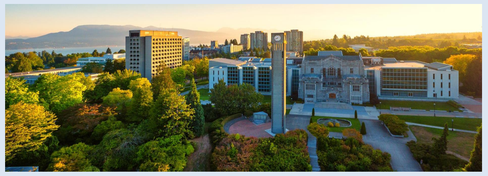


Facilitating Data Exploration, Query Composition, and Query Answer Analysis

Zainab Zolaktaf[1], Rachel Pottinger[1], Reza Babanezhad[1], and Jian Xu[2]
Email : zolaktaf@cs.ubc.ca
[1]Department of Computer Science, University of British Columbia
[2]Microsoft Corporation

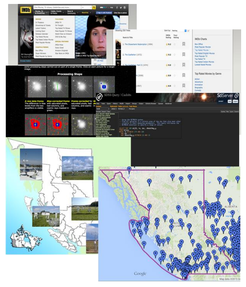
1 Facilitating User Interaction with Data

Abstract

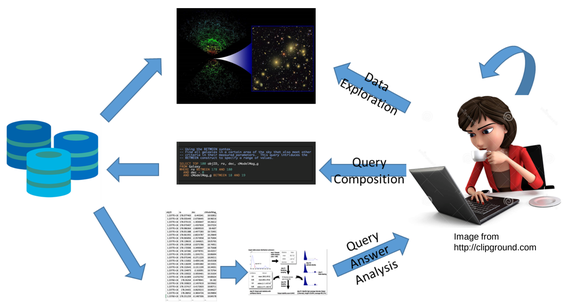
In many domains, such as scientific computing, users can directly access and query data that is stored in large, and often structured, data sources. Discovering interesting patterns and efficiently locating relevant information, however, can be challenging. Users must be aware of the data content and its structure, before they can query it. Furthermore, they have to interpret the retrieved results and possibly refine their query. Essentially, to find information, the user has to engage in a repeated cycle of data exploration, query composition, and query answer analysis. The focus of my PhD research is on designing techniques that facilitate this interaction. Specifically, I examine the utility of recommender systems for the data exploration and query composition phases, and propose techniques that assist users in the query answer analysis phase. Overall, the solutions developed in my thesis aim to increase the efficiency and decision quality of users.

User Interaction with Structured Data Sources

- Explore a database
  - Internet Movie Database (IMDb)
- Compose SQL queries
  - Sloan Digital Sky Server (SDSS)
- Analyze query results
  - Canadian Climate Data



Thesis Outline



- ✓ Facilitating Data Exploration with Recommender Systems
- ➢ Facilitating Data Exploration and Query Composition
- ✓ Facilitating Data Exploration and Query Composition

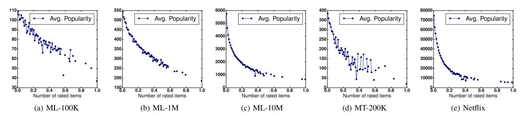
2 Facilitating Data Exploration with Recommender Systems

Problem Overview

- Research Question: How do existing recommendation models perform with regard to exploration and how can they be modified to facilitate data exploration more rigorously?
- Problem addressed: Top-N query recommendation
  - Given a log of prior interaction data, e.g., user ratings on items, recommend a set of N unseen items to each user such that recommendations
    - Are accurate
    - Are novel and include items the user is unlikely to have seen before
    - Improve overall item-space coverage

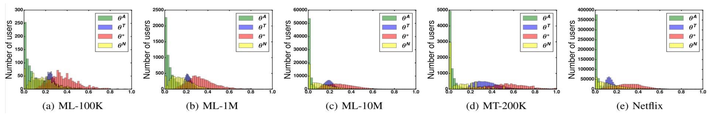
Solution Overview

- Automatically infer user long-tail novelty preference Θ_u
- Integrate Θ_u into a generic re-ranking framework to provide customized balance between coverage and accuracy
- Framework is generic
  - We can plug-in any base recommender, including Most Popular (Pop).

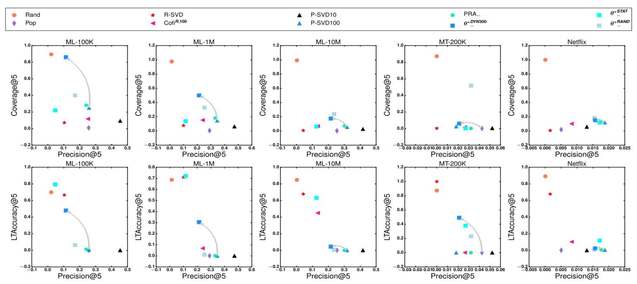


Average popularity of rated items decreases, as the number of rated items increases.

Selected Experimental Results



Histogram of long-tail novelty preference models.



Accuracy vs Coverage vs Novelty. The head of the arrow shows our algorithm, while the bottom shows the underlying accuracy recommender. Our algorithm increases coverage and novelty while maintaining reasonable levels of accuracy.

3 Facilitating Data Exploration and Query Composition

Problem Overview

- Research Question: What are the challenges of recommendation in the database context, and which algorithms are suitable for facilitating interactive exploration and navigation of relational databases?

Ongoing Work …

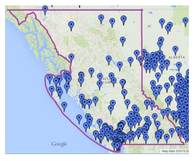
- No clear notion of item
- Instead define aspects to enable schema navigation and data space exploration?
- Problem 1: Top-N aspect recommendation
  - Given a log prior of interaction data, consisting of user SQL queries on database and schema, and a partial query, suggest top-N aspects to the user.
- Problem 2: Top-N query recommendation
  - Given a log of prior interaction data, consisting of user SQL queries on database and schema, and a partial query, suggest top-N queries to the user.

SDSS Logs Schema
SDSS Logs data is ~300 GB

4 Facilitating Query Answer Analysis

Problem Overview

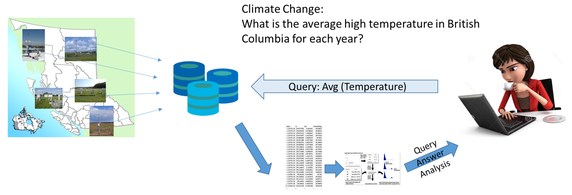
- Research Question: After a query has been submitted, how can we help the user understand and interpret the query answers?
- Problem addressed: Answering aggregate queries in integration contexts
- Value-level heterogeneity



Weather stations are not distributed uniformly across BC

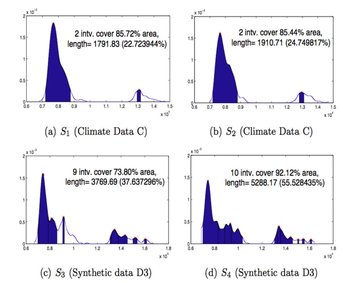
Sources have inconsistent values for the same data points

Solution Overview

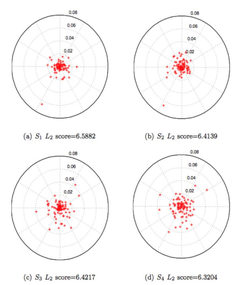


- Define aggregate answers as a distribution of viable answers
- Provide summary statistics for the viable answer distribution
  - Key point statistics
  - High coverage intervals
  - Stability
- Provide algorithms for the efficient extraction of above statistics

Selected Experimental Results



By returning dense areas, the high coverage intervals cover a small percentage of the range of data



Closer to the center, and the denser the points, the more stable the query result