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.

Introduction

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.

Data Analysis

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 c₁, c₂, c₃, ..., cₙ, 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.

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