Improving Situation Awareness In Traffic Management*

Norbert Baumgartner
team Communication
Technology Mgmt. GmbH
Goethegasse 3
1010 Vienna, Austria
norbert.baumgartner@te-am.net

Wolfgang Gottesheim
University of Linz
Altenberger Strasse 69
4040 Linz, Austria
wolfgang@tk.jku.at

Stefan Mitsch
University of Linz
Altenberger Strasse 69
4040 Linz, Austria
stefan@tk.jku.at

Werner Retschitzegger
University of Linz
Altenberger Strasse 69
4040 Linz, Austria
werner@ifs.uni-linz.ac.at

Wieland Schwinger
University of Linz
Altenberger Strasse 69
4040 Linz, Austria
wieland.schwinger@jku.at

ABSTRACT

Information overload is a severe problem for operators of large-scale control systems, as such systems typically provide a vast amount of information about a large number of real-world objects. Systems supporting situation awareness have recently gained attention as way to help operators to grasp the overall meaning of available information. To fulfill this task, data quality has to be ensured by assessment and improvement strategies. In this paper, a vision towards a methodology for data quality assessment and improvement for situation awareness systems is presented.

1. INTRODUCTION

Large-scale control systems, as needed in domains like road traffic management [14] provide a vast amount of information from multiple data sources about a large number of real-world objects. As a consequence of information overload, human operators lack awareness about the overall meaning of the available information. Following Endsley's model of situation awareness (SAW, [12]), achieving awareness about the meaning of information is a mental process spanning three consecutive levels: (i) *perceiving* the objects in the environment (e. g., a traffic jam), (ii) *comprehending* the meaning of situations in terms of assessing relations between objects (e. g., an accident near a tunnel), and (iii) *projecting* future states (e. g., an accident will cause a traffic jam).

Situation awareness systems, such as presented in our previous work [5], are gaining importance as a way to help human operators cope with information overload by providing technologies to automate at least parts of this mental process. For this, situation awareness systems make use of automated reasoning algorithms that derive additional knowledge from input information on all three levels. Hence, the system's ability to fulfill its task of effectively helping human operators depends heavily on the quality of input information provided by data sources. Unfortunately, these sources may often (i) provide identical, incomplete, and most crucial, even contradictory information, (ii) exhibit different data distributions over time and space, and (iii) be characterized by frequent updates due to object evolution. On the perception level, this raises the need to not only *measure and assess a data source's quality* during both situation awareness system development and execution, but also to *maintain a single consistent system state* across multiple sources. On the comprehension level, this entails to correctly *focus assessment over time and space* in order to achieve a proper balance between information overload and starvation. Finally, on the projection level, this requires developers to *configure evolution distances* in order to make projections which closely approximate real-world evolutions. While our current research focusses on situation assessment and projection, we will direct efforts towards improving the quality of the data underlying situation awareness. Although necessary data quality assessment and improvement measures may seem straightforward if seen isolated, allowing for their interactions across levels makes the selection and implementation of appropriate measures highly challenging.

The intention of this paper is to present our vision towards improving situation awareness systems by improving data quality on all three levels. As the basis for this, typical data quality issues in situation awareness are discussed in the next section by the example of road traffic management. From these issues, quality dimensions for assessing both the quality of data sources and the effects of data quality improvement techniques are derived in Sect. 3. These dimensions are accompanied by metrics and envisioned components for increasing data quality. Finally, evaluation strategies and further ideas are discussed in Sect. 4.

*This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FIT-IT 819577.

VLDB '10, September 13-17, 2010, Singapore

2. DATA QUALITY ISSUES

In our ongoing research project BeAware! we have identified a number of data quality issues currently present in traffic messages recorded from Austrian highways over a period of four weeks. The recorded data set consists of 28,616 distinct traffic objects, comprising 25,269 traffic jams, 820 road works, 1,803 other obstructions, 614 accidents, 46 wrong-way drivers, and 64 severe environmental conditions, such as snow or ice on the road. In this section, we briefly present our most important findings to further motivate our work on data quality enhancement.

