Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

A Survey on Query Processing Over Uncertain Data Using Top-K

Swapna Gangapuram
HoD,
Department of CSE,
Siddhartha Institute of Tech & Sciences,
Swapna.gangapuram@gmail.com.

P.Bhavana
M. Tech Student,
Department of CSE,
Siddhartha Institute of Tech & Sciences,
pagabhavana@gmail.com.

Abstract

Querying uncertain data has become a prominent application due to the proliferation of user-generated content from social media and of data streams from sensors. When data ambiguity cannot be reduced algorithmically, crowd sourcing proves a viable approach, which consists of posting tasks to humans and harnessing their judgment for improving the confidence about data values or relationships. This paper tackles the problem of processing top-K queries over uncertain data with the help of crowd sourcing for quickly converging to the real ordering of relevant results. Several offline and online approaches for addressing questions to a crowd are defined and contrasted on both synthetic and real data sets, with the aim of minimizing the crowd interactions necessary to find the real ordering of the result set.

Index Terms—User/machine systems, query processing

INTRODUCTION

Both social media and sensing infrastructures are producing 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, 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", 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).

However, in application scenarios involving uncertain data and fuzzy information needs, this does not 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.

Furthermore, uncertainty may also derive from the user's information need itself; for example, when ranking

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

apartments for sale, their value depends on the weights assigned to price, size, location, etc., which may be uncertain because they were specified only qualitatively by the user or estimated by a learning-to-rank algorithm. When either the attribute values or the scoring function are nondeterministic, there may be no consensus on a single ordering, but rather a space of possible orderings. For example, a query for the top-K most recent videos may return multiple orderings, namely all those compatible with the uncertainty of the timestamps. To determine the correct ordering, one needs to acquire additional information so as to reduce the amount of uncertainty associated with the queried data. Without this reduction, even moderate amounts of uncertainty make top-K answers become useless, since none of the returned orderings would be clearly preferred to the others.

An emerging trend in data processing is crowd sourcing, defined as the systematic engagement of humans in the resolution of tasks through online distributed work. This approach combines human and automatic computation in order to solve complex problems, and has been applied to a variety of data and query processing tasks, including multimedia content analysis, data cleaning, semantic data integration, and query answering. When data ambiguity can be resolved by human judgment, crowd sourcing becomes a viable tool for converging towards a unique or at least more determinate query result. For example, in an event detection and sorting scenario, a human could know the relative order of occurrence of two events; with this information, one could discard the incompatible orderings. However, crowd sourcing has problems of its own: the output of humans is uncertain, too, and thus additional knowledge must be properly integrated, notably by aggregating the responses of multiple contributors. Due to this redundancy, significant budget savings may be achieved by avoiding to post even a small amount of tasks.

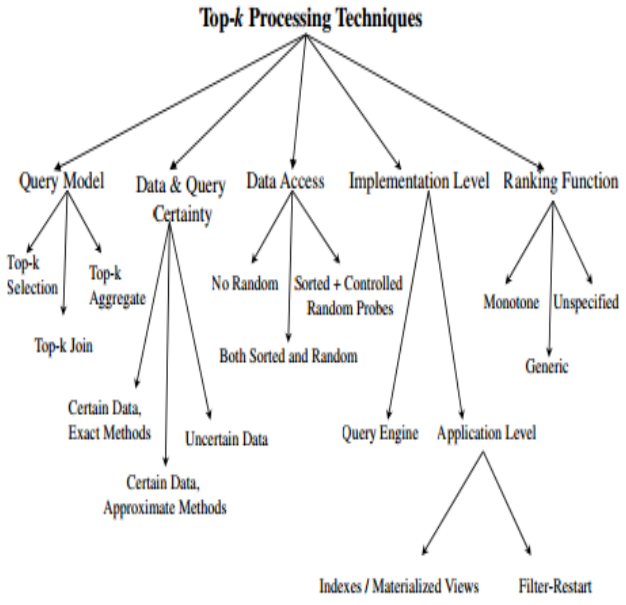
This problem requires an appropriate policy in the formulation of the tasks to submit to the crowd, aimed at reaching the maximum reduction of uncertainty with the smallest number of crowd task executions. The goal of this paper is to define and compare task selection policies for uncertainty reduction via crowd sourcing, 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. The main contributions of the paper are as follows: 1) We formalize a framework for uncertain top-K query processing, adapt to it 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. 2) We define and contrast several measures of uncertainty, either agnostic (Entropy) or dependent on the structure of the orderings. 3) We formulate the problem of Uncertainty Resolution (UR) 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. We show that no deterministic algorithm can find the optimal solution for an arbitrary UR problem. 4) We introduce two families of heuristics for question selection: offline, where all questions are selected prior to interacting with the crowd, and online, where crowd answers and question selection can intermix.

