Data Quality Inference

Raymond K. Pon and Alfonso F. Cárdenas
UCLA Computer Science
Boelter Hall 4829
Los Angeles, CA 90095
(310) 825-1770
{rpon, cardenas}@cs.ucla.edu

ABSTRACT

In the field of sensor networks, data integration and collaboration, and intelligence gathering efforts, information on the quality of data sources are important but are often not available. We describe a technique to rank data sources by observing and comparing their behavior (i.e., the data produced by data sources) to rank. Intuitively, our measure characterizes data sources that agree with accurate or high-quality data sources as likely accurate. Furthermore, our measure includes a temporal component that takes into account a data source's past accuracy in evaluating its current accuracy. Initial experimental results based on simulation data to support our hypothesis demonstrate high precision and recall on identifying the most accurate data sources.

1. INTRODUCTION

A major aspect of data provenance is the ability to track the quality of data as it is processed by various transformations, each with an associated computational or intrinsic data collection error. It is important to choose trustworthy data sources when querying over multiple data sources [1]. Users of data warehouses regard the quality of information as important and as a factor in measuring the utility of a data warehouse [2]. Furthermore, the inclusion of data quality information has an impact on decision-making and decision-support systems as well [3, 4]. Also data conflicts, occurring when heterogeneous data sources are integrated, can be resolved by considering the quality of the data sources involved [5]. The accuracy of data can also be used to rank query results as well (as opposed to the relevance of query results to the query) [6]. Clearly, this trustworthiness and quality information should be stored as part of a data item's provenance. The following is a general query in which data quality can be used to answer:

Query 1: *"Given many genomic databases where data has been collected by various means and institutions, find a DNA sequence that satisfies a condition C."* In this query, collections of data sets (i.e., DNA sequences) have been collected by different instruments and/or possibly derived by various and possibly multiple transformations. Each of these instruments and transformations has a different degree of reliability and error. Additionally, the data sets that are relevant to the query may be numerous, so it is necessary to rank data sets by their "quality" or trustworthiness (i.e., how reliable the data sets are).

IQIS 2005, June 17, 2005, Baltimore, MD, USA.

However, it is unclear as to where metadata regarding data quality comes from. User-provided ratings of data sets or sources can be used to rate the quality of data sources [1], but is clearly a subjective measure and would require large samples to get any meaningful results. Error measures can be provided by data sources providers along with data sets, but may be inaccurate, difficult to use, incomplete, or untrustworthy [7]. Thus, we describe a technique to rank data sources by observing and comparing the behavior (i.e., the data produced by data sources) to rank them in terms of their quality. Intuitively, in our measure, data sources that agree with accurate data sources are likely to be accurate. Furthermore, in our measure, data sources that have been accurate in the past are also likely to be accurate in the future. We provide some initial experimental results based on simulation data to support our hypothesis. The following sub-sections discuss motivations for this work and the related works. In section 2, we describe our technique for data source ranking. In section 3 and 4, we present our initial experimental results and possible roads of future research, respectively.

1.1 Motivation

We describe three possible application areas in which the modeling of data accuracy and trustworthiness are important.

Application 1: Sensor networks are becoming increasingly prevalent in observing wildlife, monitoring environmental conditions, monitoring of soldiers in the field, and the detection of harmful biological and chemical agents, with practical applications in homeland security [8]. The effectiveness of these sensor networks is highly dependent on the accuracy of the networks, which is a function of the current battery level of the device, interference, and intrinsic error in data collection. For example, in the near future, soldiers may be equipped with data capturing devices, making each soldier a sensor [9], to give field commanders current battlefield status reports. Data captured by soldiers may be conflicting and/or erroneous because of the human element involved in the data capturing process. It would be advantageous to be able to determine the more trustworthy "sensors" in capturing the current situation to filter out noisy data and to reduce the consumption of resources (e.g., manpower, time, and battery-life).

Application 2: In biomedical research, research facilities frequently collaborate with each other, sharing experimental data and results. In particular, comparing genome sequences from different species has become an important tool for identifying functions of genes [10]. This necessitates dynamically integrating different databases or warehousing them into a single repository. Scientists need to know how reliable the data is if they are to base their research on it. Pursuing incorrect theories cost time and

money. The obvious solution to ensure data quality is curation, but data sources are autonomous and as a result sources may provide excellent reliability in one area, but not in all data provided, and curation slows the incorporation of data. Data providers will not directly support data quality evaluations to the same degree since there is no equal motivation for them to and there are no standards in place for evaluating and comparing data quality [7]. Thus, automatic, impartial, and independent data quality evaluation would be needed.