Since systems used for road traffic management are geographically widespread, we naturally encounter a variety of data sources. Among these sources are, for example, automated traffic jam detectors used on highly frequented highway sections to automatically report on traffic flow conditions, manually managed traffic information messages broadcasted by an Austrian radio station, or scheduling information from a system for managing roadworks on highways. Since road traffic information is highly dynamic, these sources are updated frequently and describe how objects evolve over time. The granularity of the information in these sources differs with their purposes since, for example, a highway operator is interested in accurately managing roadworks, while the information on road works broadcasted by the radio station is less detailed to better suit their customer's information needs. Other examples for this are jams that are caused by accidents which are reported by the radio station as accidents with the spatial extents of the traffic jams they cause.

The most common data quality issue encountered is duplicate information that exists not only if information from different sources describe one and the same real-world object (e. g., a traffic jam is reported by the automated detection system as well as by the radio station, based on driver reports), but also within a single data source (e. g., a traffic jam is long enough to be detected by multiple automated detectors). As these data sources are independently maintained, they are charaterized by differing update intervals and delayed updates, making it necessary to align messages before being able to meaningfully compare them.

Another quality issue that occurs frequently is that information on an object's validity period is only partially available. Finally, only some messages concerning real-world objects that no longer exist such as, e. g., a resolved jam or a cleared accident, are accompanied by a dedicated clearance message, while others simply disappear from the data sources.

3. TOWARDS DATA QUALITY IMPROVEMENTS IN SITUATION AWARENESS

Based upon the issues identified in the previous section, we envision a methodology for the assessment and improvement of data quality in situation awareness. In this section, we propose first steps towards this methodology.

As a recent survey on methodologies in this area [3] shows, data quality improvement methodologies comprises phases concerning *data quality assessment* and *data quality improvement*. In the assessment phase, data quality is measured in multiple *data quality dimensions*, a term traditionally employed for measure concepts [9], along intrinsic data values and contextual information about the data. In the improvement phase, the identification of causes of errors enables the selection of strategies and techniques to deal with these errors, as well as the design of a data improvement solution. Note, that it is crucial to repeat the assessment phase after performing improvements to evaluate the effects of selected improvement strategies. Figure 1 shows an architecture based on the data quality assessment and improvement components described in this chapter.

3.1 Perception Level

Data Quality Assessment. In literature, an overwhelming number of techniques to assess data quality has been proposed (e. g., [3], [16]). Based on these works and our experience in the domain of road traffic management, we have selected the following set of data quality dimensions for their applicability to this domain. For each dimension, we provide an informal description and a metric for the computation.

Completeness describes whether data contains all required information. For example, some sources provide only partial information due to missing values for validity periods. Completeness can be computed automatically, for example as the ratio of null to non-null values. For this, necessary contextual information is schema information.

Accuracy is a two-part dimension [3]: *syntactic accuracy* describes whether data reported by a source is among expected values, while *semantic accuracy* details whether reported data matches real-world objects. As [3] shows, most methodologies focus on computing syntactic accuracy which can be easily computed automatically if validity ranges for attribute values are specified in a schema. Semantic accuracy could be assessed in three ways: (i) by exploiting known accuracy characteristics of automated sensors, (ii) by selecting a reference data source, based on other data quality dimensioms, from the list of available sources, and comparing reported data against this source, or (iii) by comparing to a test data set created by domain experts.

Timeliness is another two-part dimension consisting of *age* and *volatility.* Age describes the difference between the reporting time of information and its actual observation time. For example, traffic jam detectors might only report their data in fixed intervals to minimize transmission costs, but still generate observation timestamps. Volatility illustrates the frequency of changes within a data source and is measured as the number of changes that occurred within a defined time frame.

