The DataBridge

Arcot Rajasekar
University of
North Carolina at Chapel Hill
rajasekar@unc.edu

Hye-Chung Kum
Texas A&M University
Health Science Center
kum@srph.tamhsc.edu

Merce Crosas
Harvard University
IQSS
mcrosas@iq.harvard.edu

Jonathan Crabtree
University of
North Carolina at Chapel Hill
jonathan_crabtree@unc.edu

Sharlini Sankaran
University of
North Carolina at Chapel Hill
sharlini@unc.edu

Howard Lander
University of
North Carolina at Chapel Hill
howard_lander@unc.edu

Tom Carsey
University of
North Carolina at Chapel Hill
carsey@unc.edu

Gary King
Harvard University
IQSS
king@harvard.edu

Justin Zhan
North Carolina A&T
State University
zzhan@ncat.edu

ABSTRACT

The rapid increase in the amount and diversity of data collected in multiple scientific domains implies a corresponding increase in the potential of data to empower important new collaborative research. However, the sheer volume and diversity of these datasets present new challenges in locating data relevant to a particular line of research. The explosion of data has taken two primary forms: the emergence of extremely large individual datasets and the proliferation of a massive number of small to moderately sized datasets. This project focuses on the latter – a region sometimes called the long tail of science. Realizing the potential of data in the long tail of science requires investigating algorithms and designing tools that will enable important new multidisciplinary collaborative research at scales ranging from small teams focused on relatively simple issues to large collaborations investigating grand challenge problems. In short, we need a way to make long tail data something greater than the sum of its parts. The DataBridge is a new e-Science collaboration environment tool being designed and developed specifically for this purpose. The DataBridge will exploit a rich set of sociometric network analysis (SNA) and relevance algorithms to define different types of semantic bridges that link diverse datasets. DataBridge will enable discovery of relevant datasets and methods by computing metrics in multiple spaces of relevancy – different ways data can be related to each other – by metadata and ontology, by pattern analysis and feature extraction, through usage tools and models, and via human connections. In the initial phase of this project, we will integrate DataBridge with the Dataverse Network (the largest social science data repository) to test and validate the new tools with real-world data. In this paper, we discuss the motivation for the DataBridge project, introduce concepts in SNA relevant for long tail science data and their application in designing the DataBridge, and detail our anticipated implementation strategy.

I INTRODUCTION

The term ‘Big Data’ is often used as a synonym for ‘very large’ datasets that pose problems in management and analysis due to their sheer size. Generally these datasets are generated by a limited number of data producers and from a small number of distributed experimental or commercial sites each having a small number of homogenous formats, types, or schemas. Provenance and description of these datasets are well defined, with metadata that is sometimes generated automatically as the data is collected. The data may contain thousands of variables or more, but their description is often contained in a single coherent metadata record. The location, availability, and usage of this type of Big Data is often linked to large observatories or data centers, with dedicated staff responsible for collecting the data and managing the associated metadata. These data have real problems that require real solutions, but they represent only one aspect of the Big Data

© ASE 2013

problem. While these individually very large datasets are becoming more common, far and away the largest proportion of scientific research still uses datasets from what Palmer et al. [1, 2] described as the “long tail of science” data – the massive number of relatively small datasets.

These long tail datasets are small and easily managed individually. Collectively, however, they are quite heterogeneous and growing rapidly in number. Taken together these datasets represent a fundamental Big Data problem. A major obstacle to generating new knowledge from this type of research data is the atomistic way in which the vast majority of it is currently collected, archived, and analyzed. Long tail data suffer from sparse provenance and metadata, which even when available can be highly idiosyncratic. Moreover, these data are often highly distributed (stored in multiple locations, often in personal repositories), not very well organized or managed, and not easily re-discoverable and re-usable. In other words, while one kind of Big Data problem is the single massively large dataset, a more fundamental problem is the massive number of smaller datasets that exist in almost total isolation from one another. Building bridges between data of all types and sizes is one of the foundational challenges facing researchers in the era of Big Data.

Data from the long tail of science contains rich information that can be used and reused to enable new scientific discoveries. To maximize this reuse, it must be as easy as possible to discover, access, and analyze relevant data. Discovering long tail data is hard because the data is often distributed in personal workspaces with little attempt made at data publication. Finding relevant data is made even more problematic by the difficulty in defining relevancy metrics for scientific datasets. Accessing relevant data is not easy when the data are distributed, not well documented, and in heterogeneous and possibly unique formats. These same characteristics also inhibit the analysis of data.

DataBridge [3, 4] is an NSF-funded collaboration between University of North Carolina at Chapel Hill, Harvard University, and North Carolina A&T State University aimed at developing an e-Science collaboration environment tool designed specifically to measure the relevancy of different datasets. Measurement will be based on four general components: the contents of the data itself, contextual information about the data, producers and consumers of the data, and methods used to create and analyze the data. Using these relationships, profiles and clusters for datasets can be discovered and maintained that will help scientists seek, search, browse and identify data relevant to their research.

Even though a large number of datasets still remain only stored in personal workspaces, without formal organization and metadata, successful efforts have been made to provide centralized data repositories to properly share, manage and archive scientific datasets. These efforts include the Dataverse Network (DVN) [5, 6, 7, 8, 9] and the Integrated Rule-Oriented Data System (iRODS) [10, 11, 12, 13, 14, 15]. The initial source for example datasets and metadata for the DataBridge research effort will be the Dataverse Network (DVN) because it offers a rich set of real-world structured data and metadata that will help validate the algorithms and analysis. The DVN is an e-Science collaboration environment used to publish, share, reference, extract and analyze research data [5, 6, 7, 8]. A DVN hosts multiple, individually branded dataverses containing studies or collections of studies, and each study contains cataloging information that describes the data. The data section may contain primary and secondary data, code, and documentation files. The DVN serves many fields and uses, and complies with many content-agnostic standards. The repositories hosted at the University of North Carolina and at Harvard University hold more than 50,000 research studies with more than 700,000 data and supplementary files in social science. Recently, the software has been adopted for astronomy data and is being extended to support sensitive data in health and medical domains. One of the main advantages of the DVN is that it facilitates long term access and good archival practices for a researcher's data while the researcher retains control of, and recognition for, the data he or she deposits.