Application 3: In intelligence gathering efforts, data is often collected from many heterogeneous data sources, such as satellites, human assets, transcripts, wiretaps, etc. It is obvious that each of these data sources have different degrees of quality and trust. And with the multitude of data sources to incorporate, it is currently time-consuming to sift through each of these data sources to determine which the most accurate sources are. To make the correct decisions based on the intelligence available in a timely manner, we will need an automatic means to determine accurate data sources and to be able to detect malicious or compromised data sources to prevent them from influencing decision-making processes.

1.2 Related Works

There has been a significant amount of work in the area of information quality, ranging from techniques in assessing information quality and accuracy to building large-scale data integration systems over heterogeneous data sources. For example, the DaQuinCIS system [11-13] is a cooperative information system where data source providers are evaluated by data source users in a peer-to-peer system. Unfortunately, such a system relies heavily on the participation of users in the review of the quality of data in the system, which may not be practical in real-life environments. Users may not reliably or consistent in evaluating data sources.

Other works have taken other approaches in modeling and capturing data quality. Some have developed data models to model data quality but rely on data quality metadata being available, such as data sources publishing such information [14-19]. Unfortunately, these approaches rely on precise and accurate metadata. However, such metadata are not always available [7] and there is no single agreed upon standard in describing data lineage. Additionally, it may be possible for malicious processes to corrupt or "spam" query results by providing false metadata.

There has also been work in methodologies in assessing of the quality of data in databases [20-23]. However, such methodologies rely on human assessment of the data, which is often time-consuming and possibly error-prone.

Previous works have assumed that the metadata regarding the quality of data is available, accurate, and unbiased, either published by the data providers themselves or provided by user-rankings of the data sources. Our contribution is that we do not assume that such metadata is available and reliable. Rather, our automated approach examines how well the data sets produced by data sources agree with one another, and infer the rankings of the data sources in terms of their accuracy. We take an approach similar to Google's PageRank [24]. Instead of evaluating the popularity of web sites by measuring how many other popular websites link to them as in PageRank's approach, we evaluate the accuracy of data sources by measuring how well other data sources agree with the data they produce. This approach is automated, does not rely on possibly faulty and limited metadata, and does not require human assessment.

2. DATA SOURCE RANKING

Traditionally, the ranking of query results was based on the relevance of a user's query. However, the quality of the results could be improved if we incorporated a data quality measure in addition to their relevance to the user's query.

We wish to do following in general:

1. Rank the data sets or data sources in order of their accuracies.
2. Determine the top-k accurate data sets or data sources.

This ordering is important particularly in data integration systems, where there are numerous data sources available of varying accuracy that changes dynamically across time. Ideally, given a query, we would like to contact each data source; however, this may be prohibitively expensive if there are budget constraints such as time and network resources. Such applications can be found in sensor networks, where battery-life is limited, and intelligence-gathering efforts, where manpower and time are limited. Thus, it would be advantageous to determine the most accurate set of data sources, so that they can be contacted in answering a query. Additionally, this methodology could be used for identifying malicious or compromised data sources that are attempting to feed false information into the data integration system. We provide the following framework for ranking the accuracy or trustworthiness of data sources based on observing and comparing data source behavior without any a priori knowledge of their relative accuracies to help solve the problem. In our model, we assume that schema and data heterogeneity have been reconciled, which is beyond the scope of this work.

2.1 General Framework

Let D be a set of data sources. A data source $d_i \in D$ generates a table $T_i^t(k,v)$ for a query Q where t is the time index, k is the key column, and v is the value column of the table. We want to derive a metric $A_i^t \in [0,1]$ that measures the relative accuracy of data source d_i at time t such that $A_i^t < A_j^t$ if d_i is less accurate than d_j at time t. We define such a metric as the weighted average of the previous accuracy estimate at time index $t-1$ and the accuracy estimate derived by observing the data generated by data sources in D:

$$A_i^t = h(t)A_i^{t-1} + (1-h(t))c(i,t) \quad (1)$$

The intuition behind A_i^t is that a data source's accuracy should be a function of its past accuracy (i.e., reputation) and its current behavior. The function $h(t)$, where $0 \le h(t) \le 1$, is the historical weight function that determines the contribution of the accuracy estimate at the previous time index. The intuition behind the historical component of the accuracy measure is that a data source that has been accurate (or inaccurate) in the past should also be accurate (or inaccurate) in the near future. For simplicity, the above historical component assumes a Markovian behavior in the evolution of the data sources, where the accuracy at time t is only dependent on the value at t-1. However, it will be interesting to see if we can improve the quality of our estimation by taking into

account a sliding window of size w: $[A_i^t, A_i^{t-1}, \ldots, A_i^{t-w}]$. Thus, the historical component would consist of a weighted sum of all accuracy estimates within the last w time indexes, where each estimate is weighted with a decaying weight function. The decaying weight function would assign a higher weight to more recent estimates than older estimates. A sliding window version of the equation (1) would be of the following form, where *w(j, t – 1)* is the decaying weight function:

$$A_i^t = h(t) \sum_{j=t-w}^{t-1} w(j, t-1) A_i^j + (1 - h(t)) c(i,t) \qquad (2)$$

However, we leave this issue to future research, where we will study the most appropriate decaying weight function and the optimal sliding window size.

