How Users Explore Ontologies on the Web
A Study of NCBO's BioPortal Usage Logs

Simon Walk[1], Lisette Espín-Noboa[2], Denis Helic[3],
Markus Strohmaier[2,4] and Mark A. Musen[1]

[1]Stanford University

[2]Gesis - Leibniz Institute for the Social Sciences

[3]Graz University of Technology

[4]University of Koblenz-Landau

July 6, 2017

Motivation

- Developing and maintaining ontologies is very time consuming.
- Ontologies in the biomedical domain are highly specialized and expensive to develop.
- Many different repositories exist to foster reuse of existing ontologies.
- We do not fully understand how users are using ontology repositories to search and explore ontologies!
- Improve our understanding of **how** ontology repositories, such as BioPortal, are used by **what kind** of users.

Dataset & Preprocessing

- Logs of 51.8M requests (Apache) between 01/2016 to 07/2016
  - IP, timestamp, requested resource, useragent with browser, operating system or custom name for bot/spider

Table: Characteristics of the BioPortal dataset.

Feature	Value
Requests (unfiltered)	51.8M
Requests (filtered for bots)	16.7M
Click actions (filtered AJAX calls)	2.52M
Unique IPs	215,908
Ontologies	1,818 (517)
Unique IPs (with ≥ 2 clicks)	168,008
Observation period (ca.)	7 months

Click-Action Sequence Example

We assigned easier readable action labels to all 2.52M requests and created click-action sequences for all users.

Table: *Example of a click-action sequence.*

Timestamp	Type	Request	Action Labels (Sequence Step)
09:07:32	GET	/	Browse Main Page (1)
09:07:46	GET	/login?redirect=http%3A%2F%2Fbioportal.bioontology.org	Login (2)
09:07:48	POST	/login	Login (3)
09:07:50	GET	/	Browse Main Page (4)
09:08:04	GET	/ontologies/MCCV	Ontology Summary (5)
09:08:22	GET	/ontologies/MCCV/submissions/new	Create Ontology Submission (6)
09:09:34	POST	/ontologies/MCCV/submissions	Create Ontology Submission (7)
09:09:59	GET	/ontologies/success/MCCV	Create Ontology Submission (8)
09:10:14	GET	/ontologies/MCCV	Ontology Summary (9)

Sequence: Browse Main Page - Login - Login - Browse Main Page - Ontology Summary - Create Ontology Submission - Create Ontology Submission - Create Ontology Submission - Ontology Summary

Click-Action Labels

Table: Overview of all click actions available in the BioPortal user interface.

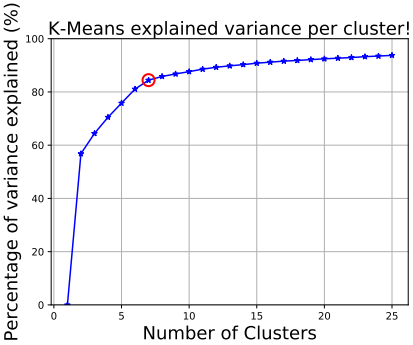
	Click-Action Labels	Description
Main Page	*Browse Main Page*, *Browse Ontologies*, *Browse Search*, *Browse Help*, *Browse Mappings*, *Browse Recommender*, *Browse Annotator*, *Browse Resource Index*, *Browse Projects*, *Browse Notes*	Browsing main areas and functionality of BioPortal.
Ontology Page	*Ontology Summary*, *Browse Ontology Classes*, *Browse Ontology Class*, *Browse Ontology Class Tree*, *Browse Ontology Mappings*, *Ontology Analytics*, *Browse Ontology Widgets*, *Browse Ontology Visualization*, *Browse Ontology Notes*, *Browse Ontology Properties*, *Browse Widgets*, *Browse Ontology Property Tree*, *Browse Class Notes*	Actions that can be performed while browsing a specific ontology on BioPortal.
Edit Content	*Create Ontology Submission*, *Validate Ontology File*, *Virtual Appliance Download*, *Browse Ontology Submission*	Actions triggered when uploading new versions of ontologies.
User Account	*Login*, *Log-Out*, *Sign-Up*, *Lost Password*, *Browse Account*, *Feedback*	Actions to manage accounts on BioPortal.
Control Term	*BREAK*	30 minutes of inactivity.

