Chapter 9

PROBABILISTIC JOIN QUERIES IN UNCERTAIN DATABASES

A Survey of Join Methods for Uncertain Data

Hans-Peter Kriegel, Thomas Bernecker, Matthias Renz and Andreas Zuefle
Ludwig-Maximilians University Munich
Oettingenstr. 67, 80538 Munich, Germany
{kriegel,bernecker,renz,zuefle}@dbs.lmu.de

Abstract
The join query is a very important database primitive. It combines two datasets \( \mathcal{R} \) and \( \mathcal{S} \) (\( \mathcal{R} = \mathcal{S} \) in case of a self-join) based on some query predicate into one set such that the new set contains pairs of objects of the two original sets. In various application areas, e.g. sensor databases, location-based services or face recognition systems, joins have to be computed based on vague and uncertain data. As a consequence, in the recent decade a lot of approaches that address the management and efficient query processing of uncertain data have been published. They mainly differ in the representation of the uncertain data, the distance measures or other types of object comparisons, the types of queries, the query predicates and the representation of the result. Only a few approaches directly address join queries on uncertain data. This chapter gives an overview of probabilistic join approaches. First, it surveys the categories that occur in general queries on uncertain data and secondly, it exemplarily sketches some join approaches on uncertain data from different categories.

Keywords: probabilistic query processing, uncertainty models, similarity join, spatial join

1. Introduction

In many modern application ranges, e.g. spatio-temporal query processing of moving objects [20], sensor databases [19] or personal identification systems [57], usually only uncertain data is available. For instance, in the area of mobile and location-based services, the objects continuously change their positions such that information about the exact location is almost impossible
to obtain. An example of a location-based service and in particular of a spatial join is to notify moving people on their cell-phone if one of their friends enters their vicinity. In the area of multimedia databases, e.g. image or music databases, or in the area of personal identification systems based on face recognition and fingerprint analysis, there often occurs the problem that a feature value cannot exactly be determined. This uncertain data can be handled by assigning confidence intervals to the feature values, or by specifying probability density functions indicating the likelihoods of certain feature values.

A join query combines two datasets \( R \) and \( S \) (\( R = S \) in case of a self-join) based on some query predicate into one set such that the new set contains pairs of objects of the two original sets. Formally,

**Definition 9.1 (Join Query)** Given two relations \( R \) and \( S \) and a predicate \( \theta : R \times S \rightarrow \{true, false\} \). A join query on \( R \) and \( S \) is defined as follows:

\[
R \bowtie \theta S = \{(r, s) \in R \times S | \theta(r, s) = true\}.
\]

In order to join uncertain objects by traditional join methods, a non-uncertain result of the join predicate is required. However, if a query predicate is applied to uncertain attributes of uncertain objects, usually no unique answer whether the query predicate is fulfilled can be given. In this case, the vague information has to be aggregated in order to make the join predicate evaluable. Obviously, aggregation goes hand in hand with information loss. For instance, we have no information about how uncertain the similarity between two uncertain objects is. Even if we had one, it would be of no use because traditional join algorithms cannot handle this additional information.

This chapter gives an overview of probabilistic join approaches. They mainly differ in the representation of the uncertain data, the distance measure or other types of object comparisons, the types of queries, the query predicates and the representation of the result. First, the following section (Section 2) gives a rough overview of traditional join methods originally defined for non-uncertain data, but which form an important foundation for several join approaches defined for uncertain data. Section 3 surveys different uncertainty models and shows how existing join queries (and queries in general) on uncertain data can be categorized. Section 4 exemplarily sketches existing probabilistic join approaches on uncertain data which are representatives of different probabilistic join categories.

2. **Traditional Join Approaches**

A variety of different algorithms have been proposed for joining relations in the traditional case where the data is not uncertain. The following section gives a brief overview over existing approaches and serves as a base for the generalization to the case of uncertain data.