The cohesion function *c(i,t)* determines the new accuracy estimate by observing data generated by data sources in *D* at the current time index and how well each data source agrees with one another. The cohesion function *c(i,t)* that we propose is the following:

$$c(i,t) = f(i,t) + \frac{(1 - f(i,t))}{|D| - 1} \sum_{d_j \in D - \{d_i\}} a(i,j,t) c(j,t) \qquad (3)$$

The function *a(i,j,t)* is the agreement function, which outputs 0 when data sources d_i and d_j are in strong disagreement regarding the data in T_i^t and T_j^t, and outputs 1 when d_i and d_j strongly agree, and values between 0 and 1 for other levels of agreement. The intuition behind *c(i,t)* is:

- If a data source agrees with an accurate data source, it should also be accurate.
- If a data source agrees with an inaccurate data source, it should also be inaccurate.
- A data source has a probability *f(i,t)* of being absolutely accurate independent of any agreement/disagreement with the other data sources.

Thus, given a system of equations of $|D|$ equations and $|D|$ variables, it is possible to determine *c(i,t)* for all $d_i \in D$. The function *f(i,t)* is the dampening factor function (similar to that defined in Google's PageRank algorithm [24]). In addition to being the probability that a data source d_i is absolutely accurate independent of its agreement with the other data sources, the function *f(i,t)* will prevent the solution to the system of equations from consisting of entirely zeroes for all *c(i,t)*.

2.2 Agreement Functions

There are several possible definitions for *a(i, j, t)*, such as the *tupleOverlap* function, which measures the proportion of tuples in approximate agreement (within some allowable difference ε) in the set of tuples whose key values are generated by both data sources d_i and d_j:

$$tupleOverlap(i,j,t) = \frac{|T_i^t \underset{T_i^t.k = T_j^t.k}{\triangleright\triangleleft}_{T_i^t.v = T_j^t.v} T_j^t|}{|T_i^t \underset{T_i^t.k = T_j^t.k}{\triangleright\triangleleft} T_j^t|} \qquad (4)$$

Another possible definition for *a(i,j,t)* is the *cosineOverlap* function, which measures the complement of the cosine distance of two sets of data over the same key values generated by d_i and d_j:

$$cosineOverlap(i,j,t) = \frac{V(i,j,t)^T V(j,i,t)}{|V(i,j,t)||V(j,i,t)|} \qquad (5)$$

The vector *V(i,j,t)* can be roughly defined as $V(i,j,t) = \pi_{T_i^t.v}(T_i^t \underset{T_i^t.k = T_j^t.k}{\triangleright\triangleleft} T_j^t)$, except it is an ordered vector in which the values stored in the vector are ordered by their corresponding key values in T_i^t. There is also a Euclidian-based function for *a(i, j, t)*, which we will discuss in further detail later.

Given this system of $|D|$ equations and $|D|$ variables, we can arrange the equations to the following form:

$$A(t) * C(t) = F(t) \qquad (6)$$

A(t) is defined as the following matrix:

$$A(t) = \frac{f(t) - 1}{|D| - 1} \begin{bmatrix} \frac{|D| - 1}{f(t) - 1} & a(1,2,t) & \cdots & a(1,|D|,t) \\ a(2,1,t) & \frac{|D| - 1}{f(t) - 1} & \cdots & a(2,|D|,t) \\ \vdots & \vdots & \ddots & \vdots \\ a(|D|,1,t) & a(|D|,2,t) & \cdots & \frac{|D| - 1}{f(t) - 1} \end{bmatrix}$$

C(t) and *F(t)* are also defined as the following matrices:

$$C(t) = \begin{bmatrix} c(1,t) \\ c(2,t) \\ \vdots \\ c(|D|,t) \end{bmatrix}, \; F(t) = f(t) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The solution to equation (6), *C(t)*, is a vector where each entry $C(t)_i$ estimates the accuracy of data source d_i. The matrix *A(t)* can also be normalized with respect to maximum or sum of the entries in each of the rows (horizontal normalization) or in each of the columns (vertical normalization). We can horizontally normalize the matrix *A(t)* by performing the following division on every entry $A(t)_{i,j}$ in row *i*, column *j*, except for entries where *i* = *j*:

$$A'(t)_{i,j} = A(t)_{i,j} \Big/ Hor(t,i)$$

Hor(t,i) can either be the sum or the maximum value of all the entries in row *i* excluding the entry $A(t)_{i,i}$. We can also similarly define a function Ver(t,j) for vertical normalization ($A(t)_{i,j} = A(t)_{i,j} \Big/ Ver(t,j)$) to be either the sum or the maximum value of all entries in column *j* excluding the entry $A(t)_{j,j}$.