A second source of datasets for this effort will be instantiations of the integrated Rule-Oriented

Data System (iRODS). iRODS [10] is a data grid middleware [11, 12, 13, 14, 15]. It provides many facilities for collection building, managing, querying, accessing, and preserving data in a distributed data grid framework. The iRODS system applies policy-based control when performing these functions. iRODS provides many advanced features, such as access to many types of storage (including third-party storage, like the Amazon Simple Storage Service, cloud computing, and the Virtual Computing Lab), data grids federation, and many diverse interfaces and access mechanisms.

The remainder of this paper proceeds as follows: In section II, we introduce the concept of sociometric network analysis used to measure the relevance relationships among scientific datasets. In section III we describe the architecture of the DataBridge system. In section IV, we present relevant work. We conclude the paper in section V.

Table 1. Similarity Measures (SM)

SM_1	The Euclidean distance (L_2 -norm)	$d_{AB}^{E} = \sum_{k=1}^{n} \sqrt{(a_k - b_k)^2}$
SM_2	The Manhattan distance (L_1 -norm)	$d_{AB}^{M} = \sum_{k=1}^{n} \lvert a_k - b_k \rvert$
SM_3	The L_∞-norm	$L_1 d_{AB}^{\infty} = \max_{k \epsilon [1,n]} \lvert a_k - b_k \rvert$
SM_4	Cosine similarity Range of ρ_{AB} is $[0,\pi)$	$\rho_{AB} = arccos \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\sqrt{\sum_{k=1}^{n} a_k^2} \sqrt{\sum_{k=1}^{n} b_k^2}}$, where $\boldsymbol{a} \cdot \boldsymbol{b}$ is the dot product of the vectors **a** and **b.**
SM_5	Dissimilarity measure in structural equivalence	$d_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2}$, where $\boldsymbol{A}$ is the adjacency matrix**.**
SM_6	Overlap in neighborhood	$\omega_{ij} = \frac{\lvert \Gamma(i) \cap \Gamma(j) \rvert}{\lvert \Gamma(i) \cup \Gamma(j) \rvert}$, neighborhoods Γ(i) and Γ(j) of vertices i and j
SM_7	Pearson correlation between columns or rows	$C_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n \sigma_i \sigma_j}$, where the averages $\mu_i = (\sum_j A_{ij})/n$ and the variances $\sigma_i = \sqrt{\sum_j (A_{ij} - \mu_i)^2 / n}$.

II SOCIOMETRIC NETWORK ANALYSIS

The DataBridge project seeks to build tools that will create a semantically-rich sociometric network of data that includes data from multiple disciplines. The founder of sociometry, Jacob L. Moreno, defined the field as “the inquiry into the evolution and organization of groups and the position of individuals within them [16]”. Moreno developed the first sociogram, a graph structure that represents social relationships by symbolizing individuals as nodes connected by links, or edges. These basic constructs have been extended in a number of ways by modern information network analysis. By applying sociometry and it’s derived techniques to the study of datasets the DataBridge project hopes to greatly increase the scientific value of data in the long tail of science.

Detecting community structure, or clustering, in real systems is of great importance in sociology, biology, and computer science; disciplines where systems are often represented as graphs. Solving the community detection problem is still a difficult challenge, despite the huge effort [17, 18, 19] of a large interdisciplinary community of scientists working on it over the past few years. We will explore a broad range of community detection methods for DataBridge.

The problem of community detection in graphs is not well defined and is therefore not amenable to the highest level of analytic rigor. The first task in graph clustering is to look for a quantitative definition for community. No definition is universally accepted [17, 18, 19] but for our purposes we define communities as groups of nodes similar to each other. The main goal is to minimize inter group node similarity and maximize intra group node similarity. The similarity between each pair of nodes can be computed in several ways; examples include with respect to a reference property, local or global, and whether or not the nodes are connected by an edge. At the end of the process, each node ends up in the cluster of nodes to which it is most similar. A set of similarity measures is listed in Table 1. If we can express the nodes in a Euclidean space, we can use the distance between every pair as a measure of their similarity. Given the two data points A and B, one could use any norm L_m, like the Euclidean distance (L_2 -norm [SM_1]), the Manhattan distance (L_1-norm [SM_2]), or the L_∞-norm [SM_3] to measure their relative similarity. The cosine similarity [SM_4] is yet another distance based norm. For nodes that cannot be embedded in space, we can derive the similarity

© ASE 2013

from the adjacency relationships between nodes. One possibility is to define a non-Euclidean distance metric between nodes [20]. For this, we use the concept of structural equivalence [SM_5] [18]. Two nodes are structurally equivalent if they have the same neighbors, whether or not they are adjacent. If i and j are structurally equivalent, by definition $d_{ij} = 0$. We can also measure the overlap between the neighborhoods Γ(i) and Γ(j) of vertices i and j, given by the ratio between the intersection and the union of the neighborhoods [SM_6]. Another possibility, similar to structural equivalence is the Pearson correlation [SM_7]. Yet another metric is the number of edge- (or node-) independent paths between two nodes. Independent paths are paths that do not share any edges; their number is a measure of the maximum flow that can be conveyed between the two nodes [21]. We could also attempt to measure all paths running between two nodes. Unfortunately the total number of paths is infinite, but we can solve this difficulty by using a weighted sum of the number of paths. For instance, paths of length L can be weighted by the factor α^l, with $\alpha < 1$. We could also follow Estrada and Hatano [22, 23], and weigh paths of length L with the inverse factorial $1/l!$. In both cases, the contribution of long paths is strongly suppressed and can be safely ignored.

