ORION: Managing Uncertain (Sensor) Data

Sunil Prabhakar
Department of Computer Science, Purdue University
West Lafayette, Indiana, USA
sunil@cs.purdue.edu http://orion.cs.purdue.edu/

1 Sensor Data Uncertainty

An important quality of sensor data is that it is often uncertain or imprecise. This uncertainty can be an inherent aspect of the data (e.g. due to known errors in the measuring device, such as the Gaussian error in GPS readings), or it may be introduced in order to achieve scalability [2, 1], or to ensure a certain level of privacy [4].

Existing database management systems provide virtually no support for handling imprecise or uncertain data. It is possible to store some types of data uncertainty as an extra attribute in a table, however, all the necessary functionality needs to be provided by the user. Furthermore, as discussed below, supporting uncertain data requires the addition of significant new functionality that is not easy to achieve simply through user-defined functionality.

In addition to sensor data, uncertainty naturally arises in a large number of applications including scientific data, web data integration, machine learning, and information retrieval. Consequently, the problem of uncertain data management is receiving renewed attention from the database community. The ORION (http://orion.cs.purdue.edu/) project is an effort aimed at developing a novel database management system with native support for uncertain data. The initial goal was to primarily handle sensor data uncertainty. Currently, the scope of ORION has been expanded to encompass more general types of uncertainty in support of applications such as those mentioned above. Below we discuss the approach being taken in developing Orion and some of the related challenges.

2 Uncertain Data Models

At the basic level, sensor data uncertainty can be handled using probabilistic distributions for attributes [1]. Under this model, the value of a given attribute is represented as a collection of alternative values, each with an associated probability (for discrete alternatives), or a range of values with an associated probability density function. A more general model of uncertainty allows multiple attributes to be jointly distributed (correlated). In the extreme case, all the attributes of a tuple may be jointly distributed. The pdf values themselves do not have to add to 1 – in this case, the missing probability indicates the likelihood that the attribute's value is NULL, or that the entire tuple may not be present in the relation (this is also called tuple uncertainty). For many applications, only tuple uncertainty is of interest.

In addition to the model for the data, it is necessary to define the semantics of relational operators and queries. The most common approach is to use *possible worlds semantics* (PWD). An alternative is to provide procedural semantics for the various operations. The earlier version of Orion supports attribute uncertainty and a limited set of operators, and procedural semantics. We are currently in the process of enhancing the uncertainty model of Orion. The new model will support attribute uncertainty, correlated attributes, and tuple uncertainty, in addition to missing values. Since possible worlds semantics are more intuitive from a user-perspective than procedural semantics, the new model will support PWD. The new model is motivated by a number of representative applications including sensor databases.

3 Database Support for Uncertain Data

In addition to the development of appropriate models, there are several key components that need to be developed in order to achieve the overall goal of a providing native support for uncertain data. We aim to address a broad range of outstanding issues and develop a prototype uncertain data management system that will be validated on a sample application. In particular, the goals of the Orion project are:

1. **Specification of probabilistic queries.** This involves the definition and semantics of relational operators over uncertain data, and evaluation algorithms. Our current work supports probability threshold versions of the basic relational operators for the earlier Orion model. There is also a need to explore the need for SQL extensions for new query types (e.g. queries that manipulate probability values directly).

2. **Quality and Reliability:** As data becomes uncertain, query results also become imprecise (probabilistic). Consequently, a new issue of the quality of query results becomes important. There is need for developing metrics for both query quality and also reliability as data becomes uncertain. We introduced query quality metrics for attribute uncertainty in [1].

3. **Efficient query evaluation techniques:** As with regular queries (even more so with uncertain data), efficient execution of queries is critical. There is a need to develop novel including indexing and query processing and optimization techniques. In earlier work we have proposed new index structures for discrete and continuous attribute uncertainty [3, 5].

4. **Prototype development:** The proposed models and query processing algorithms etc. have been implemented in Orion as an extension of PostgreSQL. This enables the validation of the new models and also highlights implementation issues. The future development of Orion will continue to build on top of PostgreSQL.

Data Privacy Although uncertainty in data is often undesirable and unavoidable, there are instances where uncertainty is advantageous. In particular, uncertainty can be introduced into data in order to provide some degree of privacy [4]. It is clear to see that this comes at the price of query quality. We believe that this is an interesting direction for work on uncertain data and plan to explore it further.

References

[1] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *Proc. SIGMOD*, 2003.

[2] R. Cheng and S. Prabhakar. Managing uncertainty in sensor databases. In *SIGMOD Record issue on Sensor Technology*, December 2003.

[3] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *Proc. VLDB*, 2004.

[4] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar. Preserving user location privacy in mobile data management infrastructures. In *Proc. of the 6th Workshop on Privacy Enhancing Technologies*, 2006.

[5] S. Singh, C. Mayfield, S. Prabhakar, Shah R., and S. E. Hambrusch. Indexing uncertain categorical data. In *Proc. of the 23rd IEEE Intl. Conf. on Data Engineering (ICDE)*, 2007.