m = -2(\overbrace{\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_k)}^{\text{Log-Likelihood}_k}) - \overbrace{\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_m))}^{\text{Log-Likelihood}_m}) \quad (3) $$

Significance test for likelihood ratios

- χ^2-CDF with ${}_k\eta_m$ and degrees of freedom $(\theta_m - \theta_k)$
- p-value defines significance of alternate model

Model selection

Akaike Information Criterion

$$AIC(k) = {}_k\eta_m - 2(|\theta_m| - |\theta_k|) \tag{4}$$

Balances model complexity (over/underfitting)

- Penalizes model parameters θ

Bayesian Information Criterion

$$BIC(k) = {}_k\eta_m - 2(|\theta_m| - |\theta_k|)\ln(n) \tag{5}$$

Additionally penalizes the number of observations n (transitions).

Model selection

Bayesian Model Selection & HypTrails

Bayes' rule for posterior distribution of θ given data D and hypothesis H.

$$\overbrace{P(\theta|D,H)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,H)}^{\text{likelihood}}\overbrace{P(\theta|H)}^{\textit{prior}}}{\underbrace{P(D|H)}_{\text{marginal likelihood}}} \tag{6}$$

$$P(D|H) = \prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \frac{\prod_j \Gamma(n_{ij} + \alpha_{ij})}{\Gamma(\sum_j (n_{ij} + \alpha_{ij}))} \tag{7}$$

- Hyperparameters α represent pseudo counts
- n_{ij} is the number of transitions between states s_i and s_j