Another set of metrics of node similarity concerns properties of random walks on graphs. Commute-time between a pair of nodes is the average number of steps needed for a random traversal to complete a round trip between a pair of nodes. Saerens and coworkers [24, 25, 26, 27] have extensively studied commute-time and it's variants as a (dis)similarity measure: the commute time is inversely related to the similarity. The commute-time [28] is closely related to the resistance distance [29], which measures the effective electrical resistance between two nodes if the graph is turned into a resistor network. White and Smyth [17, 18] used the average number of steps needed to reach the target node for the first time from the source. Harel and Koren [30] proposed to build measures out of quantities like the probability to visit a target node in no more than a given number of steps after it leaves a source node and the probability that a random walker starting at a source visits the target exactly once before hitting the source again. Another quantity used to define similarity measures is the escape probability, defined as the probability that the walker reaches the target node before coming back to the source node [31]. The escape probability is related to the effective conductance between the two nodes in the equivalent resistor network. Other authors have exploited properties of modified random walks. For instance, the algorithm in [31, 32] used similarity measures derived from Google's PageRank process [33, 34].

III THE DATABRIDGE

The DataBridge system will analyze linkages between datasets. It will gather data, metadata, usage and other information, and apply SNA algorithms to develop a mapping of datasets connected by multi-dimensional relationships. In this multi-dimensional network, sub-graphs, clusters, and cliques will be used to inform the discovery of relevant datasets. The system will eventually gather information from multiple data resources maintained by individuals, projects, regional or disciplinary repositories, and national collaboratives in order to provide a semantically-rich interface to discover relevant datasets based on relationships between data, users, methods and metadata. Internally, it will apply several relevance algorithms to build a knowledge base and generate interconnected networks. The DataBridge will provide a venue for scientists to publish, discover, and access datasets of importance and to find others engaged in similar and pertinent research.

1 DATABRIDGE ARCHITECTURE

DataBridge is an indexing mechanism for scientific datasets, similar to web search engines that help find web pages of interest. Unlike web search engines that use the textual content of a web page and hyperlinks to identify its relevance to a query, the search space for scientific datasets is quite different and will need external resources such as tags, metadata, contexts, and naming conventions to identify relevancy. Scientific datasets by themselves provide very sparse information content for search and discovery. A resource discovery system for scientific datasets should provide a

© ASE 2013

rich set of tools for mining information about the dataset and context.

Given a set of criteria such as a sample dataset, DataBridge will query its knowledge base to find relevant datasets that are close to the initial dataset based on the given criteria. To do this, the architecture of DataBridge will consist of three functional units shown in Figure 1: a *data gatherer* which interacts with data repositories to gather information about scientific datasets, a *relevance engine* which integrates information about datasets into a relevance network based on sociometric analysis, and a *web-based user interface* which searches for relevant datasets and gathers information through crowd sourcing and collaborative tagging.

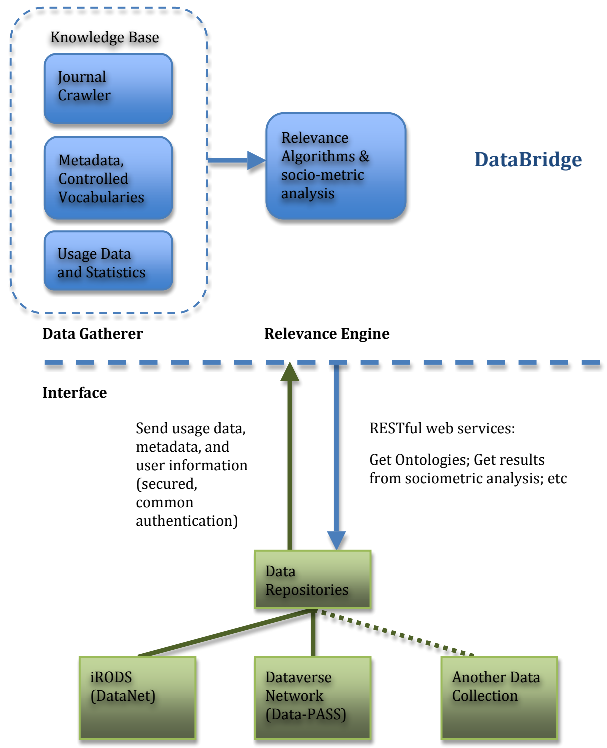
Knowledge Base
Journal Crawler
Metadata, Controlled Vocabularies
Usage Data and Statistics
Relevance Algorithms & socio-metric analysis
DataBridge
Data Gatherer
Relevance Engine
Interface
Send usage data, metadata, and user information (secured, common authentication)
RESTful web services:
Get Ontologies; Get results from sociometric analysis; etc
Data Repositories
iRODS (DataNet)
Dataverse Network (Data-PASS)
Another Data Collection

Figure 1. DataBridge Server Architecture

1.1 THE DATA GATHERER UNIT

The Data Gatherer Unit is a backend interface to data grids, data networks, and data repositories that will gather information about new and modified data from these sites. Several types of data providers and data integrators need to be queried to gather information about the scientific datasets they contain. Different data provisioning systems use different APIs, methods, and services for accessing data and metadata. Moreover, many of the datasets are located in individual labs or in personal repositories and are not available for access from the outside. As a result, DataBridge must provide a simple way to register data into its system and provide access.

For our initial research, we will gather data from two different types of data provisioning systems that have widespread adoption in the scientific community: the iRODS data grid system and the DVN system. These systems were chosen because of the diverse variety of domain data being disseminated through them and their access to rich metadata. Several of the research team members have been involved in the development of iRODS and the DVN. Another reason to choose systems such as iRODS and DVN is the support they provide for user usage information. Mining this information will provide different kinds of sociometric rankings that will help in focusing and narrowing search results.

Not all domain scientists and data providers will be using one of the above systems supported by DataBridge. To ease integration and publication through DataBridge, we will eventually provide a simple server system that can be installed by users to provide access to their data. One can view this as a self-publication mechanism for scientists to share their data (still under their control) with a wider audience by linking them through the DataBridge. We will design and develop this server system using our experience in iRODS and DVN to provide a publication service that is simple to install and maintain and captures all the sociometric data needed for our analysis.