RELATED WORK

Many works in the crowd sourcing area have studied how to exploit a crowd to obtain reliable results in uncertain scenarios. In binary questions are used to label nodes in a directed acyclic graph, showing that an accurate question selection improves upon a random one. Similarly, The aim to reduce the time and budget used for labeling objects in a set by means of an appropriate question selection. Instead, proposes an online question selection approach for finding the next most convenient question so as to identify the highest ranked object in a set. A query language where questions are asked to humans and algorithms is described;

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

humans are assumed to always answer correctly, and thus each question is asked once. All these works do not apply to a top-K setting and cannot be directly compared to our work. Uncertainty in Top-K Queries Uncertainty representation. The problem of ranking tuples in the presence of uncertainty has been addressed in several works. we based our techniques for the construction of a TPO on these works. Uncertain top-K queries on probabilistic databases. In the quality score for an uncertain top-K query on a probabilistic (i.e., uncertain) database is computed. Moreover, the authors address the problem of cleaning uncertainty to improve the quality of the query answer, by collecting multiple times data from the real world (under budget constraints), so as to confirm or refute what is stated in the database. Crowd sourcing via tuples comparison. We now discuss recent works on uncertain top-K scenarios where questions comparing tuples in a set are asked to a crowd. In the authors consider a crowd of noisy workers and tuples whose scores are totally uncertain.



This approach does not lend itself well to our scenarios, where prior knowledge on the score pdf's is assumed: for instance, when N ¼ 1;000, d ¼ 0:001 and workers answer correctly with probability 0:8, their approach would require 999 questions to determine the top-1 tuple, while 2:7 are in average sufficient with our T1 on.

The work proposes a query interface that can be used to post tasks to a crowd sourcing platform such as Amazon MTurk. When addressing a top-K query, their method first disambiguates the order of all the tuples by asking questions to the crowd, and then extracts the top-K items. This amounts to asking many questions that are irrelevant for the top-K prefix, since they could involve tuples that are ranked in lower positions. The wasted effort grows exponentially as the dataset cardinality grows. Instead, our work only considers questions that involve tuples comprised in the first K levels of the tree.

A more recent work in builds the top-K list by asking workers to sort small sets of s tuples whose scores are, again, totally uncertain. The top-K tuples are determined via a voting mechanism that refines the set of top-K candidates after each "roundtrip" of tasks, until only K tuples are left. Although when s ¼ 2 the tasks are a comparison of two tuples like in our approach, their question selection is completely agnostic of any prior knowledge on the tuples, thus resulting in a much higher overall amount of questions in scenarios like those considered in this paper. Uncertain top-K sets. In the authors propose procedures for the extraction of k objects that have a specified property.

The proposed algorithms extract sets of n objects that are analyzed in parallel by humans. At each round, n tasks are submitted to the crowd, and the objects that are recognized as relevant (i.e., objects having the specified property) are retrieved. Rounds are continuously created, until exactly K objects are retrieved. The work considers both the cases of oracles and of noisy workers. However, this work only takes care of extracting a set of objects, which remain unordered, with no guarantee to include the top K objects. Crowd sourcing via other task types.

In the authors assume that the ordering of a set of objects is known, and use a crowd sourcing-based algorithm to estimate their score values. In crowd sourcing is used to build a tree where the root represents an initial status, leaves represent a fixed objective and each path represents a sequence of actions to be performed so as to

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

meet the objective. Workers are provided with a sub-path and are required to suggest the next action in the sequence to be performed. The goal of the proposed algorithm is to retrieve the best K paths from the tree.

Uncertainty in Schema and Object Matching Schema matching. In uncertainty in schema matching is tackled by posing questions to workers. Uncertainty is measured via entropy, and two algorithms (online and offline) are proposed to select the questions reducing uncertainty the most. A similar approach is proposed in for the context of web tables schema matching, although only an online scenario is considered in this case.

We have shown that, in top-K contexts, the results obtained by measuring uncertainty via entropy are largely outperformed by the use of other criteria (e.g., UMPO). Noisy workers are used to validate schema matching's also, with emphasis on the design of questions, so as to maximize their in formativeness and reduce the noise in validations. Yet, does not present any question selection strategy, which we have shown to be a useful means to obtain good results even with a noisy crowd and simple boolean questions

EXISTING SYSTEM

- Query processing over uncertain data has become an active research field, 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 existing system, the quality score for an uncertain top-K query on a probabilistic (i.e., uncertain) database is computed. Moreover, the authors address the problem of cleaning uncertainty to improve the quality of the query answer, by collecting multiple times data from the real world (under budget constraints), so as to confirm or refute what is stated in the database.

DISADVANTAGES:

- The output of humans is uncertain, too, and thus additional knowledge must be properly integrated, notably by aggregating the responses of multiple contributors.
- These amounts to asking many questions that are irrelevant for the top-K prefix, since they could involve tuples that are ranked in lower positions.
- The wasted effort grows exponentially as the dataset cardinality grows.