Given our normalization techniques, we can now discuss in further detail the Euclidian-based function for *a(i, j, t)* mentioned briefly before. Because Euclidian-distance is unbounded, normalization would be required to describe the amount of overlap or agreement. We define the Euclidian-based function *eOverlap* for *a(i, j, t)*:

$$eOverlap(i,j,t) = 1 - eDist'(V(i,j,t), V(j,i,t)) \qquad (7)$$

The function *eDist'* is simply the Euclidian distance of the vectors *V(i,j,t)* and *V(j, i, t)*, normalized in a similar manner as described above.

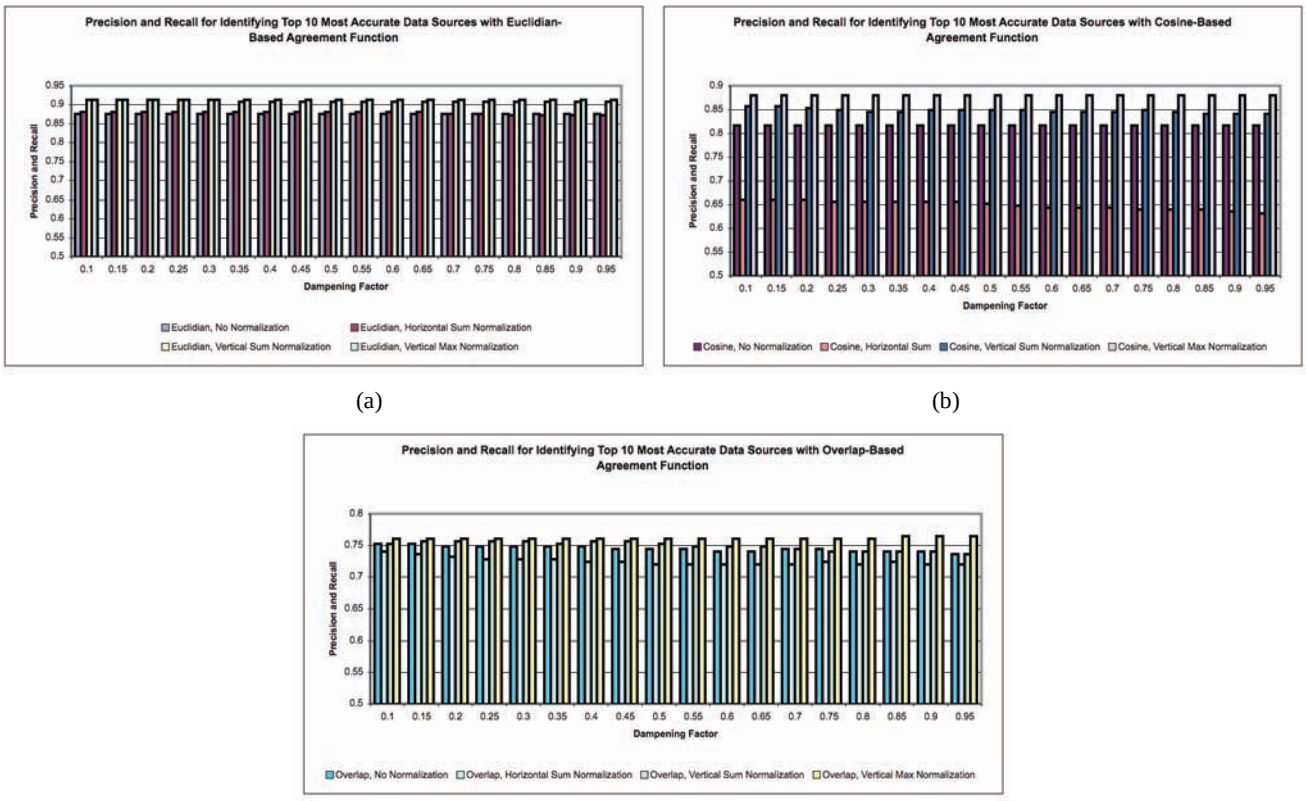

Figure 1: The precision and recall for identifying the top 10 most accurate data sources with (a) the Euclidian-based agreement functions, (b) the Cosine-based agreement functions, and (c) the Overlap-based agreement functions with = 0.1.