1.2 THE RELEVANCE ENGINE

The relevance engine is the core of the DataBridge system design. It applies the metadata stored in the knowledge base of the Data Gatherer Unit to build, store, and access a multi-dimensional relevance network for all datasets known by the DataBridge system. The relevance network consists of nodes and edges, with each dataset forming a node and

© ASE 2013

connections between them defining edges. There can be multiple edges connecting nodes representing the multiple sociometric relationships derived between the datasets. The relevance engine will be extensible and adaptive. The architecture for the relevance engine will include a rule-based system of the Event-Condition-Action (ECA) type, which has been successfully used in iRODS. The Events will include data gathering, user access, and crowd sourcing input. The Conditions will be on types of datasets, metadata conditionals, and other user-defined parameters. The actions will be the application of relevance algorithms of the proper form to rebuild the relevance network. The extensibility of the system comes from the rule-based architecture. New rules and algorithms can be added in a modular fashion. We expect DataBridge to be primarily implemented in the Java language.

We expect to build relevance networks connecting datasets to each other that will be computed multi-dimensionally based on the various criteria outlined in this paper. We will analyze and define a range of relevance dimensions for datasets across scholarly domains, based in part on input from domain scientists through crowd sourcing. Let us illustrate with some examples. Suppose dataset A and dataset B are analyzed using a particular method M. Even if they are not applied together, the two datasets have relevance because of the common application of method M. For example, if a statistical procedure available in R is applied to two datasets to obtain an aggregation measure, the two datasets become procedurally relevant (high similarity index in the method plane) even if one set of data is about rainfall and the other about astronomy. One idea that we will study is that if A and B are relevant through M, and A is also applied through another method M', then B is also a candidate for M', and M' can be suggested as a relevant method for B. The aim is to cross-pollinate methods and applications across scientific domains. Another dimension of data-to-data relevance is through users. If user U has used data A and B, and user U' is interested in using A, then the system can suggest B to user U'. We will define relevance rules (without compromising privacy) to extract such sociometric data. Relevance algorithms and methods of interest are addressed in more detail in Sections III.2 and III.3.

1.3 THE USER INTERFACE

The User Interface for DataBridge plays three roles: a query interface for users to discover relevant data, a crowd sourcing interface for users to add more information to the knowledge base and improve the relevance calculation, and an ingestion interface that can mediate between the user and data repositories and data grids. These interfaces will be web service-based so that external products and partners can integrate the DataBridge functionality into existing packages such as DVN and iRODS. We will also implement a graphical user interface that will utilize the web services.

2 RELEVANCE INFORMATION

The Data Gatherer unit will harvest data and metadata that will be analyzed to discover relevance between datasets. The gathered information will be stored in the knowledge base. The types of information that will be gathered are discussed in this section.

Datasets will be mined for important contextual information available in the format of the dataset. We will implement several “scrapers” to mine metadata information. These scrapers will be particularly useful for legacy data in spreadsheets and text documents that contain heading, title, and other information which when matched against an ontology of a particular discipline will provide contextual information for relevance. Other types of information that can be mined include temporal, spatial, and geographic information that are embedded in the datasets themselves. Specific analysis algorithms can also be performed on particular types of data, and the features of the data can be used as metadata in the knowledge base. For example, one can extract minimum, maximum, and average rainfall information from a particular dataset and associate it as metadata. This can be used to quickly find which regions have heavier rainfall on a particular time period and can be used to associate one dataset with another. Similar functional features can be detected from

© ASE 2013

environmental, astronomical, sociological, and other types of data. The Data Gathering Unit will provide the capability to register new feature detection algorithms, associate them with types of datasets, and automatically use them to cull important information into the relevance network knowledge base. These feature detection functions can be simple (identify all dates) or complex (perform a workflow to extract a hurricane from satellite images).

Metadata will also play an important role in deriving relationships between datasets. We will use structured, semi-structured, and key-value metadata to show how one dataset is connected to another. One can view such connections as semantic relationships between datasets based on common metadata. Additionally, associations between datasets can be determined based on their derivation relationship. Capturing how one dataset is derived from another will provide rich linkages. We expect to mine and apply such lineage and provenance information to define additional relationships between datasets. iRODS and DVN provide access to descriptive and systemic metadata that can be useful for relevance analysis.

Audit trails that contain user usage behavior are another source of relevance information. iRODS and DVN have user bases that are associated with their data repositories. Some of the repositories keep audit trails for user access of datasets as well as some of the operations that users perform on those datasets. These audit trails can provide rich sociometric information about what types of data are of interest to a user and for matching users with similar interests. A problem of interest is how to extract sociometric information from audit trails without compromising privacy. We plan to research and create algorithms for audit-trail anonymization and aggregation techniques for extracting sociometric data. PINQ is a good potential framework to query the audit trail with privacy protection. PINQ is a LINQ-like API for computing privacy-sensitive datasets while providing guarantees of differential privacy for the underlying records [35, 36]. Our research will investigate what types of audit trails are necessary and sufficient to extract sociometric information to improve discovery and sharing without compromising privacy.

What methods and services are used (or applicable) to a dataset provides another kind of relationship between datasets. If the same or similar method is applied to two different datasets, there may be a relationship between the two. Some types of method relationships are likely to indicate a closer relationship than others; for example, calculating a mean or variance relationship on each of two datasets likely implies less about the relationship between the two than using each of them as an input to a computational fluid dynamics model.

Finally, publications are another concrete source of relevance information about datasets that have been used for successful research. Increasingly, publishers require proper citation and public release of data sources. Although there is much room for improvement, many publications now have a proper citation for the data used in the research. The DVN supports a standard persistent data citation method proposed by [37, 38, 39], and the DVN team is involved in several efforts to help standardize it into other domains such as DataOne [40]. The use of the data in publications is ultimate evidence that these data are relevant and useful for certain users and types of research. One of the biggest challenges in machine learning algorithms is to find “gold standard” data to train models and evaluate the results. Data citations can be used to build gold standards for sociometric measurements that connect data. In addition to verifying the sociometric network built by the DataBridge, these data will be used to identify sets of data from past collaborations and to learn patterns of collaborations.