Modeling Dynamic Browsing Behavior

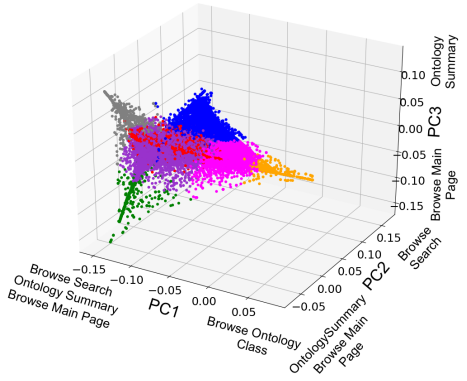
- Use click-action sequences to calculate stationary distribution to *represent their dynamic browsing behavior* on BioPortal.
  - How likely can we find a user on a given state in the limit of large number of steps.
  - On BioPortal: Given a sequence of click actions, what is the probability to find a user conducting a specific action at any point in time?

Clustering Users

To learn how users adopt BioPortal we use their stationary distribution vectors as input for K-Means.

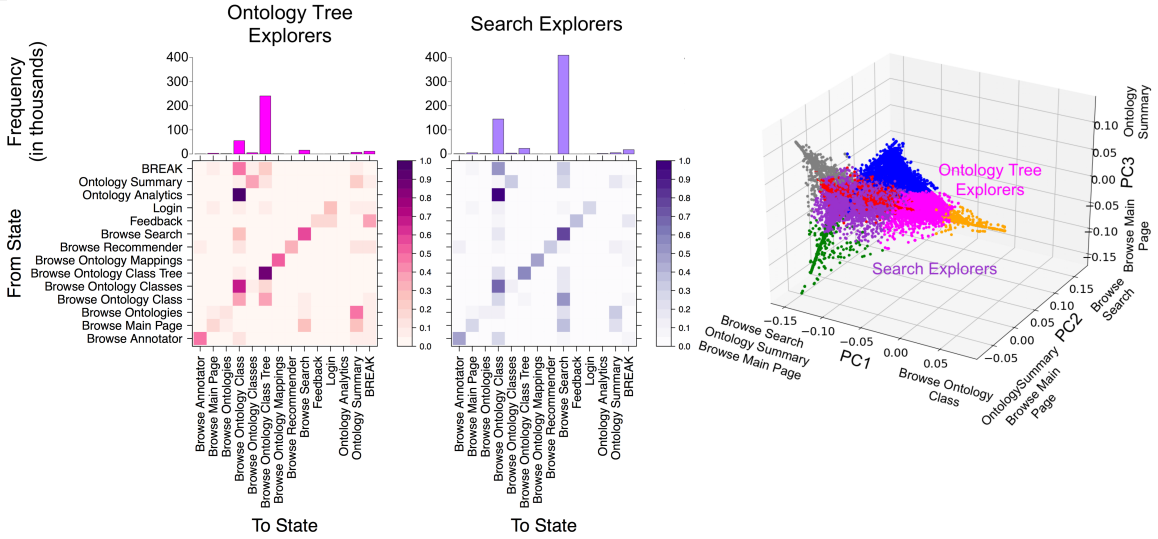


(a) K-means Validation

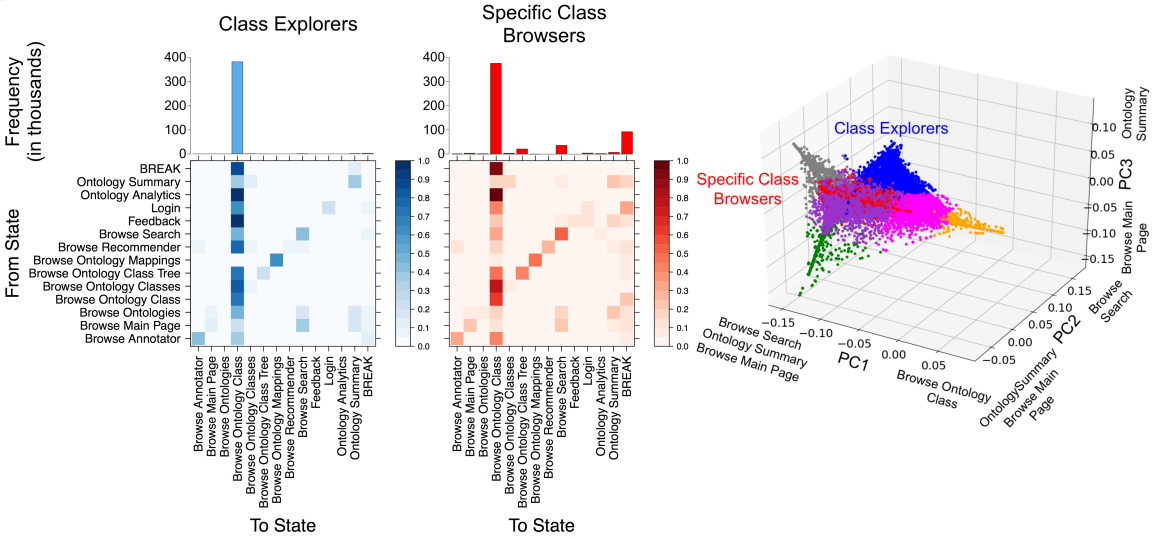


(b) K-means Clusters (PCA)