3. EXPERIMENTAL RESULTS

We hypothesize that the above framework can be used as a springboard in solving the general problem of identifying accurate data sources. To do so, we will need to identify adequate *h(t)*, *a(i,j,t)*, and *f(i,t)* functions through experimentation. Our initial experiments examine the cohesion function *c(i,t)* with a dampening factor *f* independent of time (i.e., the probability of a data source being absolutely accurate independent of all other data sources is constant), and excluding incorporation of the historical component. As a result, the combination of equations (1) and (3) reduces to the following:

$$A_i^t = c(i,t) = f + \frac{1 - f}{n - 1} \sum_{d_j \in D - \{d_i\}} a(i,j,t)c(j,t) \qquad (8)$$

We implemented a Java prototype, using JAMA (Java Matrix Package) [25] for solving the system of equations, and experimented on simulation data consisting of 100 data sources, each producing 20 different tuples, each consisting of a key (of type integer) and a value (of type double). In each run, a data set, consisting of 20 keys and values randomly assigned to each key with a uniform distribution, represent the "actual" data that each data source will attempt to report. Additionally, in each run, each data source was randomly assigned positive error values according to a Gaussian distribution with an average of 0 and a standard deviation of 1.0. For each run, we ran five iterations, where each data source produced a data set, consisting of values for each key, where each value is randomly generated with a Gaussian distribution with a standard deviation equal to that of the data source's error value and an average equal to that of the "actual" data item's value, essentially perturbing each data item's value with data source's error value. Data sources with large error values will generally generate values farther away from the "actual" value than data sources with smaller error values. We ran a total of five runs, consisting of five iterations, and averaged the results.

Figure 1 shows the precision and recall of the various agreement functions and normalizations as the dampening factor *f* is varied. Note that the overlap-based functions are using a difference margin = 0.1. The figure clearly shows that the dampening factor has very little effect in identifying the top 10 most accurate data sources. However, the figure does show that the vertical normalization with respect to the maximal value of the column yields the best performance. Additionally, the figure shows that the overlap-based functions perform the worse, with the cosine-based functions performing well and the Euclidian-based functions performing even better with a precision and recall of over 90%. The overlap-based functions suffer from having a fixed allowable difference margin that is difficult to estimate without knowing the nature of the data and the data sources a priori. The cosine-based functions perform better than the overlap-based functions because no such assumption is needed but does not accurately capture the amount of distance/overlap as Euclidian-based functions do.

To summarize, Figure 2 shows the performance of the three agreement functions with various normalizations using a fixed