3 RELEVANCE ALGORITHMS

Relevance metrics will be used to connect data in a network for discovering multi-dimensional similarities between datasets. We propose to research and implement relevance metrics across a broad range of types as described below.

3.1 DATA TO DATA CONNECTIONS

We expect to use techniques in Natural Language Processing (NLP) to measure semantic similarity of datasets to infer the topic space as latent classes. Methods such as

Probabilistic Latent Semantic Analysis (PLSA) will be used to identify both synonyms (words that refer to the same topic) and polysemy (words with multiple meanings) in a corpus of abstracts by modeling each document as a mixture of unigram model topics. The topics are modeled as latent classes that regulate the probability distribution of words in a given document as follows where $\mathrm{p}(\mathrm{z}_\mathrm{k})$ is the probability of topic z_k:

$$\mathrm{p}(\mathrm{d}_\mathrm{i}, \mathrm{w}_\mathrm{j}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{p}(\mathrm{z}_\mathrm{k})\mathrm{p}(\mathrm{d}_\mathrm{i}|\mathrm{z}_\mathrm{k})\mathrm{P}(\mathrm{w}_\mathrm{j}|\mathrm{z}_\mathrm{k}).$$

Expectation Maximization algorithms are widely used to estimate the model $\mathrm{p}\left(\mathrm{z}_\mathrm{k}|\mathrm{d}_\mathrm{i}\,, \mathrm{w}_\mathrm{j}\right)$. Once the topic space is formed for a given set of documents, queries can be handled in the projected topic space by calculating the cosine distance, a method which is more accurate than using key words. However, the parameterization of the PLSA model is susceptible to over-fitting and there is no obvious way to infer about new documents not seen in the training data. Latent Dirichlet Allocation (LDA) addresses these limitations by using a more general Bayesian probabilistic topic model [41]. There are many extensions to the PLSA and LDA models that differ in how they vary this basic generation process and what statistical assumptions are made. All current probabilistic topic models are based on the fundamental idea that documents are mixtures of topics, where a topic is represented by a multinomial distribution of words (a unigram language model). The main limitations of these methods are that they are computationally expensive and can only handle a modest sized corpus.

In DataBridge, we will compare different probabilistic topic models to determine which work best to cluster a corpus of short abstracts, explore effective metrics for sampling the abstracts to be projected together using these models, and methods to effectively scale these models (i.e. layer the topic space). We will further compare the effectiveness of these methods to random projections, which can scale to handle a large corpus. We expect to use and contribute to the open source semantic vector package in applying random projections [42].

The ability to understand and link data across the many types of data and research disciplines hinges on the quality of metadata available to describe these data. Archives and researchers are tasked with creating these valuable metadata, but the current process is very labor intensive. Archives and repositories are faced with the decision to process many datasets with minimum description or process fewer datasets and take time to create a more detailed description. The DataBridge research plan includes a RESTful web service that allows automated metadata generation based upon information discovered during relevance engine processing. This new interface will automatically suggest topics and keywords to researchers uploading and ingesting data into repositories by pre-tagging the projected topic space with appropriate terms. Researchers will be presented with easy to select metadata options that help them describe their data relative to other data discovered by the system and matched to the new data via the investigation of variable level and data abstract discovery tools. We will measure the effectiveness of this pre-tagging by tracking how many users pick one of the suggested topics and explore the appropriate range in the cosine distance for effective suggestions. Accurate suggestions will encourage accurate crowd sourced tagging of the ingested data. Since quality assurance is one of the biggest challenges of crowd sourcing, adding a good NLP module for semantic processing of the abstract is expected to have a significant impact on improving the metadata. The addition of DataBridge to the ingest process will enhance automated metadata creation and feed more precise metadata back into the process. As DataBridge builds more connections, the intelligent metadata creation process has better data to work with and the relevance engine has the ability to provide more accurate results. This researcher guided/machine-learning metadata creation environment will have a profound impact on the future of data discovery and reuse [43].

3.2 USER TO DATA CONNECTIONS

Currently searchable metadata on the DVN includes authors, producers, distributors, provenance, and geographic and time coverage of a dataset. DataBridge will extend beyond passive searching, allowing us to better understand and even predict possible future collaborations. We will crawl published papers that use data stored in DVN to identify dataset relationships that have been manifested in previous collaborations and to discover and explore collaboration patterns based on features extracted from the given datasets. Using these patterns, we will build models of collaboration to predict data connections.

We also plan to gather datasets usage patterns to connect users to datasets. If a user U uses dataset A and dataset B at a later time, one can infer a weak linkage. Based on the (semantic) types of the two data one can infer a link between user and data types. This link can be reinforced if similar usage patterns are repeated with other datasets. Similar connections can also be drawn about users and the data they submit (or own) to the DataBridge. Patterns connecting users and datasets can also provide linkages between users who have similar patterns of data usage and ownership. A similar connection can be made comparing methods and processes (or workflows) that a user applies. We believe that there are several such types of patterns that can be gleaned through usage patterns and we will characterize such a class of linkages as part of the DataBridge project.

When discovering other relevant data, data quality becomes as important as the relevance of the data to the topic. The importance of quality can be illustrated with a basic search on the web. The underlying method used in search engines such as Google and PageRank is based on measuring quality of a webpage using links [33, 34]. In direct marketing, RFM (Recency, Frequency, and Monetary value) is the accepted norm for predicting potential future customers [44]. Recent activity is the first and most important factor for predicting customers likely to make another purchase. Following these ideas, DataBridge will use metadata on download frequencies and times of download as a measure of the quality of the data. Different research topics become popular at different times. DataBridge will also measure recency as defined by when data download activity occurs. We will also explore metrics to evaluate data reliability based on authors, producers, distributors, and data provenance.

