Touring Dataland?
Automated Recommendations for the Big Data Traveler

William Agnew
Georgia Institute of Technology
wagnew3@gatech.edu

Michael Fischer
Univerity of Wisconson-Milwaukee
fisch355@uwm.edu

Kyle Chard and Ian Foster (advisors)
University of Chicago
lastname@uchicago.edu

ABSTRACT

We explore how recommendation techniques can be adapted and applied to big data science to predict data storage locations to users. Specifically, we present a collection of heuristics that use features specific to big data science. We combine these heuristics into a single recommendation engine using a deep recurrent neural network. We show, via analysis of historical Globus operations, that our approaches can predict the storage locations accessed by users with 78.2% and 95.5% accuracy for top-1 and top-3 recommendations, respectively.

1. INTRODUCTION

Big data scientists, like travelers to a new land, are faced with the daunting task of discovering which (storage) locations contain interesting attractions (i.e., research data). This task is complicated by the rapid growth of research data and the increasing number of accessible storage systems. For example, the Globus [2] research data management service provides access to more than 10,000 data locations. A new user connecting to the Globus network has few guideposts as to which of these locations may be useful for finding or placing data.

Commercial services, such as travel websites, provide valuable user-specific recommendations derived from analyzing huge amounts of usage data. These recommendations reduce the complexities associated with trawling through vast amounts of data and improve user experiences [1]. Recommendations have also been applied to support scientific users, for example by predicting workflow components [5] and coauthors [4].

We explore here how recommendation approaches can be adapted and used to recommend storage locations to users. To do so, we develop a collection of specialized heuristics that consider unique features of scientific big data. We evaluate our approach using Globus, a hosted service that provides research data management capabilities across a vast network of distributed storage locations (called "endpoints").

Supercomputing '16 Salt Lake City, UT, USA

DOI:

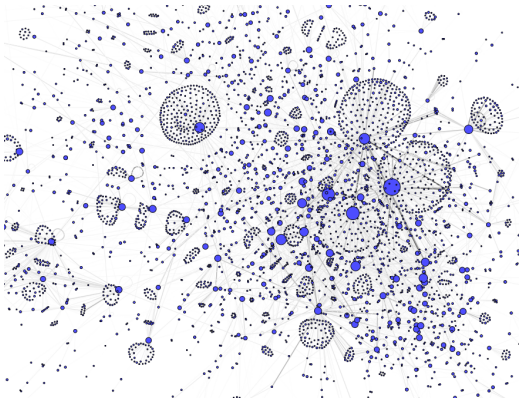


Figure 1: Globus Network. Endpoints are represented as vertices. Edges represent transfers between endpoints.

The complexity of this recommendation task is illustrated in Figure 1. The network includes nearly one million historical transfers submitted via the web interface. These transfers include 23,650 endpoints and 16,412 users, many of whom infrequently use the service. The matrix of historical user/endpoint pairs is sparse, with approximately 0.01% of potential pairs present. The long tailed usage of Globus is further highlighted in Figure 3. The goal of our work is to develop an online recommendation engine that can suggest endpoints to users, as shown in Figure 2.

2. HEURISTICS

Globus stores detailed records regarding users and their historical usage. We use this information to develop a collection of endpoint recommendation heuristics:

History: A baseline heuristic that predicts the most recently used source (S) / destination (D) endpoint.

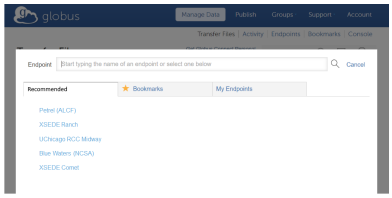
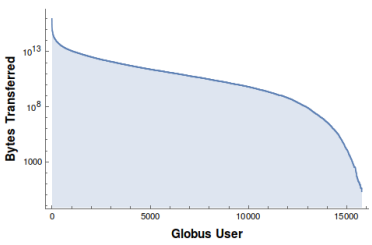
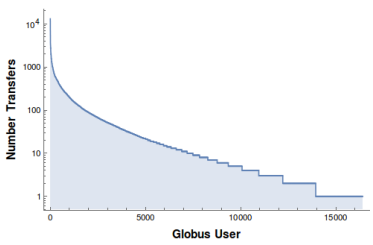


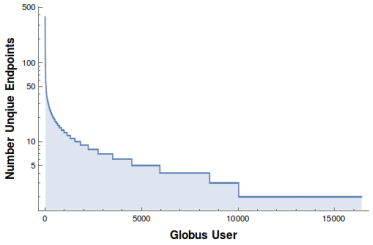
Figure 2: Mockup of recommendation interface.



