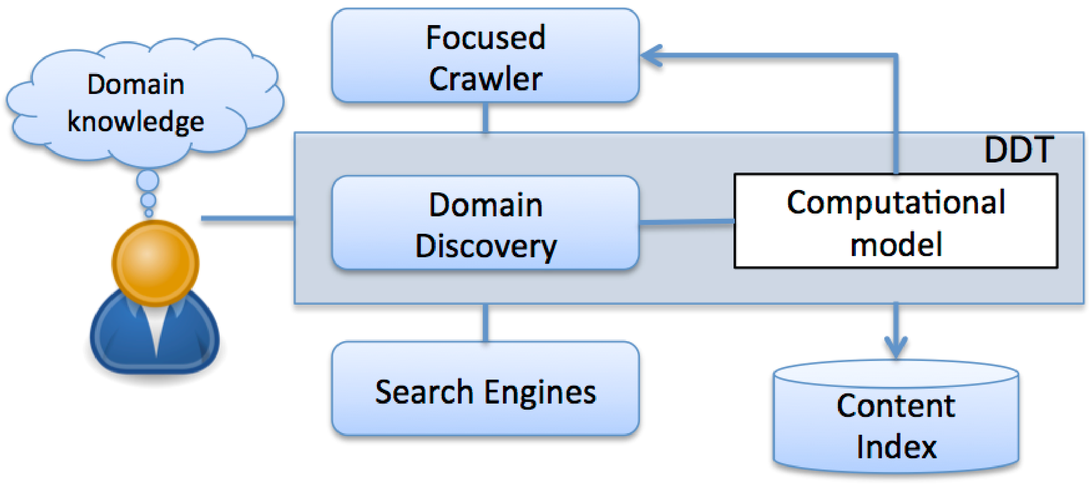
Interactive Exploration for Domain Discovery on the Web

Yamuna Krishnamurthy, Kien Pham, Aecio Santos, Juliana Freire

NYU

DOMAIN DISCOVERY (DD) PROCESS

Iterative process to identify, retrieve and learn information and sources from the Web relevant for a specific information need with a human-in-the-loop

