Provenance-Aware Automated Data Quality Control

Qing Liu, Greg Timms, Yanfeng Shu, Daniel Smith and Andrew Terhorst
Tasmanian ICT Centre, CSIRO, GPO Box 1538 Hobart TAS 7001, Qing.Liu@csiro.au

INTRODUCTION

As automated data collection has become more commonplace (e.g. through industrial and environmental sensors and sensor networks), the volume of data produced has risen exponentially. In order for this data to be shared and re-used, it is crucial that automated techniques for the assessment of data quality are also developed. Such techniques have begun to appear in the literature in recent years [1, 2], combining data statistics and domain expertise to produce data quality flags or estimates of uncertainty.

Where the quality of data is assessed by the organisation responsible for collecting the data, these approaches are relatively easy to implement. However, in many circumstances, it is unavoidable that users will have to sometimes use data provided by a third party. Therefore, knowledge of how data quality is assessed is critical for users to decide if data is trustworthy and fit-for-purpose. In this paper we discuss how data provenance can enable proper assessement of an automated quality control process.

AUTOMATED DATA QUALITY CONTROL

Automated quality control systems for streaming data take a number of inputs based on either domain expertise or data statistics, and estimate a quality metric for the data. Examples of domain expertise inputs in a marine environment include the expected time taken for a particular sensor's output to degrade due to algal growth (biofouling), and the accuracy of a particular sensor once calibrated. Examples of data statistic inputs include the rate of change of the sensor output, and correlations between the output of a sensor and other sensors nearby. The automated data QC process typically outputs a data flag (e.g. good, probably good, bad) and/or an interval estimate (error bar). However, details off the whole QC process is generally not stored and therefore, cannot be queried. In the case of third party sensor data, domain expertise may not be available, reducing the possibility of an accurate quality assessment.

PROVENANCE

Provenance is a very broad topic that has many meanings in different contexts. The W3C Provenance Incubator Group defines provenance as follows: Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility [3].

INTEGRATING PROVENANCE INTO AUTOMATED DATA QUALITY ASSESSMENT

In this work, we describe a provenance-aware framework that provides tracebility for an automated data quality control process. This allows users to make appropriate decisions on the trustworthiness and fitness for purpose of the data by better understanding sensor readings and associated quality flags and/or the estimates of uncertainty.

This is achieved by tracking the information life cycle of data and making it accessible to users. Therefore, a provenance information model needs to be designed to capture relevant information. We analyze the automated data quality control process and identify some key concepts that need to be accounted for in the information model:

- Sensor Reading: a sensor observation paired with its timestamp;
- Sensor Description: name, type, model, location measurement range, measurement accuracy and measurement resolution, calibration history and maintenance history etc. Such a description can be captured in the information model to assess the quality limits of a particular sensor series;
- Automated Quality Control Process: the methods used to pre-process the data and provide an estimate of the data quality.
- Quality Process Parameters: the set of parameters used to assess the quality of data within the quality control process. These parameters are associated with the assessment criteria (i.e. series gradient, correlation with neighouring sensors) and can take the form of thresholds, probability distributions and/or membership functions. It is particularly important to capture the parameters of the flag-based methods in order to interpret the flag meaning (i.e. what does a good quality flag or poor quality flag actually mean?);
- Data Quality Indicators: the assessments generated by the quality control methods include quality flags and/or the estimates of uncertainty. These indicators are captured in the information model to represent the sensor reading uncertainty;

Identification of the key concepts involvd in the automated data quality control process allows the design of a provenance information model that captures not only the data, but also the lineage relationships of the data.

We encode provenance by using the Web Ontology Language (OWL) for its well-defined formal semantics, which not only enables complex provenance modelling, but also facilitates provenance reasoning. To leverage existing provenance work, we extend the Open Provenance Model (OPM) for provenance representation. Domain models are also required that describe the domain knowledge, such as sensor descriptions/metadata and data quality assurance and control processes. We store provenance as RDF triples, which allows us to query provenance using SPARQL. Given the limited space, the model design will be discussed in future papers.

Based on the provenance model designed, next we present the Provenance-Aware Automated Quality Control Framework that provides provenance harvesting throughout the information life cycle, storing and querying capabilities for sensor reading and its uncertainties.

