The Quality of Temporal Information

(Practice-Oriented Paper)

Eric Hughes, Ken Smith
The MITRE Corporation
{hughes,kps}@mitre.org

Abstract

We have investigated the needs of the government and military intelligence community to manage temporal information (sequences, series, periodic events, etc.) and to document and utilize the quality of information. We find the overlap between these needs to be significant – perhaps more so than previously recognized. This position paper explores this overlap and relates our experiences to business intelligence.

Background

We are embarking on an applied research effort to help the Intelligence Community provide, maintain, and utilize information annotated with simple measures of *quality*. Intelligence domains (whether government or business) are forced to utilize and produce information of imperfect quality. Information products are frequently approximate, obscured, or of uncertain reliability. Even if all information inputs were perfect, this would still be so. For example, a business analyst may be able to use accurate and timely data on the price of a company's stock, but human judgment will still be needed to evaluate that company's financial health. In many cases we can improve information quality by simply changing the way systems and processes manage and deliver data, but we can never hope to deliver perfect information for all needs. Here we assume that information shared between producers and consumers is annotated with its quality, according to some agreed-upon metadata model. Our focus is not on the particular quality measures, but on the importance of a temporal model to both quality measures and the underlying information.

To initiate our research, we worked with a variety of organizations to identify deficiencies in information management approaches and in data quality. Somewhat to our surprise, many groups that identified a need for quality improvements and annotations also expressed a need to manage temporal information (often in the same breath). We found that these could not be addressed separately, but required a combined model addressing aspects of the temporal quality of data. While it is not surprising that such a model would address concepts like currency (denoting how long ago a value was thought to be correct), here we explore the relationship between data quality and other temporal features, such as *event orderings* and *time-variance*. Specifically, we have identified a need to annotate temporal events and time series data with both qualitative and quantitative measures of quality. Here we discuss these findings and propose to measure quality for information that includes temporal data.

The Need To Model Temporal Information

In the following, we illustrate the dependence of useful data quality measures on *explicitly* modeled temporal information, not present in traditional standard data models (such as the relational model).

Events, Duration, and Their Relationships

Consider the case of a company that occasionally acquires smaller companies. A temporal representation of these acquisitions might define events for merger talks, purchase offers, government review, etc. These events have a duration, as do the intervals between events (e.g., how long between the end of the talks and the initial purchase offer). It is *also* frequently crucial to track the temporal relationships between events. For example, it may be more significant when a purchase offer is made *before* the start of merger talks, than when an offer is made *during* talks. In fact, this relationship may be more important than the mere fact that an offer was made. Events and duration can be represented in standard data models, however with a loss of semantics. Relationships between these, such as *before* and *during*, are best represented by explicit us of temporal information models.

Next, consider the measure of the number of widgets held by a company in its warehouse. An inventory system will report a precise number for any given instant, conveying the assumption that movements into and out of the warehouse are automatically and immediately entered. However, these "precise" numbers may not reflect reality. For example, the warehouse may take deliveries in the morning and load shipments in the afternoon. Then, the standing inventory of widgets reported by the inventory system might only be truly accurate at night, or perhaps never. Although events, their duration, and their relationships are not typically explicitly modeled in applications such as this, note that critical data quality measures are nonetheless a function of such temporal information. We argue such information should therefore be explicitly modeled.

Time-Varying Data

Consider a company that regularly invests in a given technology, and whose investment varies consistently over the fiscal year. A conventional approach would record each investment event (including date and amount) as a row in a table. However, this approach does not capture the temporal knowledge of the predictable variance over time associated with the company's behavior. For example, MITRE's fiscal year typically begins October 1st, at which time our projects predictably make a large number of initial purchases. Using a temporal model, we could also define a function describing the yearly pattern of investment. For periodic data (or similar examples), the pattern can be described as a simple function (e.g., sinusoid) of time. This function can be used to derive quality measures for the original investment event data. For example, deviation from the fiscal year cycle may be as significant as investment amounts are to a business analyst because it can signal questionable quality of data, or exceptional circumstances. Explicitly representing concepts like "period" and "skew" also supports the derivation of quality metadata: a skew of less than 10% for a particular period may be deemed unimportant.

The irregularities of calendars interact with the quality of certain data in interesting ways. Deviation from expected behavior within a work-week may or may not be significant, depending on "calendar effects". For example, MITRE's fiscal year 2001 began on October 2 (a Monday) to align with the MITRE work-week (which starts on Monday). However, a one-day deviation might be significant in other years. So, quality measures for some temporal information (e.g financial data, but not weather data) must be computed with awareness of the calendar. Again, temporal models are designed to address calendar effects, while traditional models support date and time types, but it quickly becomes complicated when we consider holidays, leap years, and the like.