(a) Transfer volume per user



(b) Transfers per user



(c) Distinct endpoints per user

Figure 3: Long-tailed distributions of Globus usage

Markov Chain: Using a user/endpoint transition matrix the heuristic predicts the S/D endpoint based on the most likely endpoint transition given that user's previously used S/D endpoint.

Most Unique Users: The most likely S/D endpoint is the S/D endpoint with the most unique users.

Institution: The most likely S/D endpoint for a user is the endpoint with the most unique users that is also owned by a user belonging to the same institution.

Endpoint Ownership: The most likely endpoint is the endpoint most recently created by the user.

These heuristics model different aspects of the Globus ecosystem and therefore perform well for different classes of users. To combine the strengths of each heuristic we trained a deep recurrent neural network [3] on the series of endpoints chosen. The neural network is given the heuristics' recommendation weightings and some additional user and endpoint information, it re-weights heuristic recommendations and chooses the most highly re-weighted recommendation.

3. RESULTS

We explore recommendation accuracy by evaluating historical Globus transfers. We train our neural network on the first 500,000 transfers and evaluate accuracy using the remaining 450,000 transfers. We measure performance using two metrics. Transfer accuracy (Figure 4): the average number of endpoints predicted correctly for all transfers; and user accuracy (Figure 5): the average accuracy per user, where a user's accuracy is the percentage of that user's endpoints that are correctly predicted. We compare accuracy when predicting the top-1 and top-3 recommendations.

The *most unique users*, *institution*, and *owned endpoints* heuristics perform significantly worse than the other heuristics. However, when heuristic accuracy is compared against user history size (Figure 6), we see that the *history* and *Markov* heuristics perform poorly when users have few previous transfers. By combining heuristics, the neural network is able to outperform all individual heuristics. For example, when there is little user history, the neural network increases the weighting of the *unique users* and *institution* heuristics.

4. CONCLUSION

We have developed and evaluated a collection of heuristics for recommending data locations. By combining these heuristics using a neural network, we correctly predict endpoints with 78.2% and 95.5% accuracy for top-1 and top-3, respectively. In future work we aim to characterize, model, and improve the usage of scientific big data by analyzing the performance of heuristics for different user profiles.

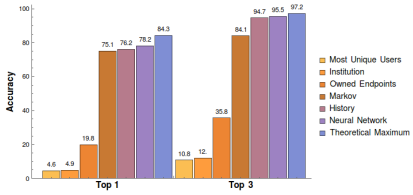


Figure 4: Transfer recommendation accuracy

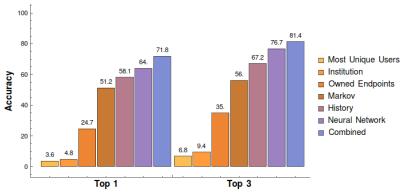


Figure 5: User recommendation accuracy

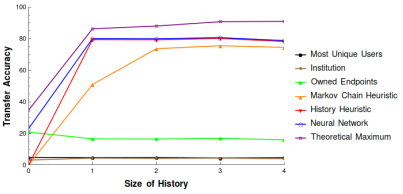


Figure 6: Top-1 transfer accuracy vs. history size

5. REFERENCES

[1] A. Ansari, S. Essegaier, and R. Kohli. Internet recommendation systems. *Journal of Marketing research*, 37(3):363–375, 2000.

[2] I. Foster. Globus online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*, 15(3):70, 2011.

[3] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[4] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 121–128, 2011.

[5] J. Zhang, W. Tan, J. Alexander, I. Foster, and R. Madduri. Recommend-as-you-go: A novel approach supporting services-oriented scientific workflow reuse. In *Proceedings of the IEEE International Conference on Services Computing (SCC)*, pages 48–55, July 2011.