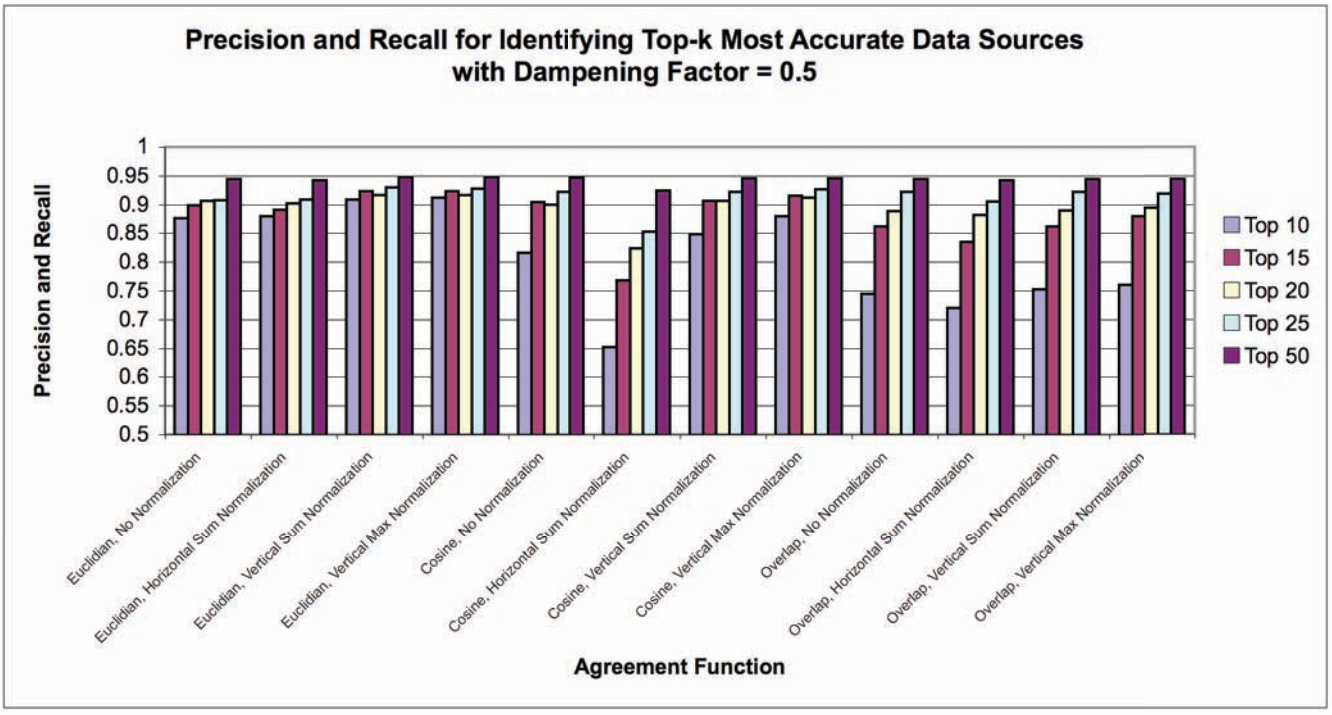

Figure 2: The precision and recall for identifying the top 10, 15, 20, 25, and 50 most accurate data sources with a dampening factor of 0.5

dampening factor of 0.5 (since current results do not definitively indicate the best value for *f*, we selected a mid-range value for *f*). The figure clearly shows that the Euclidian-based agreement function with vertical max and sum normalizations performs the best with a precision and recall of over 90%.

4. FUTURE WORK

One of the caveats of the current technique is that it relies on data sources reporting on the same set of data items. Often, it may be the case where data sources will report about different data items. Future study will have to be done to evaluate the current technique's effectiveness over incomplete and heterogeneous data sources. Additionally, the current technique may suffer from possibly expensive polling of all data sources. In future work, we will need to devise an efficient and intelligent sampling technique to alleviate such a problem while still preventing the staleness of estimates. One obvious possibility is to use the data gathered during a query (which is essentially free from the point of view of the quality estimator since such a cost will need to be incurred anyway to answer the query) to estimate a new relative accuracy measure than can be used for the next query. However, only data sources with high accuracy estimates will have their estimates updated and the estimates of data sources of low accuracy will become stale, since accurate data sources are the only data sources consistently being probed since they are selected to the answer the query. Thus, we will need to explore additional sampling techniques [26], such as polling for only small subsets of data from a majority of data sources, to solve this problem and to be able to associate a confidence metric in the ranking generated by our methodology.

Additionally, computing the solution to a set of *n* $c(i,t)$ equations with *n* variables may be computationally expensive if *n* is very large. Thus, we will also explore techniques to speed up this computation with an acceptable margin of error, such as using an iterative approach, using old $c(j,t-1)$ values for computing the new $c(i,t)$ value in equation (3). Figure 3 shows promising preliminary results regarding the performance of the iterative solution, indicating that we can arrive to a reasonably good estimation in very few iterations and that the dampening factor has some effect on how fast we can arrive to a solution. We use an initial estimate of $c(i,-1) = 1$ for all data sources and use the Euclidian-based agreement function with vertical sum normalization while varying the dampening factor. Future work will further explore the effect of the dampening factor.

In this preliminary study, we randomly assign error values to the

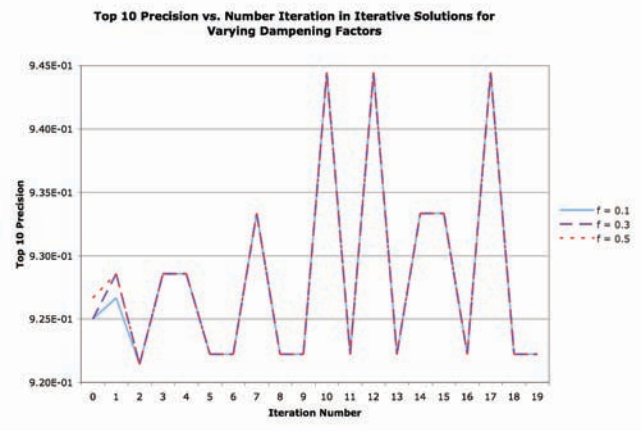

Figure 3: Performance of attaining an iterative solution using $c(i,-1) = 1$ and the Euclidian-based agreement function with vertical sum normalization.

data sources with a Gaussian distribution. Additional research will include further study on how well our cohesive function performs with other probability distributions, such as uniform distributions. We also hypothesize that such a technique can be used to automatically identify faulty or failing data sources dynamically, such as a sensor or an intelligence asset. We will need to experiment with the historical component of our accuracy measure. We will study how robust and reactive our accuracy measure will be when the accuracy of data sources becomes dynamic, as opposed to being static as in the case of this preliminary study.