Available Temporal Models

Temporal models abound [Vassilakis 1996, Sistla 1997, Bertino 1998, Kurutach 1998, Randall 1998, Bettini 2000], as do temporal query languages [Abiteboul 1996, Combi 1997, Bertino 1998, Erwig 1999]. Some approaches address the possibility that a data sample may be missing (a value is not reported or sensed), e.g., [Bertino 1998]. Others address currency [Finger 1998] or temporal precision [Combi 1997, Van Der Cruyssen 1997]. Some temporal work is directed at anomaly detection (e.g., [Lane 1999]). We are not aware of a temporal model that directly addresses quality.

Many existing database management systems (DBMSs) measure events without the benefit of a temporal model more sophisticated than a *date* data type and simple associated operators on dates (e.g. >, =). An acquisition tracking system might record the date that merger talks begin and the date that the acquisition is approved. However, these systems have no means of assessing the quality of such dates. For example, what should be recorded when an approval is dated "this week", following "3 months" of talks? Such ambiguities are common and limit the ability of conventional databases to manage intelligence information. As a result, measures of the temporal quality of the information recorded about such events is not captured, and is therefore difficult (if not impossible) to derive when event data is retrieved.

Discussion

The quality measures alluded to in these examples could be derived from conventional, non-temporal representations of the data. Relationships like *before* and *during* can often be calculated from dates. If we have an accuracy measure for start and end date elements, we can compute the accuracy of the resulting *duration*. Two dates can be compared to estimate which occurred *before* the other, again given knowledge of the accuracy of each. Calendar effects can be accounted for by standard routines, where available (e.g., UNIX functions that return the day of the week given a date). Anomalies in time-varying data can be identified by scanning the data or by keeping separate tables for missed samples (ignoring normalization concerns). Sample rates and other parameters can be recorded separately or provided as metadata.

However, we suspect that quality metadata will be more consistently used by systems with a rich temporal modeling capability. First, the information provider is more likely to report quality, since much less additional work is required to do so. Second, a powerful temporal query language is an important enabler of the exploration of temporal quality;

we would not expect an information provider to stretch the bounds of what is supported to invent such a language. Third, the simpler measures of quality supported by a temporal model are more likely to be computed correctly. Clearly, an incorrect derivation used to annotate information with quality metadata is at best counter-productive. Finally, temporal models and the mathematics that underlies them leads to more consistent representations, which in turn leads to more consistent measures of quality. In other words, it becomes possible to define standard quality measures for temporal data, and temporal models for quality measures, that are commonly understood and trusted. In essence, the information provider gives up some flexibility when choosing to use a temporal model to represent information, but this use increases the chance that common quality measures can be applied.

There are many ways that a temporal model might be used to address quality. For example, automatic extraction of merger events from natural language text would lead naturally to a *time* data element that is accurate to one time quantum (second, minute, hour, day, week, month or year). This leads to an enumerated type for *time_accuracy* that supports the math needed (e.g., to compute the time_accuracy of a sum). A temporal query language would support use of the time element, but would need to be extended to handle time_accuracy. For example, consider a query language in which "all events between March and June 2000" is interpreted as events that might fall within the range.

Position

We see quality measures as a way to improve confidence in, and indirectly improve the quality of, information provided. We recognize that over time, through experience, users develop a general understanding of the quality of an information source, but we believe that explicit representation of quality through formal annotation contributes to greater consumer confidence. This is especially true in modern enterprises characterized by data which rapidly migrates out of its original context, in which quality measures may be implicitly well-understood, and within hours or even minutes appears in multiple “foreign” contexts such as a data warehouse or a web site. Explicit representations may also clarify implicit quality information, making it more useful: as in the investor who may be more willing to buy a technology stock if it is recommended by a source with annotation of 75% accuracy on technology stock picks, than one recommended by a generally reliable source.

However, there are several practical impediments to providing *temporal* quality annotations:

1. These annotations may require a more precise understanding of the semantics of the temporal information than is normally available. For example, to compute the probability that one event happened before another, one must know that the dates recorded are accurate to within one day.
2. Conceptually simple quality measures, like the accuracy of a duration value derived from two dates, may be difficult to compute. For example, if one date is extracted from a news report, its accuracy may be a bell curve around the reported day. If the other date is provided by a company's financial data, it may represent an arbitrary

date within the reporting week. More general quality measures may be similarly complex to apply to temporal information.

3. We suspect that consumers will need to understand the model from which quality annotations are computed in order to interpret it correctly. For example, we find error bars on a time line more convincing when we know that they are computed by adding the errors of contributing dates. Similarly, qualitative quality measures need to be well-understood. We believe that consumers need insight into the computation used by a producer, so that consumer understanding of the information is improved. This insight also facilitates feedback from consumer to producer on quality. There is also some advantage to quantitative quality measures, as they permit derivation of similar measures for aggregated values.

