**Query Aware Determinization of Uncertain Objects**

**ABSTRACT**

This paper considers the problem of determinizing probabilistic data to enable such data to be stored in legacy systems that accept only deterministic input. Probabilistic data may be generated by automated data analysis/enrichment techniques such as entity resolution, information extraction, and speech processing. The legacy system may correspond to pre-existing web applications such as Flicker, Picasa, etc. The goal is to generate a deterministic representation of probabilistic data that optimizes the quality of the end-application built on deterministic data. We explore such a determinization problem in the context of two different data processing tasks—triggers and selection queries. We show that approaches such as thresholding or top-1 selection traditionally used for determinization lead to suboptimal performance for such applications. Instead, we develop a query-aware strategy and show its advantages over existing solutions through a comprehensive empirical evaluation over real and synthetic datasets.

**EXISTING SYSTEM**

In existing system, User data is generated automatically through a variety of signal processing, data analysis/enrichment techniques before being stored in the web applications. Pushing such data into web applications introduces a challenge since such automatically generated content is often ambiguous and may result in objects with probabilistic attributes. Such probabilistic data must be “determinized" before being stored in legacy web applications. The problem of mapping probabilistic data into the corresponding deterministic representation as the *determinization* problem.

**DISADVANTAGE**

* The problem of determinization has not been explored extensively in the past.
* The main drawback of this approach is that it is *unaware* of the query workload and thus does not necessarily optimize the given quality metrics, which leads to lower quality.

**PROPOSED SYSTEM**

We introduce the problem of *determinizing* probabilistic data. Given a workload of triggers/queries, the main challenge is to find the deterministic representation of the data which would optimize certain quality metrics of the answer to these triggers/queries. We propose a framework that solves the problem of determinization by minimizing the expected cost of the answer to queries. We develop a branch and- bound algorithm that finds an approximate near-optimal solution

We address the problem of determinizing a collection of objects to optimize set-based quality metrics, such as F-measure. We develop an efficient algorithm that reaches near-optimal quality. We extend the solutions to handle a data model where mutual exclusion exists among tags. We show that correlations among tags can be leveraged in our solutions to get better results. We also demonstrate that our solutions are designed to handle various types of queries.

**ADVANTAGE**

* the proposed algorithms are very efficient
* high-quality results that are very close to those of the optimal solution

**ARCHITECTURE**

