Make Informed Decisions: Understanding Query Results from Incomplete Databases

Poonam Kumari
Supervised by Dr. Oliver Kennedy
State University of New York at Buffalo
New York, United States of America
poonamku@buffalo.edu

ABSTRACT

Analyzing data has been central in making decisions whether it be a decision to buy stock or detect the chances of diabetes based on family history. Datasets used for analysis might include incomplete, inconsistent, missing data or might involve integrating two or more sources. Data quality management has been studied extensively with focus on tabular data. Lot of work has been done in terms of data curation and imputation, although visualization aspect of data quality management remains fairly unexplored. The aim of this PhD research is to focus on visualizing the imperfections in these datasets in order to help users analyze and interpret data and guide them to make informed decisions. We explore how different visualization techniques affect perceived data quality, accuracy and decision confidence.

1. INTRODUCTION

With growing data sizes and different ways of obtaining data, datasets being analyzed are prone to incomplete, inconsistent, missing data etc. These errors must be detected and corrected in order to maintain the quality and usability of data. This takes up to 30-80 percent of an analyst's time and resources [14]. Different systems have been designed to help analyst curate the data using a wide variety of methods to deal with dirty data.

For example, simple imputation techniques like hot-deck imputation substitute values from current sample whereas cold-deck imputation make use of related datasets [8] or domain heuristics [10]. Methods like linear interpolation, regression, and adaptive interpolation [7] infer missing values by using a weighted combination of available data. More complex imputation techniques estimate missing values using machine learning and related techniques [3] or [13] integrate information about the processes used to generate the dataset.

Historically, uncertain data could not be queried using classical databases. Although incomplete [9] and probabilistic databases [17] let the user query data with uncertainty, the query results might be difficult to understand.

Probabilistic databases: Probabilistic Databases (PD) make use of a user specified probability distribution function for the uncertain data. For instance, in [5] a parameter p on each tuple specifies the probability distribution for tuple existence. In [15] a user specified joint probability is used by PD to determine resulting output tuples and their associated probabilities. No explanation about query results is provided in PD's ("Why a tuple is present in the query result or why is a high probability associated with the tuple?"). Although PD's help manage uncertain data successfully, the probability distributions in query results might be difficult to understand.

Incomplete Databases: To deal with uncertain data, incomplete databases work on a set of all deterministic instances known as possible worlds. A typical query result on these databases might consist of certain answers, possible answers or both (depending on the type of incomplete database system). Database instances in fig 1 represent two possible worlds i.e ceiling mart database is one possible world and Aimpoint is another possible world. The ratings for Dell i7 and Lenovo i7 are consistent across both the possible worlds. If a user issues a query to get the ratings for the two products, the result set would consist of **certain answers** (answers in all possible worlds). Whereas the query result for getting rating of Asus i5 and Lenovo i7 would contain **possible answers** (due to missing value for HP AMD in one possible world and inconsistent rating for Asus i5).

Different approaches have been used to represent possible and certain answers. Conservative approaches [1] consider only the certain answer. For instance a query on possible worlds in fig 1 will result in two tuples Dell i7 and Lenovo i7 since the ratings are consistent in both worlds. Best guess query processing use the best possible world by making an educated guess and work exclusively with guessed world. Suppose best guess approach chooses first instance from fig 1. A query to get the ratings of products will present all certain answers ignoring the missing value and inconsistency in rating for Asus in the second instance. **(1) Conservative approach ignores the uncertainty altogether missing out on valuable information. (2) Best guess approach takes uncertainty into account but the valuable information about interpreting the uncertainty is lost [6].**

To summarize various imputation methods are used to deal with uncertain data which make a guess based on ex-

Proceedings of the VLDB 2019 PhD Workshop, August 26th, 2019. Los Angeles, California.

Ceiling Mart

Name	Rating
Dell i7	4
HP AMD	2
Asus i5	3.5
Lenovo i7	3

Aimpoint

Name	Rating
Dell i7	4
HP AMD	
Asus i5	4.5
Lenovo i7	3

Figure 1: Product rating data from ceiling mart and Aimpoint

isting values, domain heuristics or machine learning techniques. These guesses can be in the form of certain and possible answers. Incomplete and probabilistic databases help query these datasets and provide query results as tabular data. These query results might or might not contain uncertain data (possible answers) which hinders users ability to make an informed decision. Uncertainty annotated databases (UA-DB's [6]) help overcome the limitations of earlier systems caused by ignoring uncertainty or missing out on information about interpreting uncertainty. UA-DB's also help represent uncertain data effectively and distinguish between certain answers and merely possible answers.