Although we have experimented with an overlap-based function using a difference margin $\varepsilon = 0.1$ and could have used other values for ε to see the effect on the precision and recall of identifying the top-k most accurate data sources, the results indicate that that the overlap-based function performs poorly compared to the Euclidian and cosine-based functions with this value for ε. Another value for ε would have probably been better, but we hypothesize that the optimal ε is dependent upon the domain application of the data. In later work, we will examine the effect of ε when real-life data (e.g., sensor data) becomes readily available.

Currently, our accuracy measure evaluates the accuracy of data sources based a single domain of data (i.e., a single topic). However, data sources may provide data for multiple domains (i.e., multiple topics) and may be more accurate in one domain than another. There are two possible attitudes in approaching this problem. A "suspicious" attitude would suspect all data (regardless of topic) provided by a data source if a data source contradicts a more trustworthy data source. A "trusting" attitude would only suspect a minimal set of data (i.e., data from the contradicting topic) that contradicts a more trustworthy data source, which is a similar attitude taken in [27]. Future research will examine how these attitudes can be incorporated into the overall accuracy measure.

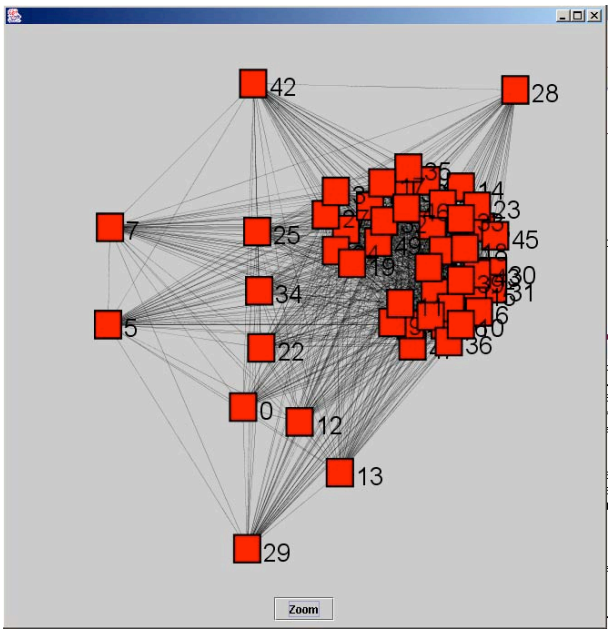

Figure 4: Network of agreeing data sources

We also envision that this technique can be used to identify communities of data sources in which members of the community share common "beliefs." In Figure 4, a graph generated with JUNG (Java Universal Network/Graph Framework) [28] consisting of 50 nodes, each representing a data source, are connected by edges, whose lengths are the Euclidian-distance of the data sets generated by the connecting nodes. It is clear from the graph that nodes that are in high agreement with one another are clustered very closely with each other; whereas, outliers in the graph disagree with the cluster and can be considered as inaccurate. Future work will include studies how clustering techniques can be used to identify communities of data sources, such as that from social network analysis [29].

5. CONCLUSION

We have presented an automated technique for inferring the quality of data sources without the luxury of metadata. Our main contribution is a framework to capture the historical accuracy of data sources and the relationship of data sources in how well they agree with one another (i.e., the cohesive function). Our second contribution is a preliminary study of the cohesive function, examining the precision and recall of identifying the top-k most accurate data sources with various agreement functions and normalizations. We have shown that the Euclidian-based agreement function vertically normalized performs the best.

We have also identified several significant challenges and future roads of research, including performance optimizations, exploring various sampling techniques, developing robust yet reactive accuracy estimations, and identifying communities of data sources.

6. ACKNOWLEDGMENT

The authors would like to thank Roderick Son, from the UCLA Medical Imaging Informatics Group, Terence Critchlow and David Buttler from the Lawrence Livermore National Laboratory, and the anonymous reviewers for their invaluable inspiration and input for this work. This work is partially funded by the National Foundation Grant # IIS 0140384.

7. REFERENCE

[1] D. Buttler, M. Coleman, T. Critchlow, R. Fileto, W. Han, C. Pu, D. Rocco, and L. Xiong, "Querying multiple bioinformatics information sources: can semantic web research help?" *SIGMOD Record*, vol. 31, pp. 59-64, 2002.

[2] A. Rudra and E. Yeo, "Issues in user perceptions of data quality and satisfaction in using a data warehouse-an Australian experience," presented at 33rd Annual Hawaii International Conference on System Sciences, 2000.