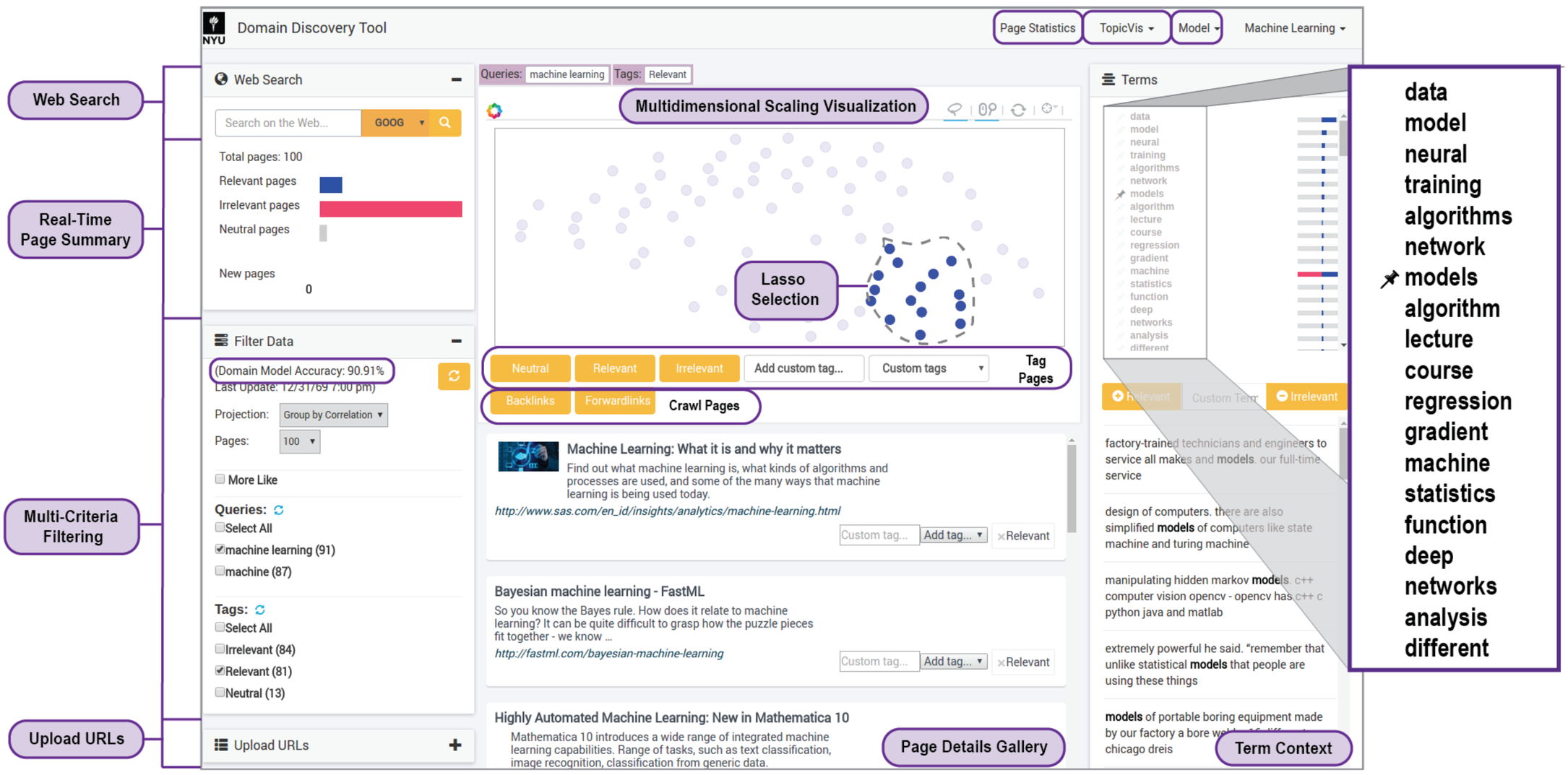


Q: How to Assist Analysts in DD Process?

A: Domain Discovery Tool

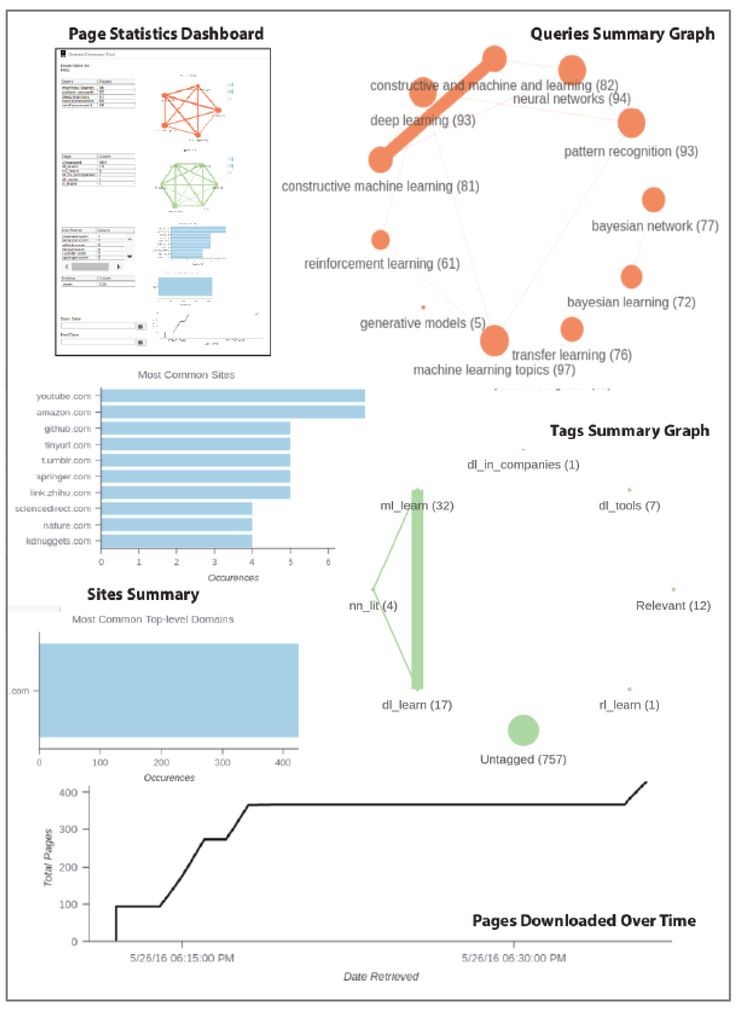
- ***Support exploratory data analysis (EDA) of web pages***
  - Multidimensional scaling visualization of pages (PCA, TSNE)
  - Maintain search context and capture analyst's feedback (Elasticsearch)
  - Summarize search results (Aggregations, Topic Modeling)
  - Streamline annotations
  - Multi-criteria filtering (By queries, tags, date/time, keywords)
- ***Translate the analyst's interactions with the Web pages into a computational model of the domain***
  - Provide quality indicator of domain model (Accuracy of online calibrated SVM with SGD training)
  - Further discovery of domain on the Web with model through:
    - Focused crawling (ACHE)
    - Automated searches (ACHE Seed Finder)