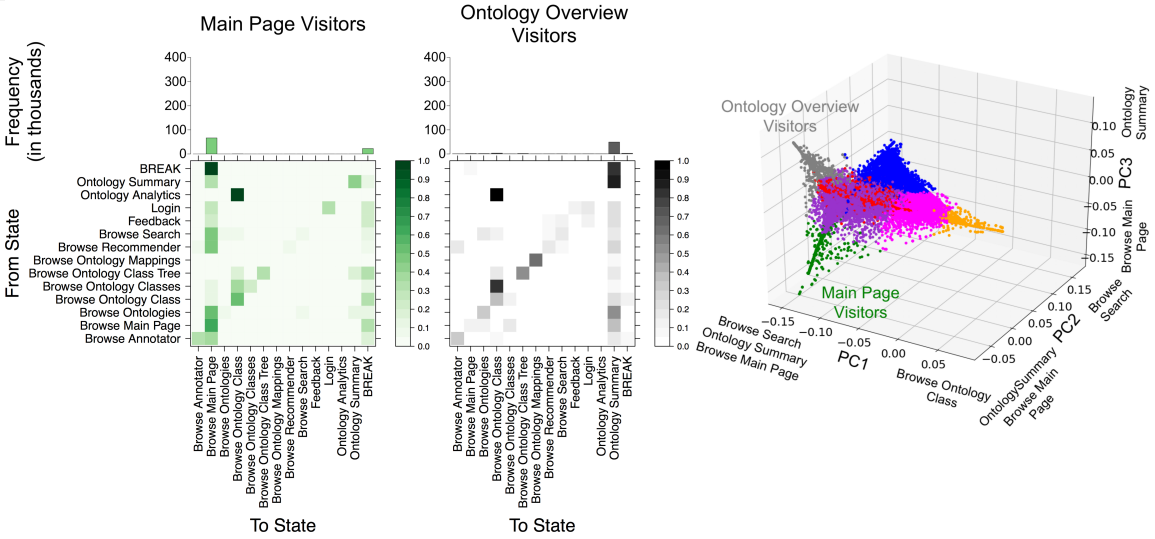
Ontology Tree & Search Explorers



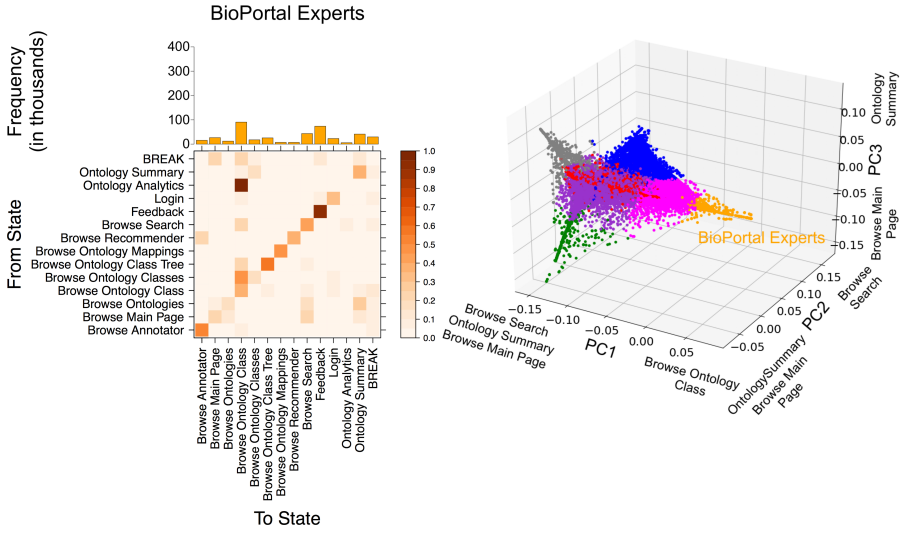
Class & Specific Class Explorers



Main Page & Ontology Overview Visitors

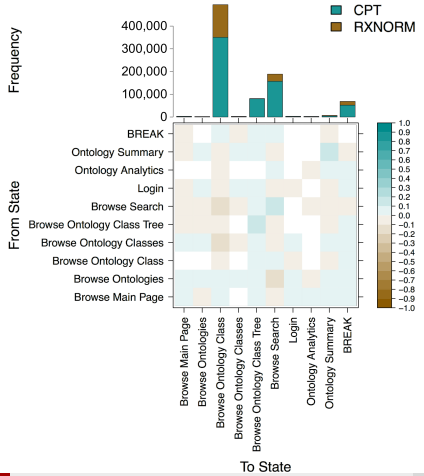


BioPortal Experts



Inspecting Browsing Behaviors for Ontologies

- Are certain browsing types more represented in a specific ontology than other types?



Conclusions & Future Work

Conclusions

- We were able to extract *seven* different browsing types.
- We discussed different exploration strategies of users for ontologies on BioPortal.
- We related differences in ontologies to different exploration strategies.

Future Work

- Comparing browsing (website) and query strategies (API) of users on BioPortal.
- Clustering ontologies according to the browsing behavior of their users to infer features of ontologies which trigger differences in the way users interact with the datasets.

Thanks!

Thanks!

Stationary Distribution Example & Comparison

Given a vector of click actions (2xA, 2xB, 2xC) in a chronologically sorted sequence "ABCABC":

- We define the weighted matrix $\boldsymbol{A}$.
- Then, we calculate the transition matrix $\boldsymbol{P}$.
- The stationary distribution $\boldsymbol{\pi}$ is the left eigenvector of $\boldsymbol{P}$ with eigenvalue 1.0 and satisfies the eigenvalue equation for the matrix $\boldsymbol{P}$: $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$.

Sequence	Weighted Adjacency Matrix	Transition Matrix	Stationary Distribution	Occurrence Count
ABCABC	$\boldsymbol{A} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 1 & 0 & 0 \end{pmatrix}$	$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$	$\boldsymbol{\pi} = \begin{pmatrix} 0.3533 \\ 0.3356 \\ 0.3111 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$
AABBCC	$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{\pi} = \begin{pmatrix} 0.5229 \\ 0.2916 \\ 0.1855 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$