Confidence describes whether information from a data source can be trusted in terms of a statistical measurement [15]. It is computed as the standard deviation within a defined time frame, assuming that the observed object did not change in the meanwhile. For example, a data source reporting traffic jams based on sensor data might yield a low confidence if the reported traffic jam length fluctuates wildly.

Coverage can be seen as *spatial coverage* and *temporal coverage.* In the domain of road traffic management, spatial coverage describes whether information available in a data source describes the road network entirely or only partially. For example, automated traffic jam detectors are only installed on highly frequented parts of highways, whereas information from a radio station covers the whole road network. Temporal coverage describes the availability of a data source, since, e. g., a radio station might only report traffic messages at peak hours. Both kinds of coverage can be com-

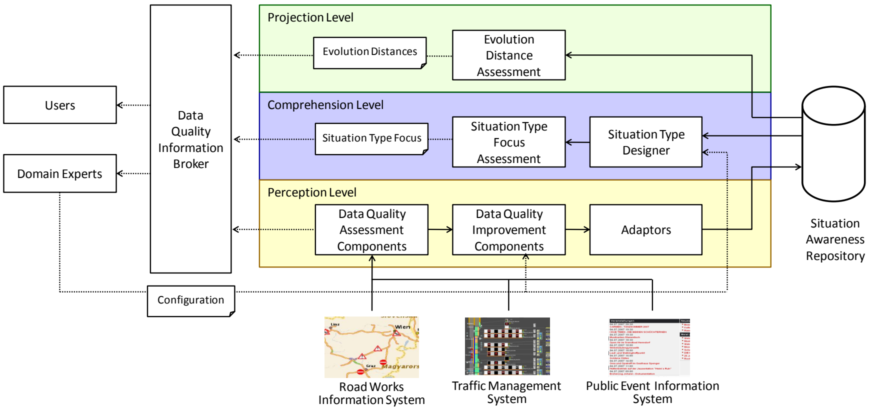
Projection Level
Evolution Distances
Evolution Distance Assessment
Comprehension Level
Situation Type Focus
Situation Type Focus Assessment
Situation Type Designer
Perception Level
Data Quality Assessment Components
Data Quality Improvement Components
Adaptors
Users
Domain Experts
Data Quality Information Broker
Configuration
Situation Awareness Repository
Road Works Information System
Traffic Management System
Public Event Information System

Figure 1: Data quality improvement architecture.

puted automatically.

Consistency denotes the extent to which domain-specific integrity constraints are violated. For example, such a constraint might describe that an accident cannot have a spatial extent larger than 50 meters. Given a list of integrity constraint, consistency can be computed automatically as the ratio of invalid data to all available data.

These quality dimensions need to be calculated during the assessment phase in order to select appropriate improvement strategies.

Data Quality Improvement. After having assessed data quality dimensions, a strategy for improving data quality has to be developed. We envision to provide a number of data-driven improvement techniques embedded into a a configurable domain-independent framework. The goal of this framework is to flexibly change components in order to meet the needs of different application domains. In its design we follow the three-step data fusion process presented in [8] comprising *schema matching*, *duplicate detection*, and *data fusion*. In the following, we concentrate on the instance level by means of duplicate detection and data fusion, assuming that structural heterogeneities have already been resolved.

Duplicate Detection. In a recent survey [4], we evaluated duplicate detection approaches from various domains for their applicability to SAW. Briefly summarized, current approaches lack comprehensive support for (i) quantitative as well as qualitative data, (ii) similarity dimensions for spatial and temporal attributes, (iii) detecting duplicates in data streams, and (iv) taking object evolution into account. We plan to re-use our situation assessment algorithms [5] to implement a duplicate detection component exploiting rules that capture knowledge from domain experts to capture clearly defined duplicates. For example, two objects are considered duplicates if they occupy the same spatial region and their validity periods overlap.

