Technofist

Contact no: 9008001602

Crowdsourcing for Top-K Query Processing over Uncertain Data

Technofist,

YES Complex, 19/3&4, 2nd Floor, Dinnur Main Road, R.T.Nagar,Bangalore-560032

Ph:080-40969981, Website:www.technofist.com. E-mail:technofist.projects@gmail.com

ABSTRACT:

Querying uncertain data has become a prominent application due to the proliferation of user-generated content from socialmedia and of data streams from sensors. When data ambiguity cannot be reduced algorithmically, crowdsourcing proves a viableapproach, which consists of posting tasks to humans and harnessing their judgment for improving the confidence about data values orrelationships. This paper tackles the problem of processing top-K queries over uncertain data with the help of crowdsourcing for quicklyconverging to the realordering of relevant results. Several offline and online approaches for addressing questions to a crowd aredefined and contrasted on both synthetic and real data sets, with the aim of minimizing the crowd interactions necessary to find therealordering of the result set.

INTRODUCTION

Both social media and sensing infrastructures areproducing an unprecedented mass of data that are at the base of numerous applications in such fields as information retrieval, data integration, location based services, monitoring and surveillance, predictive modeling of natural and economic phenomena, public health, and more. The common characteristic of both sensor data and user-generated content is their uncertain nature, due to either the noise inherent in sensors or the imprecision of human contributions. Therefore query processing over uncertain data has become an active research field [47], where solutions are being sought for coping with the two main uncertainty factors inherent in this class of applications: the approximate nature of users' information needs and the uncertainty residing in the queried data In the well-known class of applications commonly referred to as "top-K queries" [26], the objective isto find the best K objects matching the user's information need, formulated as a scoring function over the objects' attribute values. If both the data and the scoring function are deterministic, the best K objects can be univocally determined and totally ordered so as to produce a single ranked result set (as long

as ties are broken by some deterministic rule). ties are broken by some deterministic rule). However, in application scenarios involving uncertain data and fuzzy information needs, this doesnot hold. For example, in a large social network the importance of a given user may be computed as a fuzzy mixture of several characteristics, such as her network centrality, level of activity, expertise, and topical affinity.

A viral marketing campaign may try to identify the "best" K users and exploit their prominence to spread the popularity of a product . Another instance occurs when sorting videos for recency or popularity in a video sharing site for example, the video timestamps may be uncertain because the files were annotated at a coarse granularity level (e.g., the day), or perhaps because similar but not identical types of annotations are available (e.g., upload instead of creation time). Sometimes, data processing may also be a source of uncertainty; for example, when tagging images with a visual quality or representativeness index, the score may be algorithmically computed as a probability distribution, with a spread related to the confidence of the algorithm employed to estimate quality.

.