User Type Clustering to Refine Search and Browse for Educational Resources

Monika Akbar and Clifford A. Shaffer
Department of Computer Science, Virginia Tech
{amonika, shaffer}@vt.edu

Introduction

Educational portals such as Algoviz.org contain rich information resources. A key concern is directing users to specific resources that are of interest to them. While AlgoViz has significant traffic, we cannot count on active user participation in the form of explicit ratings of individual resources. Lacking active user data (e.g., user ratings on resources), we instead use log data to deduce user trends. We describe our techniques for clustering users based on the log data. We show how cluster analysis can be used to improve searching and browsing within AlgoViz. Our approach has the potential to be useful for a wide range of educational resource portals.

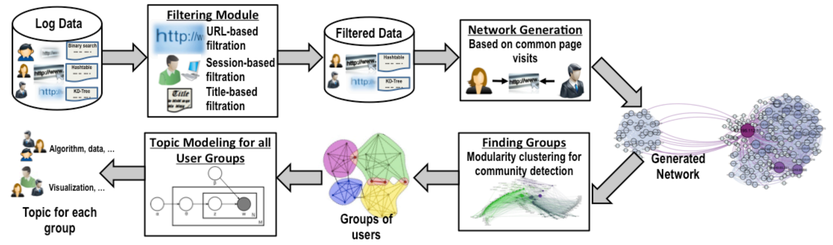
Data Analysis

- **Web metrics**
  - Raw data are stored in various places: Server log, Site log.
  - Sites such as Google Analytics provide more advanced metrics like visits, pageviews, bounce rate, time one site, etc.
- **Analysis overview**
  - Data cleaning: Remove irrelevant pages, bots, crawlers, spammers, etc.
  - Find connections between users/objects
    - Connect two users if they viewed the same pages to create a network.
    - Within this network, find possible user group(s).
  - Update default search/browse ranking based on the group characteristics.

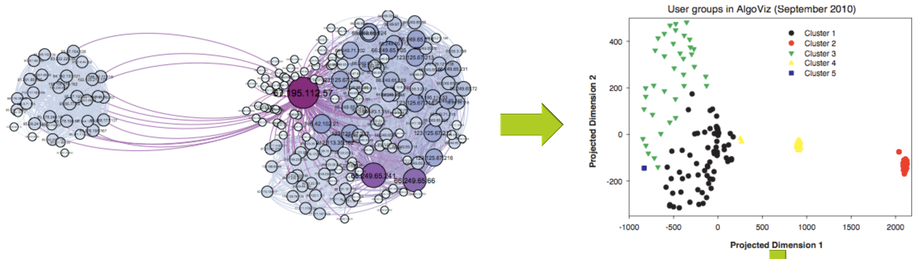
Refine Search and Browse

- **Search and browse**
  - AlgoViz uses Apache Solr to index and rank its content.
  - Not all fields within a page can be indexed by default (e.g., Works, Projects in AlgoViz Catalog Entry) .
- **Refining ranking**
  - We use a custom ranking function that places different weights on AlgoViz-specific fields of an Algorithm Visualization catalog entry.
  - Clusters representing a specific content type are used to add weight to content of that type
    - Top contents c_1, c_2, c_3,...c_n of cluster x that is dominated by a content type of y (e.g., forum, page, catalog entry, etc.), receive certain points.
  - Search results are ordered based on the ranking score.
- **Future directions**
  - For anonymous and registered users, personalize ranking based on which group s/he belongs to.
  - Evaluation: track if the highly ranked documents receive more clicks than the others.

System Overview



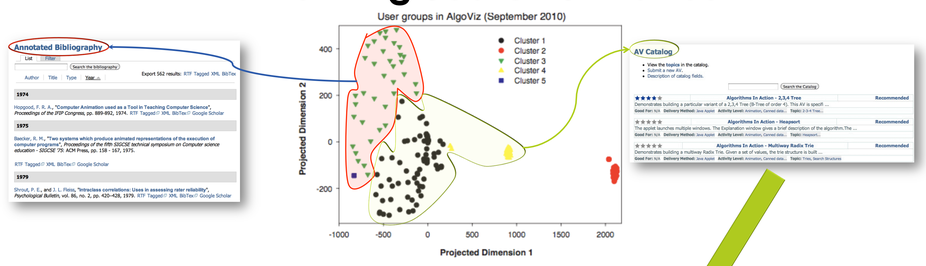
Finding User groups and Interests



Clust.	Top Topic (1)	Contrib.	Top Topic (2)	Contrib.	Top Topic (3)	Contrib.
1	3	0.667	5	0.25		
2	3	0.312	1	0.212	5	0.209
3	1	0.539	2	0.193	4	0.119
4	3	0.75	5	0.156		
5	2	0.254	1	0.216	4	0.213

Topic ID	Words in Topic
1	biblio export xml bibtex rtf set
2	biblio author bibtex export function algorithm
3	algorithms data author trees demo computer
4	visualization sort algorithm structure tree animation
5	biblio java sorting programming learning sorts

Refining Search & Browse



Catalog Entry (CE) Ranking:

$$score(CE_i) = x + y + ... + z$$

$$\text{where, } x = \begin{cases} 20 & \text{if the CE has 'Yes' in the 'Works' field} \\ 0 & \text{otherwise} \end{cases},$$

$$y = \begin{cases} 6 & \text{if the CE is 'Recommended'} \\ 0 & \text{otherwise} \end{cases}, \text{ and}$$

$$z = \begin{cases} 5 & \text{if the CE was present at least } m \text{ times, in a} \\ & \text{cluster dominated by Catalog Entry content type.} \\ 0 & \text{otherwise} \end{cases}$$

This work has been supported by NSF DUE-0836940, DUE-0937863, DUE-0840719