Data Fusion. Upon detected duplicates, the overall goal of data fusion is to increase completeness as well as accuracy of information. Quality dimensions serve as both indicators for problem areas requiring the application of data fusion strategies, and as performance measurement to judge their effect.

In order to increase completeness, rather straightforward strategies such as “Take the information”, which prefers newer values over older values, and “Trust your friends” [8], which prefers reliable data sources over others, can be viable if confirmed by quality dimensions (e. g., higher confidence and accuracy of a certain data source). Nevertheless, other approaches that are not only able to increase the completeness of information by substituting missing values with data from other sources, but also extrapolate new information from existing data, will be required. For example, the common problem of missing validity periods can be tackled by a component that computes an exponentially decreasing existence probability based on age and type.

In the domain of road traffic management, not all data sources describe all possible object types equally well. Thus, fusion strategies need to be configurable in order to, e. g., prefer information about road works from one source, but at the same time only regard traffic jams reported in another one.

3.2 Comprehension Level

In Endsley’s model people must combine and interpret information to achieve level 2 situation awareness (i. e., comprehension) , that is, according to Situation Theory [2], detect *relations* between objects by examining their properties, and upon that, pinpoint relevant *situations*. A situation awareness system can automate this process by trying to satisfy rules for particular relation and situation types [5]. As underlying relation semantics, a large body of different *relation calculi*, each focusing on a particular spatio-temporal aspect, from spatio-temporal reasoning research is suitable for situation awareness [6]. For example, a relation type PartiallyOverlapping of the Region Connection Calculus RCC [10] holds between two objects, if the spatial regions of these objects overlap. A sample situation type Traffic jam extending into a tunnel may use PartiallyOverlapping, which means that a corresponding situation should be brought to the operator’s attention in case a traffic jam object partially overlaps a tunnel object at run time. During situation type design, domain experts, however, may not be aware of the fact that particular relation types and, hence, situation types using them, are valid between a very large or a very small number of objects (i. e., the *focus* of designed situation types can easily become too broad or too narrow).

Data Quality Assessment. The contextual information for computing the focus of situation and relation types can be integrated into situation awareness systems in the form of relation calculi. By that, the quantity structure of underlying objects can be mapped onto relation calculi (for example, 20% of object pairs may partially overlap). The focus of a relation type is measured at the percentiles of the object pairs being in this relation in ratio to all object pairs. For example, the focus of a relation type may be narrow if it holds between less than 10% of all object pairs, or it may be broad if it holds between more than 50% of all object pairs. Assuming, that relation types are semantically (and, hence, statistically) independent, multiple relation type percentages can be aggregated to a situation type percentage, finally determining the situation type's focus.

Data Quality Improvement. By mapping the quantity structure of underlying object data onto relation calculi, we can bring the focus of situation types to the designer's attention already at design time, thereby leveraging negative effects at run time. Such a situation type design support component could additionally make use of relation type neighborhood [13] to find similar situation types [7] in order to suggest alternatives with a different focus to the designer.

3.3 Projection Level

The projection level of situation awareness enables timely decision making by anticipating future situations. In a situation awareness system, one possible technique is, for example, to model evolutions between landmark situations [7], i. e., knowledge in the form of rules that some particular *trigger situation* often precedes a critical *climax situation* that should be anticipated. Other approaches, such as relation neighborhood-based ones [1], start at a current state to enumerate all possible future states along the transitions in a relation neighborhood graph.

Data Quality Assessment. In any case, information on distances—such as probability and temporal distance observed in real-world evolutions—can increase the quality of projections. Such distances can at the same time be used as metrics for evaluating the quality of projections by comparing projected evolutions with observed evolutions in a test set.

4. DISCUSSION

The quality dimensions discussed in this paper play a crucial role in the evaluation of data quality improvement frameworks for SAW. Our evaluation strategy is threefold and structured along the levels of situation awareness. First, on the perception level, data quality measures evaluate the effects of our data improvement strategies and allow to compare different duplicate detection and data fusion components. On the comprehension level, we see the focus of a situation type as a possible quality dimension, subsuming other dimensions such as, e. g., the number of object comparisons required. Finally, evaluation on the projection level will consist of defining a test set containing relevant object evolutions and comparing projected evolution courses with those that actually occurred.

