International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Survey of Basic SQL Operators of Crowdsourcing System

Seema Yuvraj Shelar[1] , A.B.Rajmane[2]

[1]Ashokrao Mane Group of Institutions, Vathar, Maharashtra, India
[2]Ashokrao Mane Group of Institutions, Vathar, Maharashtra, India

Abstract: ***We study different SQL operators which are to be used in crowdsourcing system. Crowdsourcing is system which solves hard problems by combining the computer power and human logic. In crowdsourcing system, requester inputs SQL query and system processes the query, creates the execution plan and executes this plan on crowdsourcing platform. We require some SQL operators such as Select, Join, Fill, Max and Sort for the query optimization process in the crowdsourcing system. In the present paper, we studied the cited basic operations of crowdsourcing system.***

Keywords: Crowdsourcing, Query Optimization, Human Intelligence Tasks (HIT)

1. Introduction

Now a days, Crowdsourcing system is a successful software tool to getting accurate result. Crowdsourcing system uses human logic to solve hard problems which computer cannot solve correctly such as image tagging. Crowdsourcing system is an online process in which human gets work or posts work from the group of people e.g. Wikipedia [14]. Crowdsourcing system including Deco [7], CrowdDB [6], Qurk [10], and CrowdOP [9], provide an SQL query language as a declarative application to the crowd. The major objective of SQL is to eliminate the difficulties and a complexity related with crowd and gives crowdsourcing system an application that is familiar to database users.

The working of crowdsourcing system is shown in fig.1. Crowdsourcing system consists of two types users: requesters and workers. In crowdsourcing system, requester inputs query. Crowdsourcing system parses the query and with the help of query optimization process, it creates the evaluation plans. On the basis of the evaluation plan, crowdsourcing system creates human intelligence tasks (HITs) and posts on crowdsourcing platform. Human worker works on HITs and submits answer for HITs. Then an appropriate answer is returned to requester.

Deco [7] is declarative crowdsourcing system. It executes the SQL queries which are given over data collected from crowd and existing database. The present system solves the many problems which computer does not solve easily. In Deco, user submits requirements and system adopts the responsibility to output crowd source data dynamically to the user. Deco is adjustable model. It works on only missing values or records in database. Deco uses only fill operator for query optimization purpose.

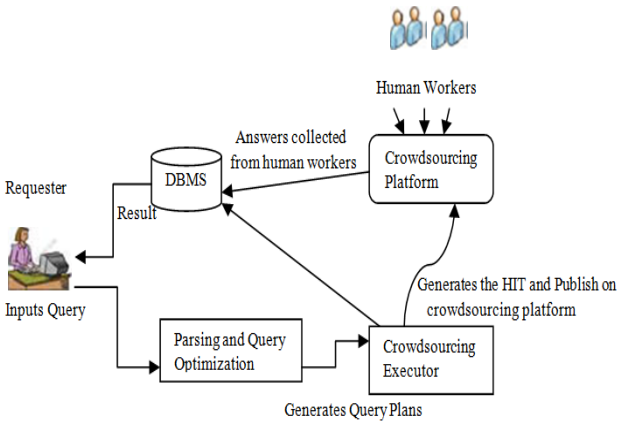

Figure 1: Workflow of Crowdsourcing system

CrowdDB [6] is another declarative crowdsourcing system. It takes the human answer through the group of people to execute the queries which are not solve by database engine. This system focuses on the following features. 1) With the help of SQL, CrowdDB creates the declarative interface. It sustains SQL syntax so that developers are introduced with familiar computational model. 2) In CrowdDB, programmer writes SQL queries without having to focus on which operation takes place in the database and which operation takes place by the crowd. If SQL queries run on CrowdDB, we get correct answers instead of running SQL queries on traditional DBMS. 3) User interface design is considered as major factor in the process of generating human tasks to be solved by people. 4) Due to the declarative interface and SQL operator based approach, CrowdDB gives the accurate cost based optimization to obtain the best query result. CrowdDB uses three operators: Crowdprobe, Crowdjoin and Crowdsompare. Crowdprobe works on missing information of database. Crowdjoin implements an indexed nested loop join over two relations. Crowdcompare implements crowdequal and crowdorder functions.