PROPOSED SYSTEM:

- The goal of this paper 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, 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.

The main contributions of the paper are as follows:

- We formalize a framework for uncertain top-K query processing, adapt to it 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.
- We define and contrast several measures of uncertainty, either agnostic (Entropy) or dependent on the structure of the orderings.
- We formulate the problem of Uncertainty Resolution (UR) 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.
- We introduce two families of heuristics for question selection: offline, where all questions

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

are selected prior to interacting with the crowd, and online, where crowd answers and question selection can intermix.

- For the offline case we define a relaxed, probabilistic version of optimality, and exhibit an algorithm that attains it as well as sub-optimal but faster algorithms. We also generalize the algorithms to the case of answers collected from noisy workers.

ADVANTAGES OF PROPOSED SYSTEM:

- We show that no deterministic algorithm can find the optimal solution for an arbitrary UR problem.
- We propose an algorithm that avoids the materialization of the entire space of possible orderings to achieve even faster results.
- We conduct an extensive experimental evaluation of several algorithms on both synthetic and real datasets, and with a real crowd, in order to assess their performance and scalability.

CONCLUSIONS AND FUTURE WORK

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 T1on and Coff algorithms offer a good tradeoff between costs and performance. With synthetic datasets, both the T1on and Coff achieve significant reductions of the number of questions wrt. the Naive algorithm. The proposed algorithms have been shown to work also with non uniform 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. Future work will focus on generalizing the UR problem and heuristics to other uncertain data and queries, for example in skill-based expert search, where queries are desired skills and results contain sequences of people sorted based on their topical expertise and skills can be endorsed by community peers .

REFERENCES

[1] M. Allahbakhsh, et al., "Quality control in crowdsourcing systems: Issues and directions," IEEE Internet Comput., vol. 17, no. 2, pp. 76–81, Mar./Apr. 2013.

[2] A. Amarilli, et al., "Uncertainty in crowd data sourcing under structural constraints," in Proc. 19th Int. Conf. Database Syst. Adv. Appl., 2014, pp. 351–359.

[3] A. Anagnostopoulos, et al., "The importance of being expert: Effi- cient max-finding in crowdsourcing," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2015, pp. 983–998.

[4] M. Cha, et al., "Analyzing the video popularity characteristics of large-scale user generated content systems," IEEE/ACM Trans. Netw., vol. 17, no. 5, pp. 1357–1370, Oct. 2009.

[5] R. Cheng, et al., "Efficient join processing over uncertain data," in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage., 2006, pp. 738–747.

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

[6] N. N. Dalvi, et al., "Aggregating crowdsourced binary ratings," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 285–294.

[7] A. Das Sarma, et al., "Crowd-powered find algorithms," in Proc. Int. Conf. Data Eng., 2014, pp. 964–975.

[8] S. B. Davidson, et al., "Top-k and clustering with noisy comparisons," ACM Trans. Database Syst., vol. 39, no. 4, pp. 35:1– 35:39, 2014.

[9] J. Fan, et al., "A hybrid machine-crowdsourcing system for matching web tables," in Proc. IEEE 30th Int. Conf. Data Eng., 2014, pp. 976–987.

[10] C. Gokhale, et al., "Corleone: Hands-off crowdsourcing for entity matching," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2014, pp. 601–612.

[11] S. Guo, et al., "So who won?: Dynamic max discovery with the crowd," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 385–396.

[12] P. Hart, et al., "A formal basis for the heuristic determination of minimum cost paths," IEEE Trans. Syst., Sci., Cybern., vol. SSC-4, no. 2, pp. 100–107, Jul. 1968.

[13] F. C. Heilbron and J. C. Niebles, "Collecting and annotating human activities in web videos," in Proc. Int. Conf. Multimedia Retrieval, 2014, p. 377.

[14] N. Hung, et al., "On leveraging crowdsourcing techniques for schema matching networks," in Proc. Int. Conf. Database Syst. Adv. Appl., 2013, pp. 139–154.

[15] P. G. Ipeirotis, et al., "Quality management on amazon mechanical turk," in Proc. SIGKDD Workshop Human Comput., 2010, pp. 64–67.

[16] P. G. Ipeirotis and E. Gabrilovich, "Quizz: Targeted crowdsourcing with a billion (potential) users," in Proc. 23rd Int. Conf. World Wide Web, 2014, pp. 143–154.

[17] K. J€arvelin and J. Kek€al€ainen, "Cumulated gain-based evaluation of ir techniques," ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, 2002.

[18] M. Joglekar, et al., "Comprehensive and reliable crowd assessment algorithms," in Proc. Int. Conf. Data Eng., 2015.

[19] H. Kaplan, I. Lotosh, T. Milo, and S. Novgorodov, "Answering planning queries with the crowd," Proc. VLDB Endowment, vol. 6, no. 9, pp. 697–708, 2013.