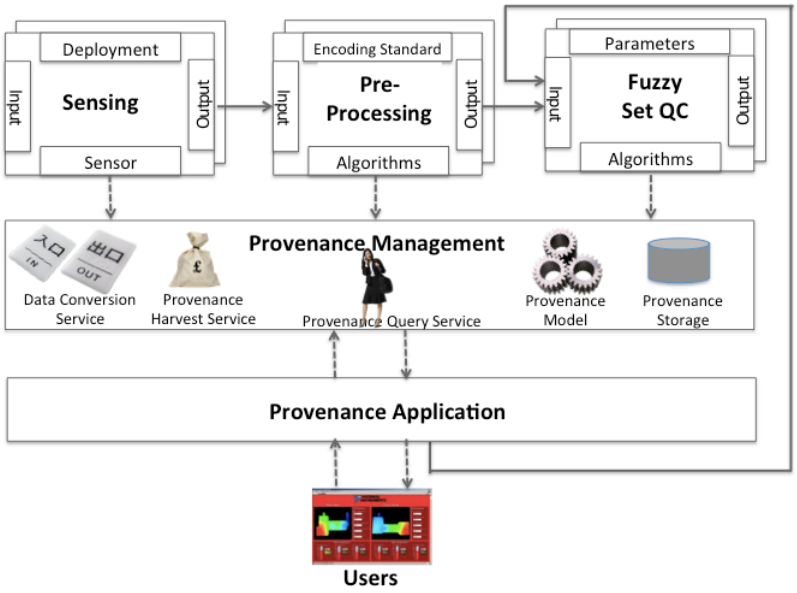

Figure 1: Provenance-Aware Automatic Quality Control Framework

Figure 1 shows the framework. The information life cycle of the automatic quality control process (from left to right) is presented at the top layer. The three grey boxes represent the high-level components involved. For each component, it includes the required input, control method, algorithm and output.

In the middle layer, the Provenance Harvest Service collects the provenance generated by the QC process and uses the Data Conversion Service to put the data into the Provenance Storage based on the provenance model developed. The Provenance Model enables the meaning of the data collected to be understood and the relationships among the data to be managed. The Proveance Query Service retrieves provenance from the Provenance Storage to answer the query received from the Provenance Application layer.

At the Provenance Application layer, the Fuzzy Set QC component needs to retrieve a sensor's calibration history and maintenance history, which is part of provenance captured through sensor description, to calcuate the sensor degradation time. Various other applications and visualization methods could be developed based on the user requirement to enable users to extract required information and discover if the data is trustworthy and fit-for-purpose.

CONCLUSIONS / FUTURE WORK

In this paper, we discuss the benefits of providing traceability of quality control process and present a generic provenance-aware framework. A detailed provenance model and semantic-based system will be implemented and presented in future. Furthermore, we will investigate how to use rich provenance to improve the automated QC process.

REFERENCES

1. Ferrero, A. and S. Salicone, *An innovative approach to the determination of uncertainty in measurements based on fuzzy variables*. IEEE Transaction on Instrumentation and Measurement, 2003, **52**: p. 1174-1181.
2. Timms, G.P., P.A. de Souza Jr., and L. Reznik, *Automated assessment of data quality in marine sensor networks*, in *Proceedings of IEEE Oceans 2010,* Sydney, Australia 24-27 May 2010, doi: 10.1109/OCEANSSYD.2010.5603827.
3. W3C Provenance Incubator Group final report. Available at: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/#A_Working_Definition_of_Provenance

ABOUT THE AUTHOR(S)

Qing Liu received her Ph.D. degrees in computer science from the University of New South Wales, Australia, in 2006. After graduated, she joined the University of Queensland as a postdoc researcher and worked in the area of spatial-temporal data management. Qing currently works as a research scientist with CSIRO since 2007. Her research interest is to develop effective and efficient solutions for managing, integrating and analysing large amount of complex datasets for the biology and hydrology applications.

Greg Timms received the B.Sc. (Hons) and Ph.D. degrees in physics from the University of Sydney, Australia, in 1993 and 1997 respectively. In 1997, he joined the Australian Nuclear Science and Technology Organisation where he spent five years investigating the environmental impacts of mining, focusing on the physical transport of reactants and pollutants within mine wastes.

Since 2002, Greg has been with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), initially engaged in research on microwave communication networks and then leading a team which developed a novel 190 GHz millimetre-wave imager in 2006.

For the past four years Greg has been based at the Tasmanian ICT Centre at CSIRO's Hobart site, where he has been part of a team developing low-cost sensor network and information system technologies for deployment in marine environments. Greg's particular interest is in the development of techniques for automated quality control of real-time streaming data.

Yanfeng Shu is a research scientist at the CSIRO ICT Centre. Her research interests include but are not limited to database systems, semantic web, Peer-to-Peer systems, and sensor networks.

Daniel Smith received the B.Eng. (Hons) and PhD in engineering from the University of Wollongong, Australia in 2002 and 2007, respectively. He is currently a research scientist at the CSIRO Tasmanian ICT Centre where he has worked on technology projects in the aquaculture and marine space. His research interests include audio processing, blind signal separation and quality control.

Andrew Terhorst is a geoscientist with considerable experience in resource management. This includes more than 10 years working as an exploration geologist, running his own spatial technologies firm for 13 years, and directing projects in government research agencies the past 7 years. Andrew has participated in a range of interesting and challenging environmental projects over the years and has a record of accomplishment in innovation and leading change. He currently leads research into next-generation Sensor Webs at the CSIRO.