Once we have built a comprehensive network of data, we will use it to connect users. Currently, iRODS has capabilities to monitor usage patterns, but the DVN is open access with no monitoring of user behavior beyond IP address tracking. Login-based user tracking will be important for connecting researchers with each other in the DataBridge system, so as part of this project, we will incorporate Shibboleth [45] into DVN for tracking users. With Shibboleth implemented, we will build user profiles to help connect users to users and users to data. Data used for building user profiles will include a researcher's download history, search history, published papers, and authorship of data uploads.

Collaborative Filtering (CF) is a technique commonly used by recommender systems for making automatic predictions (filtering) about the interests of a user by collecting preference information from many users (collaborating) [46]. In DataBridge, user interest is expressed in terms of interest in particular datasets as Y/N. Thus, we can build a user data matrix, then use the cosine distance to explore the data to data connection in the matrix. The main challenges in using CF are data sparsity in the beginning, which leads to cold start, and problems with scalability. The underlying assumption is that two researchers who have similar interests on one domain are likely to have similar interests in other domains. However, there are also many scientists who might not follow this data use pattern. For example, there could be a scientist who is the only computer scientist interested in social welfare. Indeed, how similar are scientists' data use patterns is one of the research questions we plan to investigate.

3.3 METHODS TO DATA CONNECTIONS

The use of particular models and methods to analyze datasets provides rich data for mining sociometric information. Metadata concerning the usage of datasets can be considered a measure of the similarity between the datasets. Usage methods and applications can be viewed as properties of the datasets and can be used to determine relevance between datasets. To enable this we will need to define an ontology of methods, tools, and applications. The ontology will have a hierarchical structure (possibly

© ASE 2013

shallow) and will map methodologies into a tree structure. We plan to define such an ontology using actual programs and software as elements of that ontology. In our research, we will implement relevance algorithms that will use the method ontology to help define a relevance network.

3.4 OTHER CONNECTIONS

As part of our analysis in finding linkages between data, we will derive other significant connections that will provide insight into additional sociometric relationships. The first such type of connection will be between people based solely on their interactions through common datasets. Examples of relationships that will emerge include producer-consumer (of data or methods) and same or similar data/method user. A second type of connection that we can establish through data will be between methods. Two methods applied to the same dataset will be considered to be linked semantically – either in a producer-consumer fashion, or defining similar input-output relationships or defining similar functionalities. Finally, one can also associate methods and users and derive relationships between them. Using such a relationship, the system can suggest alternative methods to a user based on similar user usage models.

4 IMPLEMENTATION

As part of our first year activities, we have begun building a prototype of the DataBridge system. For this prototype, we are focusing on the over 50,000 datasets available on the DVN hosted at the Odum Institute [47]. In the initial prototype, the DataBridge utilizes the rich metadata that is already prepopulated in this DVN. Starting with a particular DVN instance, like the one at Odum Institute, a metadata gatherer collects all metadata for all datasets in the DVN using a web crawler that can scrape the information rooted at the main public DVN website. For evaluation purposes, we have implemented a metadata and ontology database to store this information using two different graph database systems, Neo4j [48] and titan+hbase [49]. This database, however implemented, represents the unprocessed nodes of the network that are the basis of the DataBridge knowledge base. The relevance engine processes this information to calculate the similarity of the nodes using various relevance algorithms discussed in this paper. Currently we have a basic relevance engine which uses an overlap similarity algorithm to produce similarity matrices from the metadata. These similarity matrices represent the edges in the network. It is important to recognize that there can be multiple dimensions in the network which are represented by multiple edges connecting two nodes. For example, one edge might represent the Jaccard index, the number of common elements over all elements, between keywords for two datasets, while another edge might represent a similarity measure relating the original publishers of the data sets. In this manner, the DataBridge will build and store a multi-dimensional relevance network for all datasets known by the DataBridge system. Finally, this multi-dimensional relevance network information is made available to scientists via a visualization and query component currently implemented in JavaScript using the d3.js [50] library. The different components of the system are loosely coupled via a message oriented communication backbone based on the publish-notify-subscribe paradigm. The current implementation uses RabbitMQ [51], an implementation of the Advanced Message Queuing Protocol.

IV RELATED WORK

1 DATA PUBLICATION & DISTRIBUTION

Large-scale projects provide a venue for data publication and distribution. National projects such as DataONE (Dataone) [40], Datanet Federation Consortium (DFC) [52], the Consortium of Universities for the Advancement of Hydrologic Science (CUHASI) [53], iPlant Consortium (IPC) [54], and the Ocean Observatories Initiative (OOI) [55] are large consortium-based projects that collect and provide access to disciplinary data collections. What is lacking, however, is a network that connects datasets such that the whole becomes greater than the sum of its parts – connections based on similarities beyond normal textual connections within the silos of single disciplines. Scientific collaboration tools have been developed that help scientists build collections for their projects. These include middleware systems such as iRODS [10] that allow collaboration across data grids by abstracting the storage architecture and enforcing policies on the data. Additionally, communities built around repository software like DuraSpace [56], Islandora [57], or Fedora Commons [58] allow researchers and curators to build digital collections important to communities of interest for sharing and preservation. The Dataverse Network [7] is a collaboration tool that gives archivists, researchers, and publishers an open source application to publish, share, reference, extract and analyze research data. Missing from these works is an explicit focus on the relationships among datasets. DataBridge seeks to provide relationships between datasets and provide these links back to scientific collaboration environments.