CDAS [11] is developed to support the creation of various crowdsourcing systems. CDAS uses two models 1) predication model 2) verification model. Qurk [10] is also declarative crowdsourcing system. It uses sort and join operators in query optimization processes. CrowdOp [9] is another declarative crowdsourcing system. This crowdsourcing system uses Cselect, Cjoin, Cfill queries for

optimization purpose. CrowdFind [12] system uses the select operator for query optimization process. This operator is used for filter purpose.

2. Literature Survey

A. G. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom[2] proposed one database operation selection. This operator selects the tuples from database based on some condition. They proposed optimal and heuristics algorithm efficiently to find filtering strategies that result in significant cost. This algorithm is used in various crowdsourcing systems for query process. They focused on the single selection problem. Also **P.Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis [4]** introduced MAX method which finds maximum item from dataset. They proposed parameterized MAX algorithm. It considers input as a set of items and output as an item from the set. This framework supports various human errors, cost models and also tradeoff between quality cost and execution time. They described method which finds best or maximum items in a set. This method evaluates simple and used parameters like execution time, cost and quality of result.

Sorting and joining operations are proposed by **A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller [5].** Sorting is used to sort the items in some given order. For this purpose, the author proposes three approaches: comparison based, rating based and hybrid of this two approaches. Join operation is used to compare items from two tables and produces result. For this purpose, author describes three types of interface: simple, naive batching and smart batching. These batching interface decrements the total count of HITs to solve the join by an order of magnitude. **A. Marcus, D. R. Karger, S. Madden, R. Miller, and S. Oh [3]** described count operator. Count is used to calculating the number of items in the dataset that satisfies specific condition. For this purpose, they used two methods one is count based and second is label based. Label based approach is samples tuple and ask to crowd to label the category assigned to each tuple until user get a desired output. The count based approach displays a collection of items to a worker and counts how many of items fall into a particular category.

3. SQL Operators and Syntax of Crowdsourcing System

The following SQL operators are used in crowdsourcing system. All example of query executes using auto import dataset [1].

1. **Select Operator:** This operator finds the specific tuples from dataset by satisfying certain condition. Usually, input of select operator consists of set of tuples „T‟ and collection of selection condition „S‟ and output is „t‟ which is subset of „T‟ such that all tuples in „t‟ satisfies the condition in „S‟. An example query for finding car having make is jaguar. It can be expressed in query Q1.
   Q1: select * from auto where make=”jaguar”
2. **Fill Operator:** This operator is used to find out unknown field to computer but can be identified by human. Fill operator can fill the missing values in dataset. Specifically, input to the Fill operator is group of pairs of tuple set and attributes, $\{<t_1,a_1>,<t_2,a_2>\ldots\ldots\ldots<t_n, a_{n>}\}$ and output is tuples sets $\{t_1, t_2,\ldots\ldots\ldots t_n\}$ such that any tuples t‟€ t_n has its attributes $< t‟,a>$.
   E.g. In auto dataset price attribute have missing value. It is expressed by „?‟. By using Fill operator user can fill by value.
3. **Count Operator:** Count operator is used to calculate the number of items in the given attributes from given dataset. Usually, input to the operator is tuples„t‟ and output will give actual count of attributes which is present in the tuples „t‟. An example query for finding total car having make is jaguar. It can be expressed in query Q2.
   Q2: Select Count (*) from auto where make=”jaguar”
4. **Max Operator:** This operator is used to extracts the maximum items from a dataset in crowdsourcing environment. Specifically, input to this operator is set of tuples „T‟ of attributes „a‟ and output is a tuple „t‟ which is maximum tuple from specific attribute but „t‟ is not belongs to tuples „T‟. An example for query finding more expensive car. It can be expressed in query Q3.
   Q3: Select Max (price) from auto;
5. **Sort Operator:** This operator is used to order the dataset item in some specific order. Formally, input to this operator is set of tuples $\{t_1, t_2,\ldots\ldots\ldots t_n\}$ of attribute „a‟ and output is set of tuples $\{t_1, t_2,\ldots\ldots\ldots t_n\}$ in ascending or descending order. Here, we use order by clause. An example for query sorting car price in the ascending order of price. It can be expressed in query Q4
   Q4: Select price from auto order by price asce.
6. **Join Operator:** Join operator is used to combine the objects from two relations according to certain conditions. Usually, input of a join operator consists of two tuples sets „T_1‟ and „T_2‟ and collections „S‟ of join conditions. The output is set $\{<t_1 , t_2>\}$ which subset of T_1 x T_2. An example query finding car image whose color is red and quality is high. It can be presented in query Q6.
   Q6: Select R1.*, R2.image From R1 auto, R2 image Where R2.color=”red” AND R2.quality=”high” AND R1.make=R2.make AND R1.model=R2.model.

