Computers and Mathematics with Applications 62 (2011) 2755–2769

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

P-top-k queries in a probabilistic framework from information extraction models

Ming He [*], Yong-ping Du

College of Computer Science, Beijing University of Technology, Beijing, China

ARTICLE INFO

Keywords:
Probabilistic databases
Uncertain data
Information extraction
Conditional random fields

ABSTRACT

Many applications today need to manage uncertain data, such as information extraction (IE), data integration, sensor RFID networks, and scientific experiments. Top-k queries are often natural and useful in analyzing uncertain data in those applications. In this paper, we study the problem of answering top-k queries in a probabilistic framework from a state-of-the-art statistical IE model—semi-conditional random fields (CRFs)—in the setting of probabilistic databases that treat statistical models as first-class data objects. We investigate the problem of ranking the answers to probabilistic database queries. We present an efficient algorithm for finding the best approximating parameters in such a framework for efficiently retrieving the top-k ranked results. An empirical study using real data sets demonstrates the effectiveness of probabilistic top-k queries and the efficiency of our method.

1. Introduction

In recent years, uncertain data management has arisen in many applications. The reasons for uncertainty in data are as various as the applications themselves: in sensor and RFID data, uncertainty arises as a result of measurement errors [1,2]; in information extraction, uncertainty comes from the inherent ambiguity in natural-language text [3,4]; and in business intelligence, uncertainty is used to decrease the cost of data cleaning [5]. In some applications, such as privacy, it is a special requirement that the data be less precise. For example, uncertainty is intentionally inserted to hide sensitive attributes of individuals so that the data may be published [6–8]. In other cases, the data points may correspond to objects which are only vaguely specified, and are therefore considered uncertain in their representation. Similarly, surveys and imputation techniques create sets of data which are uncertain in nature. This has created a need for uncertain data management algorithms and applications [9,10]. The field of uncertain data management poses a number of unique challenges on several fronts. The two broad issues are those of modeling the uncertain data, and, then, leveraging it to work with a variety of applications. A number of issues and working models for uncertain data have been discussed in [9,11]. The second issue is that of adapting data management and mining applications to work with the uncertain data. The main areas of research in the field are including modeling of uncertain data, uncertain data management and uncertain data mining [10]. Unfortunately, traditional precise database management systems (DBMSs) do not support uncertain data due to data records being typically represented by probability distributions rather than deterministic values. Hence, it is necessary to develop data management techniques for managing probabilistic data.

A probabilistic database, or PDB, is a system that stores large volumes of probabilistic data and supports complex queries. A PDB may also need to perform some additional tasks, such as updates or recovery, but these do not differ from those in

[*] Correspondence to: Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore 639798, Singapore. Tel.: +65 6790 6862; fax: +65 6793 3318.

E-mail addresses: heming@bjut.edu.cn (M. He), ypdu@bjut.edu.cn (Y.-p. Du).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.06.064

conventional database management systems and will not be discussed here. The major challenge in a PDB is that it needs both to scale to large data volumes, a core competence of database management systems, and to do probabilistic inference, which is a problem researched in AI. While many scalable data management systems exist, probabilistic inference is in general a hard problem [12], and current systems do not scale to the same extent as data management systems do. To address this challenge, researchers have focused on the specific nature of relational probabilistic data, and exploited the special form of probabilistic inference that occurs during query evaluation. A number of such results have emerged recently: lineage-based representations [13], safe plans [14], algorithms for top-k queries [15–20] (also known as ranking queries), and representations of views over probabilistic data [21,22]. What is common to all these results is that they apply and extend well known concepts that are fundamental to data management, such as the separation of the query and data when analyzing complexity [23], incomplete databases [24], the threshold algorithm [25], and the use of materialized views to answer queries [26,27].

The goal of information extraction is to extract structured data from a collection of unstructured text documents. Usually the schema is given in advance by the user, and the extractor is tailored to that specific schema [28]. In information extraction, imprecision comes from the inherent ambiguity in natural-language text. Automatically extracting structured entities from unstructured text is a challenging problem, and has a history of attempts spanning early rule-based systems like Rapier [29] to the later statistical methods like that of conditional random fields (CRFs) [3]. Gupta and Sarawagi [30] recently described how to use a probabilistic database to store the result of text segmentation with conditional random fields.

Given this background, it is natural to consider constructing a unified database system that enables well-specified information extraction tasks, and provides a probabilistic framework for top-k queries. This is especially natural for leading information extraction approaches like that of CRFs that are themselves probabilistic machine learning methods. The query language of the PDB should be able to capture the models and methods inherent in these probabilistic information extraction techniques.

Inspired by that work, in this paper we introduce a state-of-the-art method called the semi-CRF method for information extraction and provide a probabilistic framework enabling opportunities for top-k queries. Our technique is based on two major observations. First, all approaches to information extraction are imprecise, and most often can associate a probability score with the item extracted. In the database community, information extraction has been a major motivating application for the recent groundswell of work on PDB, which can model the uncertainty inherent in information extraction outputs, and enable users to get probabilistic answers. Second, top-k queries are often natural and useful in analyzing uncertain data, and CRFs can be very naturally modeled as first-class data in a relational database, in the spirit of recent PDB like BayesStore [31] and the work of Sen and Deshpande [32]. Similarly, text data can be captured relationally via the inverted file representation commonly used in information retrieval.

Our approach in this paper is to process the score and uncertainty in one framework, leveraging current probabilistic database storage and query processing capabilities. Our contributions are summarized as follows:

- New query definitions: top-k processing on uncertain data has, we believe, grown in importance in a large number of real-world applications; we propose new formulations for P-top-k queries in uncertain databases.
- We study an alternative model for representing imprecision in a database that captures the original distribution.
- The processing framework: we construct a framework integrating a state-of-the-art statistical model—semi-conditional random fields (CRFs)—and data access methods bridging the gap between a semi-CRF-based information extraction model and a probabilistic database.
- Experimental study: we conduct an extensive experimental study to evaluate our techniques under different data sets.

This paper is organized as follows. In Section 2, we will examine the issue of uncertain data representation and modeling. In Section 3, we describe our problem and present a background of state-of-the-art probabilistic models for information extraction. In Section 4, we describe the design of a probabilistic database that can support probabilistic information extraction models, and a framework is presented for inference queries. In Section 5 we present experimental evaluation of the accuracy and efficiency of our model. Related work and conclusions appear in Sections 6 and 7 respectively.

2. Uncertain data representation and modeling

The problem of modeling uncertain data has been studied extensively in the literature [33–37]. A database that provides incomplete information consists of a set of possible instances of the database. It is important to distinguish between incomplete databases and probabilistic data, since the latter provide a more specific definition which creates database models with crisp probabilistic quantification.

2.1. Probabilistic database definitions

A probabilistic database is defined [34] as follows:

Definition 1. A *probabilistic information database* is a finite probability space whose outcomes are all possible database instances consistent with a given schema. This can be represented as the pair (χ, p), where χ is a finite set of possible database instances consistent with a given schema, and $p(I)$ is the probability associated with any instance $I \in \chi$. We