User Driven Architecture for Data Discovery

NSF Award No. 1443070 (2014 – 2017)

PI: Giridhar Manepalli (gmanepalli@cnri.reston.va.us)

Co-PI: Allison Powell (apowell@cnri.reston.va.us)

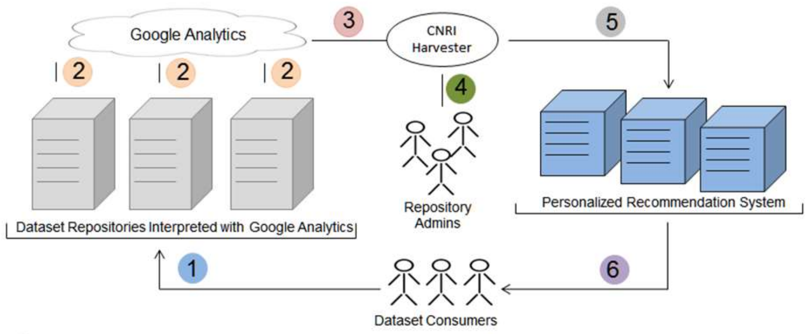
Major Findings, New Capabilities, and Scientific Advancements

Challenge: Make Scientific Datasets Discoverable with Little Effort from Dataset Producers

- Crawling and indexing datasets (e.g., tabular data) do not aid discovery of those datasets; datasets consist of just numbers in many cases.
- Personalized recommendation systems that make use of users' past interests and activities are therefore pivotal because they rely on information held outside of datasets.
- However, significant effort is needed for repository administrators to integrate usage data with recommendation systems.

Our Approach: Automatic Extraction of Dataset Usage Information from Existing Repositories

- Little to zero investment (time & money) expected from repositories to make use of our approach.
- Our harvester extracts access information from Google Analytics into 'items', 'users', and 'clicks' (on repository website), and feeds that data to our recommendation system.
- Repository administrators can configure what constitutes an 'item', a 'user' and a 'click' for determining what exactly is to be recommended to consumers.


Google Analytics
3
CNRI Harvester
5
2
2
2
4
Dataset Repositories Interpreted with Google Analytics
Repository Admins
Personalized Recommendation System
1
6
Dataset Consumers

1 Consumers interact with repositories.

2 Google Analytics tracks usage.

3 Our harvester pulls usage data from repositories.

4 Repository administrators control and configure what information is available to Recommendation System.

5 Our harvester makes curated usage information available to Recommendation System.

6 Recommendation System aids consumers by providing them meaningful dataset recommendations.

Community Integration and Broader Impact

Community Integration

- Collaborating with Vermont Monitoring Cooperative to apply our advancements to recommendation systems within the 'forested ecosystem' community.
- Under discussion with University of Arizona to apply our advancements within their Cyverse System.

Recommendation System for DOI-based Datasets

- DOIs are accessed globally and by users from disparate communities.
- Pilot makes use of access logs to produce personalized recommendations.
- Pilot is made available here: https://datacriterion.org

Technology and Software

Several Algorithmic Improvements for Personalized Recommendations are Currently Studied

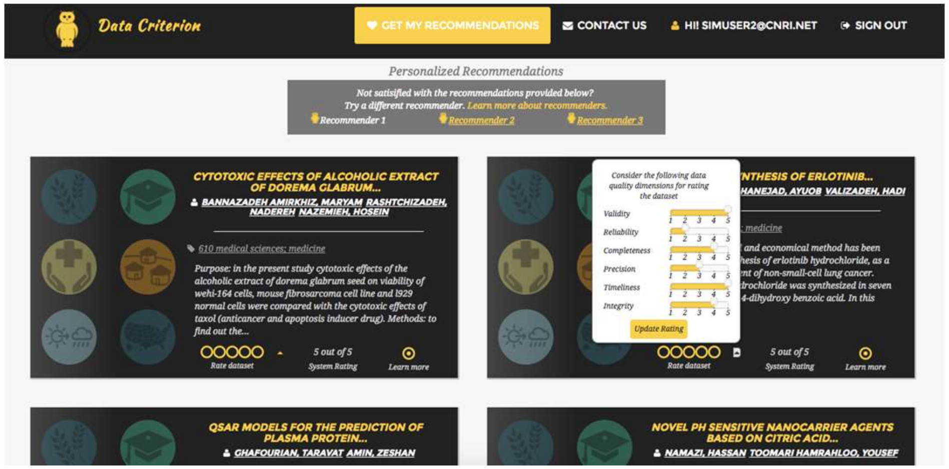
Recommendation Algorithm	How it works?	Our Improvement
Latent Factor Approach for Implicit Feedback	Extracts user interest from clicks made by users on web pages about a given item.	Not all web pages about a dataset are related to the dataset the same way; we are leveraging that variance in web page differences by inferring different weights for different pages through regression techniques.
Content based Approach	Uses metadata about items to match user interests and items.	We are casting a wider net by matching interests of 'similar users' against items' metadata.
Implicit Feedback Normalization	Different users exhibit different 'click' behavior, which requires normalization.	We are extracting the behavioral uniqueness of a user compared to fellow users by normalizing click-through rate using approaches adapted from information retrieval.

Software Development

- We developed a scalable recommendation system that can produce personalized recommendations over millions of datasets.
- Software will be released open source by Fall 2017.
- System can provide recommendations using multiple approaches; some are currently being added:

User Interface Design in the Pilot

- Many scientific dataset discovery websites are text heavy; we designed a tile-based user interface in our pilot that balances graphics with text.
- Rating widgets in several websites do not provide guidance to users on how to rate; we designed a mechanism to convert user opinion about validity, reliability, dataset completeness, etc., into a numerical rating scale.
- We provided options to switch between different recommendation algorithms.



Personalized Recommendations

Not satisified with the recommendations provided below?
Try a different recommender. Learn more about recommenders.
Recommender 1 Recommender 2 Recommender 3

CYTOTOXIC EFFECTS OF ALCOHOLIC EXTRACT OF DOREMA GLABRUM...
BANNAZADEH AMIRKHIZ, MARYAM RASHTCHIZADEH, NADEREH NAZEMIEH, HOSEIN

610 medical sciences; medicine

Purpose: in the present study cytotoxic effects of the alcoholic extract of dorema glabrum seed on viability of wehi-164 cells, mouse fibrosarcoma cell line and l929 normal cells were compared with the cytotoxic effects of taxol (anticancer and apoptosis inducer drug). Methods: to find out the...

5 out of 5

Consider the following data quality dimensions for rating the dataset

Validity	1 2 3 4 5
Reliability	1 2 3 4 5
Completeness	1 2 3 4 5
Precision	1 2 3 4 5
Timeliness	1 2 3 4 5
Integrity	1 2 3 4 5

NTHESIS OF ERLOTINIB...
HANEJAD, AYUOB VALIZADEH, HADI

; medicine

and economical method has been hesis of erlotinib hydrochloride, as a nt of non-small-cell lung cancer. drochloride was synthesized in seven 4-dihydroxy benzoic acid. In this

5 out of 5

QSAR MODELS FOR THE PREDICTION OF PLASMA PROTEIN...
GHAFOURIAN, TARAVAT AMIN, ZESHAN

NOVEL PH SENSITIVE NANOCARRIER AGENTS BASED ON CITRIC ACID...
NAMAZI, HASSAN TOOMARI HAMRAHLOO, YOUSEF

Corporation for National Research Initiatives

NSF DIBBS PI Workshop – January 11 – 12, 2017 – Arlington, VA