2. MOTIVATING EXAMPLE

ABC corp. is a sales company which helps user select products like laptops based on based on a large database of crowd-sourced and/or web-scraped reviews of those products. Alice is the customer service representative. Bob is an analyst who maintains the database. Bob is working on integrating instances shown in fig 1 containing laptop ratings from different vendors. Bob needs to clean the data first (missing value and inconsistent ratings) and load it, which will enable Alice to query the database and make a suggestion to the customer.

During data imputation the system decides to ignore the missing value in case of HP AMD and take an average of ratings in case of Asus i5. Integrated dataset (table 1) is passed on to Alice for analysis.

Name	Rating
Dell i7	4
HP AMD	2
Asus i5	4
Lenovo i7	3

Table 1: Integrated Product rating data from Ceiling Mart and Aimpoint

In the above example table 1 represents an incomplete database. Ratings for Dell i7 and Lenovo i7 are certain answers(answers in all possible worlds) where as ratings for HP AMD and Asus i5 are possible answers (uncertain) due to the system making a guess about the missing and inconsistent value.

3. RESEARCH QUESTIONS

Why is a distinction between certain and uncertain answers required? And how this distinction would help user asses relevant information and make an informed decision based on it?

In the earlier example, Bob had completed the data cleaning task and the database queried by Alice to obtain a tabular query result(table 1) containing both certain and uncertain answers.

- If the conservative approach is used then Alice is left with just two products and the user might choose Dell i7. In this method the user misses out on comparing Dell i7 and Asus i5 which has a higher rating, although uncertain.

- If the best guess approach is considered, Alice has all the 4 ratings to choose from. Since the distinction between certain and possible answers is not clear and valuable information about the possible answer is lost, the user might end up with Asus i5.

A lot of time and effort is put into cleaning the data, making guesses and calculating the best possible world. Data cleaning forms a large chunk in the data management life cycle. After all this effort what if the query results are not understood by the user. For instance, classical probabilistic databases represent query results in the form of certain answers or probability distribution which might overwhelm a naive user. Just having a probability distribution or possible answers for query results is insufficient: *the uncertainty must be communicated to the users who will ultimately decide the relevant information (in the results) pertaining to their task and make an informed decision* [6].

Incomplete databases cannot decide whether the data presented as part of query results are relevant for user's decision task. Alice is helping user make a decision in choosing a laptop based on ratings presented in table 1. Uncertain answers in the query result pose an important question. *Are uncertain answers reliable as they are ultimately a guess made by the system?*. [18] conducted a case study with real world data to demonstrate the usefulness of discovering knowledge about the patterns of missing values through classification. The data mining task was to find how important a role the race factor played in the home loan assessment process. The classifier for the data without the race factor had 64.1% accuracy for the training data set and 64.2% accuracy for the test data set, producing an overall 64.2% accuracy. In the medical domain [20] uses naive credal classifier which extends the discrete naive Bayes classifier to imprecise probabilities. The diagnostic tool delivers upto 95% correct predictions and also proves to be effective in discriminating between Alzheimers disease and dementia with Lewy bodies. Although different imputation methods are used and the system makes an educated guess, the guesses about possible answers are reliable. And excluding possible answers from the query result might result in losing valuable information.

Since the uncertain answers are reliable, *what should the user do when they see an uncertain value?* Users can take the conservative approach and ignore the uncertain values. Limitations of this strategy are well known [6]. Second approach is to consider uncertain answers for decision task. In table 1 the uncertain rating for HP AMD might not be relevant to the user since there are higher rated products. But the uncertain rating in case of Asus might be relevant to the user, since the user has to choose between a certain 4 for Dell i7 and an uncertain 4.5 for Asus i5. The system cannot decide whether the values are relevant to the user task, the user has to understand and make this decision. We believe

that providing additional information about the uncertain data will guide user to make an informed decision.

The focus of this research is **to provide guidelines and best practices to visualize uncertainty in incomplete databases.** For example we would like to **help users to visually distinguish between certain and uncertain answers in query result for incomplete databases.** As another example, simply knowing that an answer is uncertain may not be enough and we would like **to provide additional contextual hints explaining uncertainty.**

4. PRELIMINARY STUDY

Why is there a need to visualize uncertainty? We have already established that presenting possible answers do aid users in making a decision. [16] conducted a pair of crowdsourced studies to measure influence of methods used to impute and visualize missing data on an analysts perception of data quality. The methods used also affected conclusions. The study concluded that highlighting imputed values led to higher perceived confidence, credibility and data quality. Whereas not visualizing the missing values, downplaying visual encodings, filling out missing values with zero (zero-filling) lead to lower subjective perceived measurements.