2 DATA DISCOVERY

Other researchers have begun to explore how to link datasets. Tools such as SciVal, CiteSeerX and DataCite, for example, are designed around metadata schemas. CiteSeerX includes some additional techniques such as automatic metadata harvesting from indexed articles and crowd sourcing of opinions about articles, but still does not address datasets. DataCite is designed to make it easier for researchers to find relevant datasets by collecting metadata, providing a search capability, and assigning persistent data identifiers to assist in citing and publication, but it lacks any of the sociometric infrastructure we intend to build in the DataBridge. In relation to scholarly communication, the role of automatic metadata generation is being researched and tested as a method to help increase discoverability, access, and efficiency [59, 60, 61]. In today’s digital information age, many now consider datasets unique units of scholarly communication in their own right [62, 63]. SEAD [64] is also concerned with sustainability of datasets in the long tail of science and proposes to provide a data repository for scientists to manage, share, and link their data with others. However, the linking is done manually by users and not through automatic analysis as proposed here. The Linked Data project (linkeddata.org) links data through the use of a data description language, but the description is not generated by analysis as we propose. Google Scholar allows the user to search a wide variety of sources, including books and some web sites, but has neither a sociometric component nor a focus on datasets. Google does focus on the sociometry of datasets, but these datasets are limited to what are presented on HTTP servers that can be crawled and are normally unstructured text or images. Another related project is the CASRAI [65] program, which is focused on developing common metadata standards to allow for linking research information to facilitate data exchange, collaboration, and interaction with funding agencies.

3 COMPUTER-ASSISTED CLUSTERING

Computer-assisted methods to discover clusters (or partitions) in large corpora of unstructured text [66] differs from the clustering methods discussed in section II as they allow users to interact with the clustering space until they find a result useful and tuned to their needs. The computer-assisted method encompasses five main steps: 1) calculate the word count matrix of an entire document set, 2) apply more than one hundred clustering methods to the numerical representation calculated in the first step and obtain the respective clustering solutions, 3) calculate a similarity distance between clusters based on the number of pairs of documents not placed together from one cluster to another, 4) project this matrix of distances across all clusters into a two-dimensional Euclidean space, 5) calculate new clustering solutions at a given point in the space of clusters from a weighted average of the nearby pre-calculated clusters. Separately, several of us (King and Crosas) are developing a clustering discovery and exploration tool (Consilience.com) that uses this method to interactively explore and select insightful clustering solutions.

Both types of methods - fully automated and computer-assisted - can be applied to categorize collections of datasets. The fully automated methods generate general solutions for all users, while the computer-assisted method offers a thorough exploration of the clustering space to obtain the most optimal

© ASE 2013

solution for each user. In the case of the computer-assisted method described here, the text input will be all the metadata associated with each dataset in the collection.

V CONCLUSION

This paper outlines the design philosophy behind DataBridge – an e-Science collaboration system that enables researchers to discover new modalities of linkages between datasets. The main thrust of the project is to apply SNA and other relevance algorithms to define different types of semantic bridges that link large quantities of diverse datasets in the long tail of science.

ACKNOWLEDGMENT

The efforts described in this paper are funded by the NSF Office of Cyberinfrastructure OCI-1247652, OCI-1247602, OCI-1247663 grant, “BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: DataBridge - A Sociometric System for Long Tail Science Data Collections”, (2012-2015).

REFERENCES

[1] C. Palmer, M. Cragin, P. Heidorn, and L. Smith, “Data curation for the long tail of science: The case of environmental sciences.” Paper presented at the Third International Digital Curation Conference,Washington, DC., Dec. 2007.

[2] C. Palmer and C. Faloutsos, “Electricity based external similarity of categorical attributes,” in: Proceedings of PAKDD, 2003, pp. 486-500.

[3] A. Rajasekar, S. Sankaran, H. Lander, T. Carsey, J. Crabtree, H. Kum, M. Crosas, G. King, and J. Zhan. “Sociometric methods for relevancy analysis of Long Tail Science Data,” 2013 ASE/IEEE International Conference on Big Data.

[4] http://databridge.web.unc.edu/

[5] M. Crosas, “The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data,” D-Lib Magazine, Volume 17 Number 1/2, 2007.

[6] M. Crosas, “A Data Sharing Story,” 2012, J eScience Librarianship 1(3): Article 7, 2012.

[7] DVN, The Dataverse Network, http://thedata.org

[8] G. King, “An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, Sociological Methods & Research,” vol 36(2), 2007, pp. 173-199.

[9] M. Altman, and G. King, “A proposed standard for the Scholarly Citation of Quantitative Data,” DLib Magazine 13 (3/4). 1082–9873. 2007.

[10] iRODS, integrated Rule Oriented Data System, http://www.irods.org

[11] R. Moore, and A. Rajasekar, “Rule-Based Distributed Data Management Grid,” 2007 IEEE/ACM International Conference on Grid Computing.

[12] R. Moore, A. Rajasekar, A., and A. de Torcy, A.. “Policy-based Digital Library Management,” International Conference on Digital Libraries, Delhi, India, Feb. 2009.

[13] A. Rajasekar, M. Wan, R. Moore, and W. Schroeder, “A Prototype Rule-based Distributed Data Management System,” HPDC workshop on Next Generation Distributed Data Management, Paris, France, 2006.

[14] A. Rajasekar, R. Moore, R., M. Wan, W. Schroeder, and A. Hasan, “Applying Rules As Policies for Large-Scale Data Sharing,” 1st International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, UK., Jan. 2010.

[15] M. Wan, R. Moore, and A. Rajasekar, “Integration of Cloud Storage with Data Grids,” The Third International Conference on the Virtual Computing Initiative, Research Triangle Park, NC, Oct. 2009.

[16] J. Moreno, Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation, Beacon House, Beacon, New York, 1951.

[17] S. White and P. Smyth, “Algorithms for estimating relative importance in networks,” in: KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2003, pp. 266-275.

[18] F. Lorrain and H. White, “Structural equivalence of individuals in social networks,” J. Math. Sociol. Vol. 1 1971, pp. 49-80.

[19] S. Wasserman and K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, UK, 1994.

[20] R. Burt, “Positions in networks,” Soc. Forces vol 55, 1976 pp. 93-122.

[21] P. Elias, A. Feinstein, and C.E. Shannon, “Note on maximum flow through a network,” IRE Trans. Inf. Theory IT- 2, 1956, pp. 117-119.

© ASE 2013

[22] E. Estrada and N. Hatano, "Communicability in complex networks," Phys. Rev. E vol. 77 (3), 2008, pp. 036-111.

