Missing Value Evaluation in SQL Queries: A Survey

Smruti Mule[1], Antara Bhattacharya[2]

[1]Student, Department Of Computer Science and Engineering ,G.H.Raisoni Institute of Engineering and Technology, Nagpur, India, smrutimule@gmail.com
[2] Assistant Professor, Dept. of Computer Science and Engineering,G.H.Raisoni Institute of Engineering and Technology, Nagpur ,India ,antara.bhattacharya@raisoni.net

---***---

Abstract - *After decades have been passed of taking efforts on performance of database, the usability and quality of database systems have gained more importance in recent years. However, answering to why-not questions i.e evaluating missing answers in SQL Queries after doing a lot of work has also gained more attention. The main goal of this research paper is evaluating missing value in the results obtained with respect to different SQL Queries. At the same time, this research paper fulfills the following goals: (i) surveying the problem of evaluating the missing values i.e. why-not questions in SQL queries; (ii) searching the techniques for giving answers to such type of questions using different numeric and non-numeric data and (iii) comparing those efficient strategies. This research paper also gives attention towards related work which were done so far .*

***Key Words*: Missing answers, usability, SQL, top-k**

1. INTRODUCTION

After decades, database community is taking efforts on evaluation of missing values i.e answering to why-not questions and comparing techniques used for evaluation of this type of missing answers. The systems which are used today are more efficient. However, these types of systems are prominent in determining evaluation of management of data and evaluation of query [2]. But to the same degree, these systems are not suitable for end users. Now a day's , users are expecting that systems should be easy to interact and understand. It means, users are not agreed upon the results obtained from such type of systems. Users are more interested in knowing reasons for why the current set of result does not match their expectation i.e. why current set of objects are returned in the result by these systems. In particular, users may interested in knowing the reason for missing expected data object in result set and also interested in knowing why unexpected data objects appear in result received from the system. As a next step, users may also find proper explanations for these types of questions. Any system that can provide best explanations for type of questions mentioned above can be very helpful for users to better understand their information needs and also to make system more transparent and interactive to users[3],[5]. At present, traditional database systems are unable to provide any kind of exploratory data analysis facilities to support above types of why and why-not questions. The studies focusing on improvement of database usability (e.g., keyword search [2], similar graph matching [4], and spatial keywords[5]), explaining the feature of missing tuples which are not present in the result of query, are getting more importance. A why-not question [1], [2] is being posed when a user is interested in knowing why their expected tuples are not present in query result. Recently users are unable to sift directly in the set of data to examine "why-not?", due to the reason that interface of query i.e web forms are restricted by the types of queries expressed by them. When end users fires SQL query to get data from database and ask "why-not?"and are unable to search the possible ways fir getting explanation by means of query interface, easily cause the situation where users does not use the tool anymore. This would be the bad situation for database developers who give their most of the time to develop such applications. At the same time, supporting different aspects of giving explanation for missing answers [1], knowledge of algorithms which are based on query evaluation is required, that is out of scope for most database developers. Recently, community of database started the research on techniques to evaluate missing answers. Out of this, recent works focuses on giving answers to why-not questions. In this research paper, answering both why and why-not questions are addressed for numeric and non-numeric data present in SQL queries. In this research paper , aim is to evaluate missing answers in SQL queries in terms of the above mentioned aspects in different numeric and non-numeric data that have not been investigated by others. Rests of the sections of this paper are as follows: Section 2 describes related work; Section 3 presents comparison between strategies; Section 4 outlines future work; and Section 5 concludes our paper.

2. RELATED WORK

Previous studies [1],[4],[6],[7],[2],[3]and [5] have done research on problems of evaluating missing answers in SQL queries in terms of various different perspective. Xu et al.

explains to a user why their expected answers are not present in the current set of objects in query result and returns a refined query that includes expected missing answers back to the result. Query refinement method [1] is used which includes numeric attributes. Drawback of this method is that it is not useful for non-numeric data. Islam et al. explains the problem of evaluating missing answers in matching of similar graph which are used for graph databases. To address this problem, they have proposed an approximate solution approach as computing the exact solution is NPhard [4]. The search space for the new query graph is also established. Drawback is only suitable for graph databases. He et al. explains the problem of answering why-not question on two types of top-k queries: the basic top-k query where the users need to specify the set of weighting[6] and the top-k dominating query where users do not need to specify the set of weightings as the ranking function ranks an object higher if it dominate more objects. Drawback is not suitable for non-numeric data. Saiful et al. proposes technique that aims at evaluating the why-not questions in queries which are reverse skyline [7]. Also technique to explain modification of why-not and query point which includes why-not point in reverse skyline of the point called as query point. It also explains position of query point anywhere in a region without disturbing existing points that are reverse skyline. Drawback is only suitable for points of data whose dynamic skyline contains query points. Chen et al. proposed that keywords which are special in top-k queries retrieves the k objects which are best as per the function which considers both distance which is called a special distance and similarity of text. Algorithm which is having optimization sets that performs sequential examination of sets of candidate keywords is developed. Also index-based bound-and-prune algorithm [2] is used. Drawback is only suitable for initial set of query keywords. Geo et al. define and offer solutions to why-not questions on MPRQ [3]. He have given a proposal of a framework which are having three solutions that are efficient as follows : one which involves modification of original query, one which involves modification of why-not set, and last that involves both modification of original query and why-not set. Time required is more as experiments are performed using data sets which are synthetic and original. Chen et al. addresses problem of evaluating the missing answers in terms of keywords top- k queries by performing the refinement of keywords which are original that provides user with those keywords which explains their intention of query. Also the algorithm having different optimized techniques [5] is proposed that searches the better solution which is based on sets of keyword tested one by one. In some cases, identification of keywords becomes difficult for users.