Apart from improving decision-making and increase in perceived confidence, research carried out in several domains such as health, weather prediction, transportation, and more, indicates displaying uncertainty helps in improving trust placed on the system. A simple feedback mechanism in context-aware systems was evaluated in [4]. The results suggest that human performance in memory-bounded tasks increases by explicitly displaying uncertainty information.

To visually distinguish between certain and uncertain answers in query result for incomplete databases. Most of the imputation methods, require the system to make a guess and form certain and uncertain answers. The type of system decides whether uncertainty in the data should be presented to the user or not. We believe that uncertain data should be presented and uncertainty in the data should be effectively communicated in order to help users interpret the results and decide whether and how to act on the results given. [12] presents our initial efforts in communicating uncertainty about query results in On Demand Curation Tools. A preliminary user study was conducted to evaluate the cognitive burden and expressiveness of four representations of "attribute-level" uncertainty. Uncertain data was annotated using simple one bit representation (asterisk, colored text and color background) and confidence interval (Figure 2).

Product	CeilingMart	Aimpoint	Ibibo
HP	4.5	3.0	3.5±1
Asus	2.5	2.5	3.0
Dell	5.0*	3.5	5.0

Figure 2: **Example uncertainty representations.**

Participants were presented with a task to rank three different products based on the ratings provided. Product selection, re-ordering the product list, and submitting the participant's final order were logged along with timestamps as part of interactions with the web form. Think-aloud protocol was also used in the experiment in order to transcribe participants thought process while making a decision. The study aimed at answering two primary questions: (1) Is the representation *effective* at communicating uncertainty, and (2) What is the *cognitive burden* of interpreting the representation? Results showed an insignificant differences in time taken to interpret uncertainty by the user. And a change in the ways people interpreted and reacted to data based on change in uncertainty was observed. Colored text and color coding significantly altered participant behavior which is consistent with coloring signaling significant errors. Participants requested additional information when asterisk was used to represent uncertain data.

4.1 Follow up Study

Through the previous study we have established the need of representing and ways to represent uncertainty in incomplete databases. The next question is to help user understand the reason for data being uncertain. **To provide additional contextual hints explaining uncertainty**. A follow up study was conducted to explore this task using a lighter-weight, two-level interface for presenting uncertain query results to users. First, a preliminary annotation (same as preliminary study) notifies users about the presence of uncertainty. If they deem it relevant, users can then interactively explore the uncertainty to obtain additional detail. **Why would user need additional information?** One of the limitation of both PD's and incomplete databases is lack of information about the probability/uncertain answer. Output tuples in existing systems like TRIO [2] do contain lineage/provenance along with output probabilities. Lineage refers to a boolean formula which qualitatively explains the reasons for occurrence of the output tuple. However it is not informative in case of multiple output tuples. The case of projection of a million tuples on to a single tuple results in a vary large lineage formula of size one million. This can be difficult for the user to obtain any information from. We believe that information regarding uncertain data can be displayed in the form of small contextual hints. The information should be presented to the user on demand. For instance, in table 1 user might not need this contextual information related to HP AMD laptop, but this additional information might prove helpful in case of choice between Dell and Asus.

5. RESEARCH PLAN

The current systems cannot decide whether an uncertain value is relevant to the user taks (e.g: Ranking task based on data in table 1). [11] talks about the problem of determining the sensitive input tuples for the given query in PD's. Sensitive tuples refer to the one's that can substantially alter output, when their probabilities are modified. Similar strategy can be used in case of incomplete databases. Our next steps would be: **(1) To help the system identify relevancy of uncertain answers to the user query.** E.g. for the ranking task HP rating can be considered irrelevant and Asus rating as relevant. This can be done by identifying input tuples which might affect the output. An algorithm can be developed for identifying such tuples for various known queries like sum, count, min and max. **(2) Incorporate the results from the user study and relevancy algorithm into an existing system Mimir [19].** The findings of preliminary study have been incorporated into Mimir which uses red text to display uncertain answers.