[23] E. Estrada and N. Hatano, "Communicability graph and community structures in complex networks," Appl. Math. Comput. Vol. 214, 2009, pp. 500-511.

[24] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," Knowledge and IEEE Transactions on Data Engineering, , vol.19, no.3, Mar. 2007, pp.355-369.

[25] M. Saerens, F. Fouss, L. Yen, and P. Dupont, "The principal component analysis of a graph and its relationships to spectral clustering," in: Proc. Eur. Conf. on Machine Learning, 2004, citeseer.ist.psu.edu/saerens04principal.html.

[26] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering with the sigmoid commute-time kernel: A comparative study," Data Knowl. Eng. Vol. 68 (3), 2009, pp. 338-361.

[27] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering based on the commute-time kernel," in: PAKDD, pp. 1037-1045.

[28] A. Chandra, P. Raghavan, W.L. Ruzzo, and R. Smolensky, "The electrical resistance of a graph captures its commute and cover times," in: STOC '89: Proceedings of the twenty-first annual ACM symposium on Theory of computing, ACM, New York, NY, USA, 1989, pp. 574-586.

[29] D. Klein, and M. Randic. (1983). Resistance distance, J. Math. Chem. 12 pp. 81-95.

[30] D. Harel and Y. Kore. (2001). On clustering using random walks, in: FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science, Springer-Verlag, London, UK, pp. 18-41.

[31] H. Tong, C. Faloutsos, and J. Pan. (2008). Random walk with restart: Fast solutions and applications, Knowl. Inf. Syst. 14 (3) pp. 327-346.

[32] M. Gori, and A. Pucci. (2007). Itemrank: A random-walk based scoring algorithm for] recommender engines, in: IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 2766-2771.

[33] S. Brin and L. Page. (1998). The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN 30 pp. 107-117.

[34] L. Page, S. Brin, R. Motwani and T. Winograd. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[35] R. Gross, and A. Acquisti, "Information revelation and privacy in online social networks," In Pre-proceedings version. ACM Workshop on Privacy in the Electronic Society (WPES), 2005.

[36] F. McSherry. (2009, June). Privacy Integrated Queries, in Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD), ACM, Inc.

[37] M. Altman. (2007). A Fingerprint Method for Ve] rification of Scientific Data, in Advances in Systems, Computing Sciences and Software Engineering (Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering 2007), Springer Verlag.

[38] M. Altman, et al (2009). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. The American Archivist. 72(1): pp. 169-182

[39] M. Altman, et al (2001). A Digital Library for the Dissemination and Replication of Quantitative Social Science Research, Social Science Computer Review 19(4): pp. 458-71.

[40] Dataone, DataONE, http://www.dataone.org

[41] D. Blei, A. Ng, and M. Jordan. (2003). Latent dirichlet allocation, Journal of Machine Learning Research, 3, pp. 993–1022.

[42] D. Widdows and F. Ferraro. (2008). Semantic vectors: a scalable open source package and online technology management application. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp. 1183-1190.

[43] T. Hofmann. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning Journal, 42(1), pp.177-196.

[44] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen, and G. Dedene. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing, European Journal of Operational Research, Volume 138, Issue 1, pp. 191-211.

[45] B. Morgan, S. Cantor, S. Carmody, W. Hoehn and K. Klingenstein. (2004). Federated Security: the Shibboleth Approach. Educause Quarterly, 4(4), pp. 12-17.

[46] S. Xiaoyuan and T. Khoshgoftaar. (2009). A survey of collaborative filtering techniques, Advances in Artificial Intelligence archive.

[47] Odum Dataverse, http://arc.irss.unc.edu/dvn/

[48] M. Hunger and P. Rathle. (2012). NoSQL, Big Data and Graphs. http://info.neotechnology.com/rs/neotechnology/images/NBDAG WP.pdf

[49] M. Broecheler, and MA. Rodriquez. (2012). Titan: The Rise of Big Graph Data. Public Lecture at Jive Software, Palo Alto

[50] M. Bostock, V. Ogievetsky, and J. Heer "D3: Data-Driven Documents", IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011.

[51] RabbitMQ, http://www.rabbitmq.com

[52] DFC, Datanet Federation Consortium, http://www.feddata.org

[53] CUHASI, Consortium of Universities for the Advancement of Hydrologic Science, http://www.cuahsi.org/

[54] IPC, The iPlant Collaborative, http://www.iplantcollaborative.org/

[55] Ocean Observatories Initiative (OOI) http://www.oceanobservatories.org/

[56] DuraSpace, The DuraSpace Project, http://www.duraspace.org

[57] Islandora, The Islandora Project, http://islandora.ca

[58] Fedora Commons, The Fedora Commons Project, www.fedora-commons.org

[59] K. Calhoun. (2006). The Changing Nature of the Catalog and its Integration with Other Discovery Tools. Library of Congress. Retrieved from http://www.loc.gov/catdir/calhoun-report-final.pdf.

[60] J. Greenberg. (2004). Metadata Extraction and Harvesting. Journal of Internet Cataloging, 6(4), pp. 59–82. doi:10.1300/J141v06n04_05.

[61] J. Greenberg, H. White, C. Carrier, and R. Scherle. (2009). A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata, 9(3-4), pp. 194–212. doi:10.1080/19386380903405090.

[62] H. Davis and J. Vickery. (2007). Datasets, a Shift in the Currency of Scholarly Communication: Implications for Library Collections and Acquisitions. Serials Review, 33(1), pp. 26–32. doi:10.1016/j.serrev.2006.11.004

[63] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner. (2004). Rethinking scholarly communication: Building the System that Scholars Deserve. D-Lib Magazine, 10(9). doi:10.1045/september2004-vandesompel

[64] SEAD, http://sead-data.net

[65] CASRAI, http://casrai.org/

[66] J. Grimmer and G. King. General Purpose Computer-Assisted Clustering and Conceptualization. 2011. PNAS

© ASE 2013