DOMAIN DISCOVERY TOOL (DDT)

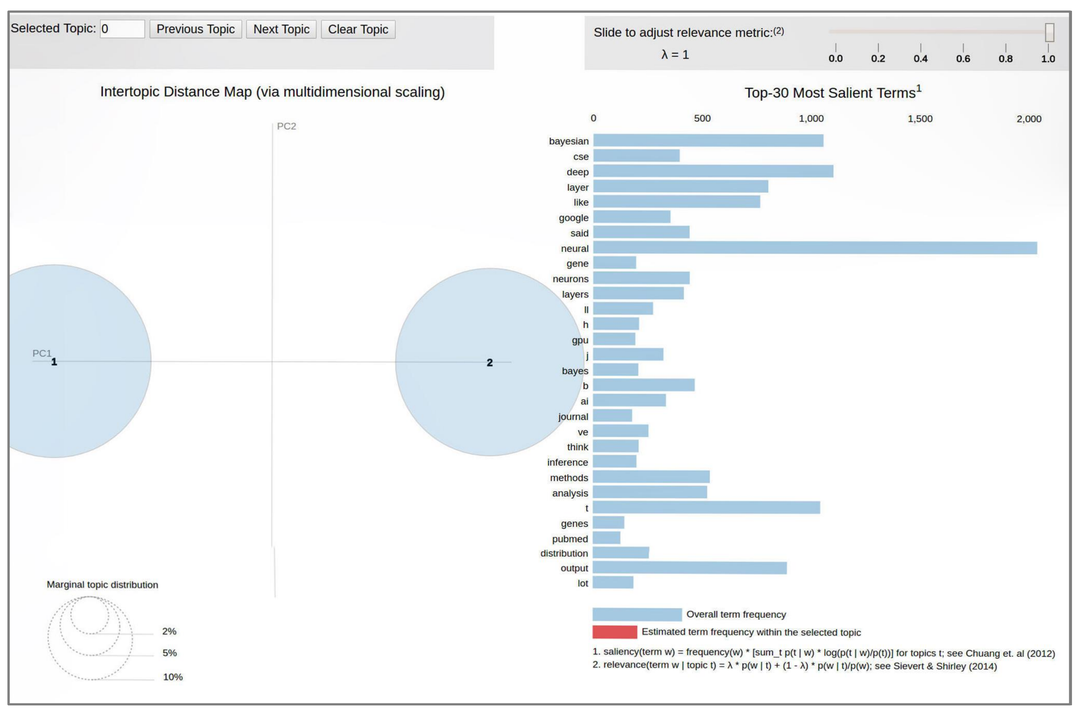


DDT SUMMARIZATION OF SEARCH RESULTS

- ***Page Statistics*** – Aggregations of pages by top level domain, search queries, tags, date/time with cross filtering
- ***Topic Distribution –*** Topics discovered in collected domain pages using LDA or PLSA

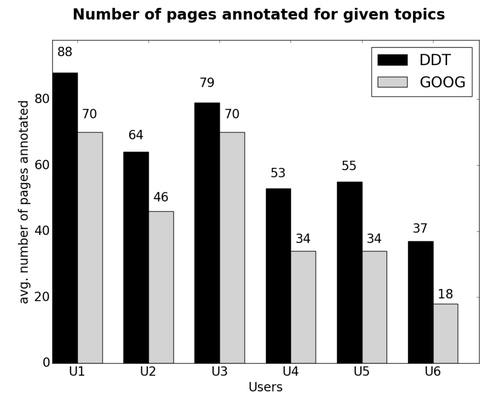


Page Statistics Dashboard

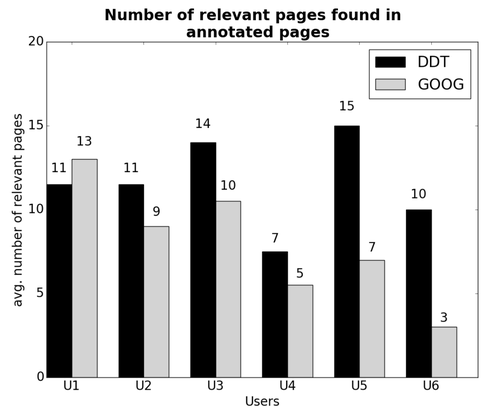


Topic Distribution Dashboard

USER EVALUATION



Users were able to annotate more pages using DDT than Google



Majority of the users found more relevant pages with DDT than Google

TRY DDT

https://github.com/ViDA-NYU/domain_discovery_tool

ACKNOWLEDGEMENTS

This work was funded by the Defense Advanced Research Projects Agency (DARPA) MEMEX program