Chapter 34
Modeling, Querying, and Mining Uncertain XML Data

Evgeny Kharlamov
Free University of Bozen-Bolzano, Italy & INRIA Saclay, France

Pierre Senellart
Télécom ParisTech, France

ABSTRACT

This chapter deals with data mining in uncertain XML data models, whose uncertainty typically comes from imprecise automatic processes. We first review the literature on modeling uncertain data, starting with well-studied relational models and moving then to their semistructured counterparts. We focus on a specific probabilistic XML model, which allows representing arbitrary finite distributions of XML documents, and has been extended to also allow continuous distributions of data values. We summarize previous work on querying this uncertain data model and show how to apply the corresponding techniques to several data mining tasks, exemplified through use cases on two running examples.

INTRODUCTION

Though traditional database applications, for instance, bank account management, have no room for uncertainty, more recent applications, such as information extraction from the Web, automatic schema matching in information integration, or information gathering from sensor networks are inherently imprecise. This uncertainty is sometimes represented as the *probability* that the data is correct, as with conditional random fields in information extraction (Lafferty, McCallum, & Pereira, 2001), or uncertain schema mappings in information integration (Dong, Halevy, & Yu, 2009). In other cases, only *confidence* in the information is provided by the system, which can be seen after renormalization as an approximation of the probability. More rarely, some applications do not provide any form of preference among possible uncertain choices (think, for example, of missing data in a data recovery application), or only some unweighted preferences (like the core solution in data exchange (Fagin, Kolaitis,

DOI: 10.4018/978-1-4666-2455-9.ch034

& Popa, 2005) or a minimal repair in managing inconsistent databases (Chomicki & Libkin, 2000; Lopatenko & Bertossi, 2007)).

Usually, data uncertainty is not formally taken into account: only the most likely interpretation is kept for future processing, or all probable choices above a threshold are maintained. We claim this is not sufficient. There is a need for managing the imprecision in this data more rigorously. The need is even stronger when the uncertain data is manipulated by other systems, potentially uncertain themselves. A good example of that is data mining. Consider a scenario where some dataset (say, a list of emails) was acquired, cleaned, and enriched, by a variety of systems (information extraction, deduplication, data integration, natural language analysis, sentiment analysis, etc.). We now want to mine this dataset, for instance to construct from it a list of popular keywords, or to build a social network of individuals, where the friendship links between two persons is derived from their recorded interactions. An application that would make use of the inherent uncertainty in the dataset would be able to discover much more knowledge than one that would ignore it altogether. Besides, in the mining task the confidence annotation in the data could also be used to derive the confidence of the resulting (mined) data.

A number of models and systems for managing uncertain data have been proposed in the literature and a high-level picture of some of them is presented in this chapter. We focus, however, on the particular case of XML data, adapted in the cases where the information is either not strictly constrained by a schema (e.g., Web data), or inherently tree-like (mailing lists, parse trees of natural language sentences, etc.). We also mostly discuss probabilistic models, which have the advantage, in addition to being suited to a number of tasks that provide probability or probability-like confidence scores, of allowing extensive mathematical manipulations (more so than models based on fuzzy logic (Galindo, Urrutia, & Piattini, 2006), that are not discussed in this chapter).

The objective of our chapter is thus to bridge the studies on uncertain XML and data mining. On the one hand, we want to introduce different models of uncertain data to the data mining community. On the other hand, we want to study different data mining tasks for probabilistic XML. Recent studies of probabilistic XML (Abiteboul, Kimelfeld, Sagiv, & Senellart, 2009; Kimelfeld, Kosharovsky, & Sagiv, 2009; Kharlamov, Nutt, & Senellart, 2010) focus on query answering and updates, but mining, that has been studied in the context of relational probabilistic data (Aggarwal, 2009; Bernecker, Kriegel, Renz, Verhein, & Züfle, 2009), has not received attention in the semistructured case. Note that the change of representation format from tables to trees also makes data mining tasks different (Nayak, 2005). In this chapter we propose methods for mining probabilistic XML data (frequent items, correlations, summaries of data values, etc.) that rely on the existing literature on probabilistic XML querying (Kimelfeld et al., 2009; Abiteboul, Chan, Kharlamov, Nutt, & Senellart, 2010).

In the following part of this chapter we discuss several main approaches to uncertainty modeling. We start with uncertain relational databases and present examples and intuitions of incomplete and probabilistic tables. We discuss how these approaches were adapted to the semistructured setting and illustrate incomplete XML trees and two probabilistic XML models: with local and global probabilistic relationships. The next section is devoted to a formal presentation of these probabilistic XML models; we present the syntax and semantics of discrete and continuous probabilistic XML. We then summarize known results about probabilistic XML querying, both for Boolean and aggregate queries, that are at the basis of the data mining approaches we present in a subsequent section, where we give examples and develop computation techniques for mining frequent, co-occurring, or popular items, or for summarizing continuous distributions.