

Towards a Quality Model for Data Selection in Collaboratories

Yogesh L. Simmhan
Beth Plale & Dennis Gannon
Computer Science Department
Indiana University

IEEE Workshop on Workflow & Data Flow for Scientific Applications (SciFlow)

2006-04-08

Motivation: Data Driven Apps

- Compute-intensive '*in-silico*' experiments used in large collaboratories
- Built as dataflow applications using *Workflow Frameworks* that run on *Grids*
- **Data Driven Applications** composed of *hundreds of tasks* that produce *hundreds of data products in the order of terabytes*
- Generate *derived data products* that are reused by others in the *Virtual Organization*

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (SciFlow '06)

2

Data Selection Challenge

- Workflows need input from raw instrument data and derived data (or model data) from other experiments
- Scientific data products are queried based on metadata
 - Metadata is *complex* & not all metadata is exposed for query
 - Query results are *rarely exact* & manual filtering needed
 - Multiple “*similar*” data products with *different qualities* for an application available: A goodness-of-fit problem

Problem: *How does a scientist select the best quality dataset(s) for their application from multiple that qualify?*

Solution: *Translate subjective perception to tangible quantitative quality score for data using quality metrics*

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

3

Outline

- Use Cases from Meteorology
- Quality Model
 - Quality Metrics
 - Quality Profile
 - Evaluating Data Quality
- Implementation
- Related Work
- Future Work

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

4

Use Cases from Meteorology

- Linked Environments for Atmospheric Discovery (LEAD)
Meteorology research & education project
- **Intrinsic Metadata**
 - An atmospheric scientist performing *24 hour* weather simulations over the *continental US* needs:
 - *Current weather data* obtained from
 - seven *instrument data* sources
 - based on *quality control flags* meeting a certain threshold
 - provided by *Forecast Systems Lab*
- **Provenance**
 - A meteorologist compares the results of a prototype weather prediction model with a standard model's result for:
 - the hurricane *Katrina*
 - prefers *derived* data generated by the *WRF* forecasting model over the *ARPS* prediction model
 - needs it configured with a *smaller grid size*

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

5

Use Cases from Meteorology...

- **Quality of Service**
 - A postdoctoral student needs to select one of *several data products*, some of which are replicas:
 - from *different data repositories*
 - would like to balance the *timeliness & reliability* of staging the data products/replicas
 - against their *usability* for his experiment
- **Community Knowledge**
 - A high-school student, as part of her homework, needs to visualize *precipitation data*. She's not sure which one of *several sources* to use, and
 - would like to use an *expert user's opinion*

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

6

Quality Model

- **Quality Factor:** Metadata attributes upon which users base their data quality expectation E.g. *publisher, grid-resolution, throughput*
- Metadata attributes for a data product are a set of *name–value(s)*
- **Quality Score:** A standardized *numerical scale* (e.g. –7 to +7) that is a quality measurement for a quality factor (per user & application)
- **Quality Metric:** A function over a set of related quality factor scores that aggregates them into a quality score for that metric.
- Four classes of metrics are proposed, as motivated by the use cases:
 - **Intrinsic Metadata**
 - **Data Provenance¹**
 - **Quality of Service**
 - **Community Perception**
- Similar to content-based, economic & social filtering
- Each metric defined over a (possibly overlapping) set of attributes

¹ A Survey of Data Provenance in e-Science, Simmhan, et al., *Sigmod Record*, 34(3) 2005

Intrinsic Metadata

- **Quality Constraint:** A user-specific rule used to map a quality factor's value to a *quality score* and its *importance* (weight).
- Intrinsic metadata metric is defined over inherent properties of the data. E.g. *temporal range, domain keywords, quality flags*
- User constraints map an attribute's value to its quality score
- Similar to traditional querying along with weights

```
switch (publisher) { // Intrinsic Metadata Constraint
  case == 'National Weather Service' : return qualityScore=+7;
  case IN {'NCEP', 'Unidata'} : return qualityScore=+5;
  default : return qualityScore=0;
} && return weight=0.3;

switch (startDate) {
  case < 2005-11-31 : return qualityScore=-7;
  default : return qualityScore=0;
} && return weight=0.1;
```

Data Provenance

- Metric defined over the derivation history of data, with attributes
 - *creating process*, (e.g. WRF model, ADAS pre-processor)
 - its *input data products* (e.g. guid:493a12ec-619d-4081-b751...)
 - *configuration parameters* (e.g. gridSize=0.1, threshold=29.2)
- Can use user-defined constraints on these attributes, but *non-intuitive*
- Process can be thought as a *quality transformation function* that uses input data (or their quality score) & configuration parameters to generate a quality score for output data products
- Use *machine learning* to derive the process' quality function
 - Provide sample data products with known scores for a process
 - Build a *decision tree* for the process with the input data & parameters as variables

```
switch (provenanceProcess) { // Provenance Constraint
  case IN {'WRF'} : return qualityScore=+7;
  default : return qualityScore=-7;
} && return weight=0.1;

return qualityScore=provenanceScore() && weight=0.2;
```

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

9

Quality of Service

- Metric based on ability to *access & transfer* data for an application use at a certain *cost*
- *Availability, reliability, timeliness* (throughput+data size), *resource cost, accessibility*
- Allow users to trade-off QoS with other quality factors
- QoS metrics already exist for resource brokering, but we expose this to user

```
switch (transferTimeSecs) { // QoS Constraint
  case < 10 : return qualityScore=+6;
  case > 60 : return qualityScore=-7;
  default : return qualityScore=+2;
} && return weight=0.4;

switch (resourceCost) {
  case <= 1000 : return qualityScore=+7;
  case <= 2500 : return qualityScore=0;
  case > 2500 : return qualityScore=-5;
} && return weight=0.4;
```