We expect that the explicit use of temporal information models to support the derivation of quality measures will avoid many of these impediments for a significant fraction of intelligence information. Based on this overlap of interests, one might extend a temporal information model to define quality metadata, such as currency, temporal accuracy, etc. It is important to be able to compute this metadata easily, so that it does not delay wider adoption of temporal models and quality metadata.

An analogy can be drawn to spatio-temporal models, which combine solutions to the seemingly orthogonal concerns of space and time. Such models have received attention because they address an important cross-section of the needs of certain domains. Separating these concerns is possible but does not always serve to simplify the overall solution. As a result, many spatio-temporal models have been proposed for a variety of domains [Nozawa 1999, Roddick 1999, Theriault 1999]. Interestingly, our arguments on the utility of temporal models for deriving quality annotations also apply to the use of spatio-temporal models.

In conclusion, we seek ways to help humans distill concise, relevant information from widely-varying measurements on the state of the world. Virtually all such measurements have temporal aspects, from the simplest measurement event to complex models of the measurement and the state being measured. We advocate the use of temporal models in unison with computed quality annotations. We also advocate informing information consumers of these models and computations to improve understanding of quality and to indirectly encourage improvements in information quality.

References

[Abiteboul 1996] Serge Abiteboul, Laurent Herr and Jan Van den Bussche, "Temporal Versus First-Order Logic to Query Temporal Databases", Symposium on Principles of Database Systems, 1996, p. 49 - 57.

[Bertino 1998] Bertino, E.; Ferrari, E.; Guerrini, G., “An Approach to Model and Query Event-Based Temporal Data”, Proc. Wkshp. on Temporal Representation and Reasoning, 1998, p. 122-131.

[Bettini 2000] Claudio Bettini, Sushil Jajodia, X. Sean Wang, "Time Granularities in Databases, Data Mining, and Temporal Reasoning", Springer-Verlag, 2000.

[Combi 1997] Combi, C.; Cucchi, G., "GCH-OSQL: A Temporally-Oriented Object-Oriented Query Language Based on a Three-Valued Logic", TIME, 1997, p. 119-126.

[Erwig 1999] Erwig, M.; Schneider, M., “Developments in Spatio-Temporal Query Languages”, DEXA, 1999, p. 441-449.

[Finger 1998] Finger, M.; Da Silva, F.S., "Temporal Data Obsolescence: Modelling Problems", Proc. Wkshp. on Temporal Representation and Reasoning, 1998, p. 45-50.

[Kurutach 1998] W. Kurutach, "Processing of Binary Temporal Constructors for Fuzzily-Bounded Time Intervals in Temporal Databases", NAFIPS, 1998, p. 156-160.

[Lane 1999] Terran Lane and Carla E. Brodley, "Temporal Sequence Learning and Data Reduction for Anomaly Detection", Trans. Inf. Syst. Secur., Aug. 1999, p. 295 - 331.

[Nozawa 1999] H. Nozawa, N. Saiwaki, S. Nishida, "Spatio-Temporal Indexing Methods for Moving Objects for Highly Interactive Environment", Conf. on Systems, Man, and Cybernetics, vol. 6, 1999, p. 7-12.

[Randall 1998] Randall, D.J.; Hamilton, H.J.; Hilderman, R.J., "Generalization for Calendar Attributes Using Domain Generalization Graphs", Proc. Wkshp. on Temporal Representation and Reasoning, 1998, p. 177-184.

[Roddick 1999] Roddick, J.F.; Grandi, F.; Mandreoli, F.; Scalas, M.R., "Towards a Model for Spatio-Temporal Schema Selection", DEXA, 1999, p. 434-440.

[Sistla 1997] Prasad Sistla, A.; Wolfson, O.; Chamberlain, S.; Dao, S., "Modeling and Querying Moving Objects", Proc. Data Engineering, 1997, p. 422-432.

[Theriault 1999] Theriault, M.; Seguin, A.-M.; Aube, Y.; Villeneuve, P.Y., "A spatio-temporal data model for analysing personal biographies", DEXA, 1999, p. 410-418.

[van der Cruyssen 1997] van der Cruyssen, B.; De Caluwe, R.; De Tre, G.; Vandycke, D., "A Theoretical Model for Crisp and Fuzzy Time", NAFIPS, 1997, p. 63-67.

[Vassilakis 1996] Vassilakis, C.; Georgiadis, P.; Sotiropoulou, A., "A Comparative Study of Temporal DBMS Architectures", DEXA, 1996, p. 153-164.