International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

Crowdsourcing For Top-K Query Processing Over Uncertain Data Base Paper

Sushma Peruwala ; Bollikunta,Warangal ; V.Janaki
[1]M.Tech,Department of CSE,Vaagdevi College of Engineering,Telangana,
Mail id :sushma.peruvala@gmail.com
[2]Asssoc.prof. k Pavani Department of CSE,Vaagdevi College of Engineering,Bollikunta,
Warangal,Telangana,
[3] HOD prof. Department of CSE,Vaagdevi College of Engineering,Bollikunta
Warangal,Telangana.

ABSTRACT

Uncertain data are inherent in some important application. Although a considerable amount of research has been dedicated to modeling uncertain data and answering some types of queries on uncertain data, how to conduct advanced analysis on uncertain data remains an open problem at large. Crowdsourcing has emerged as an effective way to perform tasks that are easy for humans but remain difficult for computers. When data uncertainty cannot be reduced algorithmically, crowdsourcing proves a feasible method, which consists in posting tasks to humans and attaching their decision for improving the confidence about data values or relations. This project tackles the problem of processing Top-K queries over uncertain data with the help of crowdsourcing for quickly converging to the real ordering of relevant results. Several offline and online approaches for addressing questions to a crowd are defined and distinguished on both artificial and real data sets, with the purpose of minimizing the crowd relations necessary to find the real ordering of the result set. Keywords—User/Machine Systems, Query processing.

I. Introduction

In recent years, crowdsourcing techniques have attracted a lot of attention due to their effectiveness in real-life applications. They tackle the tasks including image tagging, decision making and natural language processing, which are hard for computers, but relatively easy for human workers. Some successful crowdsourcing examples for question answering tasks that appear include Quora, Yahoo! Answer and Stack Overflow, where users submit questions and get answers from the

crowd. Using crowdsourcing techniques, these tasks can be solved well by human workers. Despite the success of crowdsourcing techniques, the crowd-selection still remains challenging. Earlier approaches usually focus on the problem of crowd-selection for simple tasks where the selection procedure is based on the trustworthiness of workers. All aspects of query processing over uncertain data, and in particular its complexity and existing techniques, highly depend on data representation. Since it is prohibitively expensive to explicitly represent the extremely large set of all possible worlds of a probabilistic database, one has to settle for succinct data representations [1].One common way to identify the top-k objects is scoring all objects based on some scoring function. An object score acts as a valuation for that object according to its characteristics (e.g., price and size of house objects in a real estate database, or color and texture of images in a multimedia database). Data objects are usually evaluated by multiple scoring predicates that contribute to the total object score. A scoring function is therefore usually defined as an aggregation over partial scores. Top-k processing connects to many database research areas including query optimization, indexing methods and query languages. As a consequence, the impact of efficient top-k processing is becoming evident in an increasing number of applications. The objective is to 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). The goal is to define and compare task selection policies for uncertainty reduction via crowdsourcing, with emphasis on the case of Top-K queries. Given a data set with uncertain values, the objective is to pose to a crowd the set of questions that, within an allowed budget, minimizes the expected residual uncertainty of the result, possibly leading to a unique ordering of the top K results.

II. Related work

Both social media and sensing infrastructures are producing unprecendented mass of data that is at the base of many applications in many fields like information retreival, data integration, location based services, monitoring and surveillance, predictive modeling of natural and economic phenomenon, public health and more. The common characteristic of sensor data and user generated content is their uncertain nature due to either the noise inherent in the sensor or

the imprecisions of human contributions. Therefore query processing has become an active research field where solutions are being developed for coping with two main uncertainty factors inherent in the class of applications: the approximate nature of users' information needs and the uncertainty residing in the queried data. Many systems have been developed for processing the data and many advanced systems are being developed. We will discuss some previous systems which were developed for data processing. In 2016, Eleonora Ciceri ET AL [1] proposed a system, ‖Crowdsourcing for Top-K Query Processing over Uncertain Data‖. They proposed that for query processing using crowdsourcing technique. They have developed this system for processing top-k queries over uncertain data. Their objective was that a data set with uncertain value is posed to a crowd then find out the set of questions that will minimize the expected residual uncertainty of the result, possibly leading to a unique ordering of the top K results. In 2013, Nilesh Dalvi ET AL [5] proposed a system, ―Aggregating Crowdsourced Binary Ratings―. They formulated a matrix completion problem and presented two eigenvectors based algorithms that have guaranteed error bounds when the resulting user–user rating graph satisfies expansion properties. In 2014, Susan Davidson ET AL [2] proposed a work ―Top-k and Clustering with Noisy Comparisons―. They proposed a Bayesian model of how the workers can approach clustering and error model motivated by human behavior which we call the variable error model. In 2014, Ju Fan ET AL [10] proposed a system, ―A Hybrid Machine-Crowdsourcing System for Matching Web Tables‖. They proposed an online scheme that crowdsources questions in instead of crowdsourcing all questions in one single phase. In the existing system Querying uncertain data has developed a bulging application due to the production of user-generated content from social media and of data streams from sensors. When data uncertainty cannot be reduced algorithmically, to 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 to produce a single ranked result set. The Proposed system`s aim to define and compare task selection policies for uncertainty reduction via crowdsourcing, with emphasis on the case of Top-K queries. Given a data set with uncertain values, the objective is to pose to a crowd the set of questions that, minimizes the expected residual uncertainty of the result,

possibly leading to a unique ordering of the top K results.