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

10

Community Perception

- Different user levels possible within organization. E.g. *Expert, Peer, Student*
- Leverage community knowledge about data quality available within users' profiles
 - What is the quality score assigned by a certain class of users to this data?
- Metric based on User constraints over
 - Aggregate quality score for a data product by a user group
 - Frequency of usage of data
- Balance privacy of user profiles against sharing of knowledge

```
// Community Perception Constraint
return qualityScore=expertUserScore() && weight=0.5;

switch (usageFrequency) { // QoS Constraint
  case < 10 : return qualityScore=-6;
  case < 50 : return qualityScore=+1;
  case < 100 : return qualityScore=+5;
  default : return qualityScore=+7;
} && return weight=0.4;
```

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

11

Evaluating the Quality Model

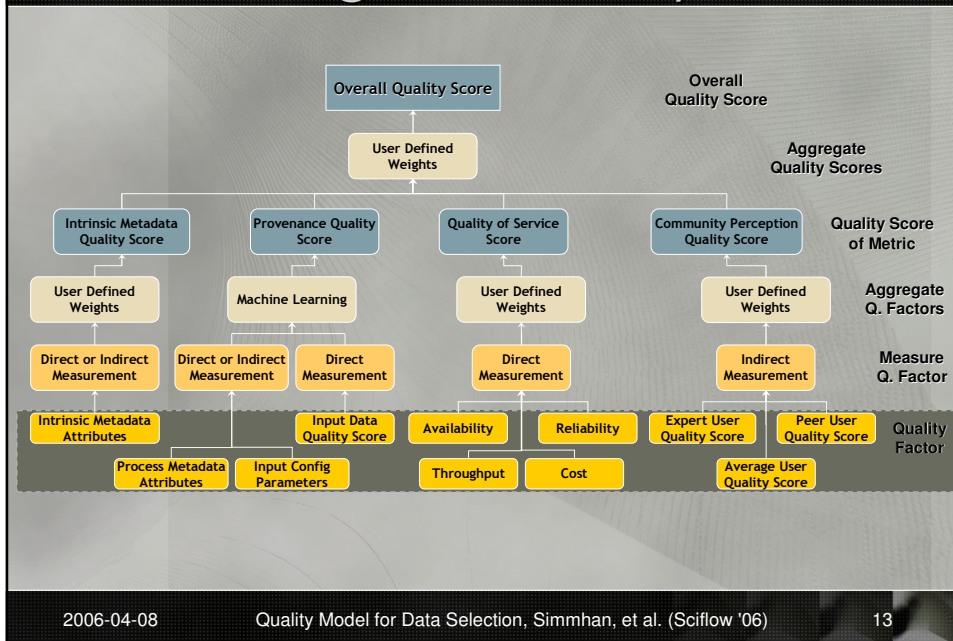
- **Quality Profile:** A set of quality constraints for a user or their application that captures all quality requirements
 - Global quality profile for user
 - Application specific quality profiles
 - Inheritance properties to make profiles reusable
- **Overall Quality Score:** The aggregated quality score evaluated for a data product by applying the quality metrics and quality profile to its metadata.
- Ways to convert, measure & aggregate quality factors into quality scores, per metric and for the data product

2006-04-08

Quality Model for Data Selection, Simmhan, et al. (Sciflow '06)

12

Evaluating the Quality Model



Implementation

- Currently implementing a Virtual Organization-wide *Data Quality Broker* to provide quality scores for data products
- Applies users' XML quality profiles to data product metadata to evaluate quality score
- Caches quality scores for data products, their metrics & decision trees created for processes
- Interacts with external information services like the *myLEAD*¹ metadata catalog, *Karma*² provenance service, and resource brokers
- Will be used for automated data selection of inputs for LEAD workflows at runtime

¹ Active Management of Scientific Data, Plale, et al., *IEEE Internet Computing*, 9(1) 2005

² Performance Evaluation of the Karma Provenance Framework, Simmhan, et al., *Intl' Provenance & Annotation Workshop (IPAW)*, 2006

Related Work

- Querying on metadata catalogs
 - Queries usually restricted to intrinsic metadata without holistic search over all quality metrics
 - Limited to sorting on attribute values...sophisticated ranking algorithms missing.
 - Some automation using metadata ontologies
- *DaQuinCIS*¹ data quality system
 - User feedback used to evaluate credibility of publishers
 - Ranking of results on publisher-trust & accuracy; no provenance, quality metrics
- Total data quality management (TDQM)
 - Surveys to evaluate quality perception in business
 - Error detection in data warehouses used to reconcile data from different sources
- Statistical quality control used with raw instrument data to find deviations & to flag outliers

¹ The DaQuinCIS Architecture: A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems, Seannapieco, et al., *Information Systems*, 2005

Conclusion & Future Work

- Motivated the need for data quality metrics & quality metrics that are used
- Defined a quality model to evaluate quality constraints to provide a quality score
- Intuitive way for users to provide quality constraints. Can it be automated?
- Using quality scores
 - Automated data selection
 - Data custodians to decide on caching and archiving
 - Provide feedback to data creators
 - Economic model for data providers to price commercial data
- Evaluate the usefulness & effectiveness of quality metrics
 - How to handle missing metadata?
 - Use of confidence intervals in quality prediction