(3) Visualize the effects of user choice on query result Mimir provides feedback on each guessed datapoint and user can choose to approve or fix the datapoint manually. Changes in query result can be visualized based on information from the algorithm in step 1 as the user makes a decision on the feedback provided. This would help the user to visually inspect the effects of their decision before the changes are applied. **(4) Visualize uncertain answers in query results using data plots** Visualization of results will aid data analysis by making it easier for the user to identify trends and outliers in the data. These data plots can be presented to domain experts for further inspection of data requiring domain knowledge. **(5) Visualize missing values in raw data using data plots.** Users can visualize the data and then inspect each data point in raw data by clicking on the data plot and accepting the feedback provided by the system or fixing the uncertainty manually. Similar uncertain data, for e.g. missing values in a column can be fixed in groups based on the feedback provided.

6. CONCLUSION

Increasing data sizes pose the problem of uncertainty in data. Several data curation techniques have been developed along with databases (PD's and incomplete databases) to help query this data. Although data cleaning is studied extensively, we need to focus on visualizing the query results for a better understanding. We described a user study as part of our initial effort and next steps to help us design guidelines for visualizing uncertainty in incomplete databases.

7. REFERENCES

[1] S. Abiteboul, P. Kanellakis, and G. Grahne. On the representation and querying of sets of possible worlds. *Theoretical computer science*, 78(1):159–187, 1991.

[2] C. C. Aggarwal. Trio a system for data uncertainty and lineage. In *Managing and Mining Uncertain Data*, pages 1–35. Springer, 2009.

[3] S. Ahuja, M. Roth, R. Gangadharaiah, P. Schwarz, and R. Bastidas. Using machine learning to accelerate data wrangling. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 343–349. IEEE, 2016.

[4] S. Antifakos, A. Schwaninger, and B. Schiele. Evaluating the effects of displaying uncertainty in context-aware applications. In *International Conference on Ubiquitous Computing*, pages 54–69. Springer, 2004.

[5] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB JournalThe International Journal on Very Large Data Bases*, 16(4):523–544, 2007.

[6] S. Feng, A. Huber, B. Glavic, and O. Kennedy. Uncertainty annotated databases-a lightweight approach for approximating certain answers (extended version). *arXiv preprint arXiv:1904.00234*, 2019.

[7] J. Gao. Adaptive interpolation algorithms for temporal-oriented datasets. In *Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)*, pages 145–151. IEEE, 2006.

[8] W. Githungo, S. Otengi, J. Wakhungu, and E. Masibayi. Infilling monthly rain gauge data gaps with satellite estimates for asal of kenya. *Hydrology*, 3(4):40, 2016.

[9] J. Grant. Incomplete information in a relational database. *FUND. INFO.*, 3(3):363–378, 1980.

[10] K. Gülensoy, C. Gawrilow, and T. von Landesberger. Visual exploration of dirty activity sensor and emotional state data from psychological experiments. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, page 19. Citeseer, 2014.

[11] B. Kanagal, J. Li, and A. Deshpande. Sensitivity analysis and explanations for robust query evaluation in probabilistic databases. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 841–852. ACM, 2011.

[12] P. Kumari, S. Achmiz, and O. Kennedy. Communicating data quality in on-demand curation. *arXiv preprint arXiv:1606.02250*, 2016.

[13] S. Rässler. Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33(1&2):153–171, 2004.

[14] B. Saha and D. Srivastava. Data quality: The other face of big data. In *2014 IEEE 30th International Conference on Data Engineering*, pages 1294–1297. IEEE, 2014.

[15] P. Sen, A. Deshpande, and L. Getoor. Prdb: managing and exploiting rich correlations in probabilistic databases. *The VLDB JournalThe International Journal on Very Large Data Bases*, 18(5):1065–1090, 2009.

[16] H. Song and D. A. Szafir. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics*, 25(1):914–924, 2019.

[17] D. Suciu, D. Olteanu, C. Ré, and C. Koch. Probabilistic databases, synthesis lectures on data management. *Morgan & Claypool*, 2011.

[18] H. Wang and S. Wang. Mining incomplete survey data through classification. *Knowledge and information systems*, 24(2):221–233, 2010.

[19] Y. Yang, N. Meneghetti, R. Fehling, Z. H. Liu, and O. Kennedy. Lenses: An on-demand approach to etl. *Proceedings of the VLDB Endowment*, 8(12):1578–1589, 2015.

[20] M. Zaffalon, K. Wesnes, and O. Petrini. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial intelligence in medicine*, 29(1-2):61–79, 2003.