[3] I. N. Chengular-Smith, D. P. Ballou, and H. L. Pazer, "The impact of data quality information on decision making: an exploratory analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, pp. 853-864, 1999.

[4] R. A. Dillard, "Using data quality measures in decision-making algorithms," *IEEE Expert*, vol. 7, pp. 63-72, 1992.

[5] F. Naumann, "From databases to information systems - information quality makes the difference," presented at the International Conference on Information Quality (IQ 2001), Cambridge, MA, 2001.

[6] M. Gertz, M. T. Ozsu, G. Saake, and K. U. Sattler, "Report on the Dagstuhl seminar: 'data quality on the web'," *SIGMOD Record*, vol. 33, pp. 127-132, 2004.

[7] T. Critchlow, L. Liu, D. Buttler, D. Rocco, and C. Pu, "Towards Automatic Discovery and Identification of Bioinformatics Web Interfaces," [Online] Available: http://sirius.cs.ucdavis.edu/Dagstuhl03/presentations/03362.CritchlowTerence.Slides.ppt, 2003.
[8] V. Kumar (editor), "Special Issue on Sensor Network Technology and Sensor Data Management," *SIGMOD Record*, vol. 32, 2003.
[9] F. Donovan, "Army to deploy hand-held devices to make every soldier into a sensor," [Online] Available: http://www.aviationnow.com/avnow/news/channel_netdefense_story.jsp?id=news/arm04294.xml, 2004.
[10] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature*, vol. 422, pp. 835-847, 2003.
[11] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini, "Managing data quality in cooperative information systems," in *Lecture Notes in Computer Science 2519*, 2002, pp. 486-502.
[12] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni, "The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems," *Information Systems*, vol. 29, pp. 551-582, 2004.
[13] L. D. Santis, M. Scannapieco, and T. Catarci, "Trusting data quality in cooperative information systems," presented at CoopIS 2003, 2003.
[14] J. Widom, "Trio: a system for integrated management of data, accuracy, and lineage," presented at CIDR 2005, Pacific Grove, California, 2005.
[15] G. A. Mihaila, L. Raschid, and M.-E. Vidal, "Using quality of data metadata for source selection and ranking," presented at Third International Workshop on the Web and Databases, WebDB'2000, Dallax, TX, 2000.
[16] G. A. Mihaila, L. Raschid, and M.-E. Vidal, "Source selection and ranking in the websemantics architecture using quality of data metadata," *Advances in Computers*, vol. 55, pp. 87-118, 2002.
[17] M. Gertz, "Managing data quality and integrity in federated databases," presented at IFIP TC11 Working Group 11.5, Second Working Conference on Integrity and Internal Control in Information Systems: Bridging Business Requirements and Research Results, 1998.
[18] F. Naumann, J. C. Freytag, and U. Leser, "Completeness of integrated information sources," *Information Systems*, vol. 29, pp. 583-615, 2004.
[19] F. Naumann, "Quality-Driven Query Answering for Integrated Information Systems," in *Lecture Notes in Computer Science*, G. Goos, J. Hartmanis, and J. v. Leeuwen, Eds. Berlin, Germany: Springer-Verlag, 2002, pp. 166.
[20] A. Motro and I. Rakov, "Estimating the quality of databases," presented at 1996 Conference on Information Quality, Cambridge, MA, 1996.
[21] M. Bobrowski, M. Marre, and D. Yankelevich, "A homogeneous framework to measure data quality," presented at IQ 1999, Cambridge, MA, 1999.
[22] B. Pernici and M. Scannapieco, "Data quality in web information systems," presented at ER 2002, 2002.
[23] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information Systems*, vol. 29, pp. 133-146, 2004.
[24] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," presented at 7th World Wide Web Conference (WWW7), 1998.
[25] J. Hicklin, C. Moler, P. Webb, R. F. Boisvert, B. Miller, R. Pozo, and K. Remington, "JAMA: Java Matrix Package," [Online] Available: http://math.nist.gov/javanumerics/jama/, 2005.
[26] J. Cho and A. Ntoulas, "Effective Change Detection using Sampling," presented at VLDB Conference, Hong Kong, China, 2002.
[27] L. Cholvy and C. Garion, "Querying several conflicting databases," presented at ECSQARU-03 Workshop Uncertainity, Incompleteness, Imprecision, and Conflict in Multiple Data Sources, Aalborg, 2003.
[28] Jung Framework Development Team, "JUNG: Java Universal Network/Graph Framework," [Online] Available: http://jung.sourceforge.net/index.html, 2005.
[29] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. R. Vallacher, "Social Networks Applied," *Intelligent Systems, IEEE [see also IEEE Expert]*, vol. 20, pp. 80-93, 2005.