While data quality dimensions on the perception level have to be calculated during the assessment phase, some of them, like timeliness and confidence, can provide valuable information at runtime as they allow human operators and higher situation awareness levels to evaluate possibly inconsistent information.

Since duplicate detection results in pairwise comparisons to calculate similarity between data, strategies to reduce the number of necessary comparisons and thereby increase performance are crucial to fulfill the requirement of duplicate detection on data streams. We plan to include an extensible set of strategies comprising different clustering techniques as well as a sliding window approach ([11]) into our framework.

Based upon the vision presented in this paper, our future work will comprise an evaluation of current tools for assessing data quality dimensions for their applicability to situation awareness. Furthermore, a formal definition and a comprehensive validation of proposed data quality dimensions will be conducted in the form of a case study.

5. ADDITIONAL AUTHORS

6. REFERENCES

[1] K. R. Apt and S. Brand. Constraint-based qualitative simulation. In *Proceedings of the 12th International Symposium on Temporal Representation and Reasoning*, pages 26–34. IEEE, 2005.

[2] J. Barwise and J. Perry. *Situations and Attitudes.* MIT Press, 1983.

[3] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):1–52, 2009.

[4] N. Baumgartner, W. Gottesheim, S. Mitsch, W. Retschitzegger, and W. Schwinger. "Same, Same but Different"—A Survey on Duplicate Detection Methods for Situation Awareness. In *Proceedings of the 8th International Conference on Ontologies, DataBases and Applications of Semantics*, Vilamoura, Portugal, 2009. Springer.

[5] N. Baumgartner, W. Gottesheim, S. Mitsch, W. Retschitzegger, and W. Schwinger. BeAware!—situation awareness, the ontology-driven way. *accepted for publication in: International Journal of Data and Knowledge Engineering*, 2010.

[6] N. Baumgartner, W. Retschitzegger, and W. Schwinger. Lost in time, space, and meaning—an ontology-based approach to road traffic situation awareness. In *Proceedings of the 3rd Workshop on Context Awareness for Proactive Systems (CAPS)*, Guildford, UK, 2007.

[7] N. Baumgartner, W. Retschitzegger, W. Schwinger, G. Kotsis, and C. Schwietering. Of situations and their neighbors—Evolution and Similarity in Ontology-Based Approaches to Situation Awareness. In *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT)*, pages 29–42, Roskilde, Denmark, 2007. Springer.

[8] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1), 2008.

[9] I. Caballero, E. Verbo, C. Calero, and M. Piattini. A data quality measurement information model based on iso/iec 15939. In *ICIQ*, pages 393–408, 2007.

[10] A. G. Cohn, B. Bennett, J. M. Gooday, and N. Gotts. RCC: A calculus for region based qualitative spatial reasoning. *GeoInformatica*, 1:275–316, 1997.

[11] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining stream statistics over sliding windows. *SIAM J. Comput.*, 31(6):1794–1813, 2002.

[12] M. Endsley. *Situation Awareness Analysis and Measurement*, chapter Theoretical Underpinnings of Situation Awareness: A Critical Review, pages 3–33. Lawrence Erlbaum Associates, New Jersey, USA, 2000.

[13] C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1):199–227, 1992.

[14] H. Kirschfink, J. Hernandez, and M. Boero. Intelligent traffic management models. In *Proccedings of the European Symposium on Intelligent Techniques (ESIT)*, pages 36–45, Aachen, Germany, September 2000.

[15] A. Klein and W. Lehner. Representing data quality in sensor data streaming environments. *Journal of Data and Information Quality (JDIQ)*, 1(2):1–28, 2009.

[16] R. Y. Wang and D. M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996.