3. COMPARISON BETWEEN DIFFERENT STRATEGIES

SR.NO	STRATEGY	ADVANTAGE	DISADVANTAGE
1	Query refinement method	Used for finding missing values which include numeric attributes	Not suitable for non-numeric data.
2	Index-based bound-and-prune algorithm	Evaluate the sequence of keywords sequentially.	Only suitable for examination of keywords present in query.
3	Metric probabilistic range queries	Define and offer solutions to why-not questions o n MPRQ.	Time required is more as experiments are performed using both real and synthetic data sets.
4	NPhard	Explains the problem of evaluating missing values in matching similar graph.	Only suitable for graph databases.
5	Optimizatio n Techniques	Determines the good solution totally based on keywords which are tested at once.	Identification of exact keywords is a difficult task for the users.

6	Ranking Function	address the problem of answering why-not questions on two types of top-k queries: the top-k dominating queries and basic top-k query.	Not suitable for non-numeric data.
7	Queries called as Reverse skyline.	Describes how to update the points called as why-not points and also the query point.	Only suitable for points of data whose dynamic skylines contains query points.

4. FUTURE WORK

The problem of evaluating missing answers in SQL Quires i.e answering to why and why-not questions in other data settings including social networks are studying currently. In particular, working is going on the following type of queries to answer the why and why-not questions.

Social and Graph Queries: Due to the emerging websites of social networking and their greater impact on our daily life, there is urgency for social queries on such networks. Social networks data are generally represented as graphs in databases. Many websites of social networking develop recommendation that are automatic over different items like giving suggestions on making new friends, events etc. Hence, the feedback for such automated recommendation is of more importance if user is not satisfied with them always. Any social networking websites that can answer such type of why and why-not questions will be more interesting to their users. In future work, this issue can be studied on queries including data types like Binary Large Object (BLOB), Boolean and others for missing value restoration and thereby making the system flexible for maximum databases.

5.CONCLUSION

This paper presents the research agendas for evaluating the missing answers in SQL Queries. Also shown why it is worth conducting such research and outlined the various techniques of giving answers to such type of why-not questions. These papers have also summarized the related work done in this area and the future research agendas. Finally, contributions made so far are presented. Currently work is in progress and focusing on future research problems mentioned in this paper.

ACKNOWLEDGEMENT

The Research presented in this paper is possible due to the data which is available through various reference papers mentioned in this research paper. I take this opportunity to express my gratitude towards Asst. Prof. Ms. Antara Bhattacharya for continuous supporting and encouragement. I would like to thank our Head Of the Department V.M.Sahare for giving proper guidance to submit this paper.

REFERENCES

[1] Wenjian Xu, Zhian He, Eric Lo, and Chi-Yin Chow, “Explaining Missing Answers to Top-k SQL Queries,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 2071–2085, July 2016.

[2] Lei Chen , Jianliang Xu, Xin Lin, Christian S.Jensen and Haibo Hu, “Answering why-not spatial keyword top-k queries via keyword adaption,” in Proc. IEEE 32nd Int. Conf. Data Eng.,2016, pp. 697-708.

[3] Yunjun Gao, Kai Wang, Christian S. Jensen and Gang Chen, “Answering why-not questions on metric probabilistic range queries,” in Proc. IEEE 32nd Int. Conf. Data Eng.,2016, pp. 767-778.

[4] M. S. Islam, C. Liu, and J. Li, “Efficient answering of why-not questions in similar graph matching,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 10, pp. 2672–2686, Oct. 2015.

[5] L. Chen, X. Lin, H. Hu, C. S. Jensen, and J. Xu, “Answering why not questions on spatial keyword top-k queries,” in Proc. IEEE 31st Int. Conf. Data Eng. , 2015, pp. 279–290.

[6] Z. He and E. Lo, “Answering why-not questions on top-k queries,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp. 1300–1315, Jun 2014

[7] Z. He and E. Lo, “Answering why-not questions on top-k queries,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp. 13Md. Saiful, Z. Rui, and L. Chengfei, “On answering why-not questions in reverse skyline queries,” in Proc. IEEE 28th Int. Conf. Data Eng., 2013, pp. 973–984.