III. Methodology

The task selection policies are defined and compared for uncertainty reduction via crowdsourcing, with emphasis on the case of Top-K queries. Given a data set with uncertain values, our objective is to pose to a crowd the set of questions that, within an allowed budget, minimizes the expected residual uncertainty of the result, possibly leading to a unique ordering of the top K results [1]. A framework is formalized for uncertain Top-K query processing to adapt to its existing techniques for computing the possible orderings, and introduce a procedure for removing unsuitable orderings, given new knowledge on the relative order of the objects. Several measures of uncertainty are defined and contrasted, either agnostic (Entropy) or dependent on the structure of the orderings. The problem of Uncertainty Resolution (UR) is formulated in the context of Top-K query processing over uncertain data with crowd support. The UR problem amounts to identifying the shortest sequence of questions that, when submitted to the crowd, ensures the convergence to a unique, or at least more determinate, sorted result set. It shows that no deterministic algorithm can find the optimal solution for an arbitrary UR problem. Fig: framework for crowdsourcing The user query is preprocessed and loaded into the database. The uncertainty is removed using algorithms. After removing the uncertainy the top-k queries are fired and the data is displayed to the user. The uncertain data is the input to the system and the sorted top-k queries is the output of the system. Two algorithms are used to remove the uncertainty :

1. Top-1 online algorithm

2. Incremental algorithm

1. TOP-1 ONLINE ALGORITHM This algorithm builds the sequence of questions Q* iteratively until the budget B is exhausted (line 2). At each iteration, the algorithm selects the best (Top-1) unasked question, i.e., the one that minimizes the expected residual uncertainty with budget B=1 (line4). The selected question Reserved 75 qi* is then appended to Q* and asked to the crowd. Depending on the answer, the TPO Tk is updated to the subtree that agrees with the answer to qi*(line7). Early termination may occur if all uncertainty is removed, i.e., the tree is left with a single path (line3).

2. INCREMENTAL ALGORITHM The number of orderings in a TPO can be large if there are many overlaps in the tuple score distributions, thereby affecting the execution

time of our algorithms. This algorithm does not receive as input a TPO Tk. Instead, it builds the TPO incrementally, one level at a time, by alternating tree construction with around of n questions and tree pruning. The number n of questions posed at each round is between 1 and B therefore incr can be considered a hybrid between an online and an offline algorithm. Each TPO Tk, $1 \leq k \leq K$,is built by adding one level to the previous TPO Tk-1 (line10), i.e., by attaching to each ordering ωk-1 in Tk-1 the unused sources as leaves. We only build new levels if there are not enough questions to ask (line5), i.e., n questions for all the rounds but the last one, where B mod n questions are asked (line 3). Then, we select the best questions, pose them to the crowd, collect the answers and apply the pruning accordingly (lines6-9), until either the budget B is exhausted or the TPO is entirely built (line2). We thus keep the TPO as pruned as possible, and only proceed to computing the next level when the uncertainty in the previous levels is so low that it does not require n questions to ask.

IV CONCLUSION

In this paper we have introduced Uncertainty Resolution, which is the problem of identifying the minimal set of questions to be submitted to a crowd in order to reduce the uncertainty in the ordering of top-K query results. First of all, we proved that measures of uncertainty that take into

account the structure of the tree in addition to ordering probabilities (i.e., UMPO, UHw and UORA) achieve better performance than state-of-the-art measures (i.e., UH). Moreover, since UR does not admit deterministic optimal algorithms, we have introduced two families of heuristics (offline and online, plus a hybrid thereof) capable of reducing the expected residual uncertainty of the result set. The proposed algorithms have been evaluated experimentally on both synthetic and real data sets, against baselines that select questions either randomly or focusing on tuples with an ambiguous order. The experiments show that offline and online best-first search algorithms achieve the best performance, but are computationally impractical. Conversely, the T1_on and C_off algorithms offer a good tradeoff between costs and performance. With synthetic datasets, both the T1_on and C_off achieve significant reductions of

the number of questions wrt. the Naive algorithm. The proposed algorithms have been shown to work also with nonuniform tuple score distributions and with noisy crowds. Much lower CPU times are possible with the incr algorithm,

with slightly lower quality (which makes incr suited
for large, highly uncertain datasets). These trends are further validated on the real datasets.

.

References

1. ElenoraCiceri, PieroFraternali, DavideMartinenghi and Marco Tagliasacchi, ―Crowdsourcing for Top-K Query Processing over Uncertain Data‖, IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 1, January 2016.

2. NileshDalvi, Troo.ly Inc., ―Query Processing over Uncertain Data‖, Dan Olteanu University of Oxford. International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 04, Issue 10, [October–2017]ISSN (Online):2349–9745; ISSN (Print):2393-8161 @IJMTER-2017, All rights Reserved 77

3. Ihab F. Ilyas, George Beskales and Mohamed A. Soliman, ―A Survey of Top-k Query Processing Techniques in Relational Database Systems‖, David R. Cheriton School of Computer Science University of Waterloo, ACM Journal Name, Vol. V, No. N, Month 20YY, Pages 1–61. 4. J. Li et al., ―A unified approach to ranking in probabilistic databases‖. PVLDB, 2(1):502–513, 2009.

5. N. N. Dalvi et al., ―Aggregating crowdsourced binary ratings‖, In WWW, pages 285–294, 2013.

6. S. B. Davidson et al, ―Top-k and clustering with noisy comparisons‖. ACM Trans. Database System. 39(4): 35:1–35:39, 2014. A. Anagnostopoulos et al., ―The importance of being expert: Efficient max-finding in crowdsourcing‖, In SIGMOD, 2015.

7. J. Fan et al. ―A hybrid machine-crowdsourcing system for matching web tables‖. ICDE, 2014.

8. F. C. Heilbron and J. C. Niebles. ―Collecting and annotating human activities in web videos‖. In ICMR, page 377, 2014. 9. J. Fan et al. ―A hybrid machine-crowdsourcing system for matching web tables‖. ICDE, 2014.