4. Application and Example of Crowdsourcing System

An example of crowdsourcing system is Amazon Mechanical Turk (AMT) [8]. A large number of experiments were conducted in Amazon‟s site. Crowdsourcing system has some following application.

1) Voting System [13]: In this type of crowdsourcing system, a user is required to select an answer from number of choices. The answer that the most of users select is considered to be correct. Voting is used as device to obtain the correctness of answer from the crowd.

2) Information Sharing System [13]: Website is used to share information between internet users. Some crowdsourcing systems aim at sharing various types of information among the crowd. A famous information sharing systems were launched on the Internet as shown in the following:
   - Wikipedia [14] is online information system in which internet users writes work or gets work from the group of peoples.
   - Yahoo! Answers [15] is identified as a general question- answering website which gives the human abstracted data to the user.
3) Creative System [13]: In creativity mode, the contribution of human work cannot be replaced by an advanced technology. The creative task of human cannot be done by computer or any advanced technology, such as coding and drawing. As a result, some researchers do creative task for crowdsourcing workers to reduce the production costs. An example is Sheep Market [16]. It is website in which lots of workers can creates the database of drawing.

5. Conclusion and Future Enhancement

We have surveyed various crowdsourcing operators such as Select, Join, Fill, Sort and Max. These operators are very essential in crowdsourcing system for query optimization process. The Crowdsourcing system is efficient way to process the query that cannot processed by computer. In addition, we have surveyed various declarative crowdsourcing systems. This survey provides better understanding about SQL operators of various crowdsourcing systems.

Future Enhancement: For the enhancement of these existing system, with the use of SQL operators we can design system which will be able to receive query from user, parse and optimize the query, generates execution plan with low latency. With respect to above result, it generates task on crowdsourcing platform and collects the answer for this task from workers and give it to user for his further processing.

References

[1] https://archive.ics.uci.edu/ml/ datasets/Automobile.
[2] A.G. Parameswaran, H. Garcia-Milina, H. Park, N. Polyzotis, A. Ram and J. Windom, "CrowdScreen: Algorithms for filtering data with humans," in Proc ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 361-372.
[3] A. Marcus, D.R. Karger, S. Madden, R. Miller, and S. Oh, "Counting with the crowd", Proc. VLDB Endowment, vol. 6, no. 2, pp. 109-120, 2012.
[4] P. Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis, "Max algorithms in crowdsourcuing Environment", in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 989-998.
[5] A. Marcus, E. Wu, D.R.Karger, S.Madden, and R.C.Miller, "Human- powered sort and joins", Proc. VLDB Endowment, vol. 5, no. 1, pp. 13- 24, 2011.
[6] M.J.Franklin, D.Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB: Answering queries with crowdsoucing", in Proc.ACM SIGMOD Int. Conf. Manage. Data,2011, pp. 61-72.
[7] A.G.Parameswaran, H.Park, H.Garcia-Moline, N. Polyzotis, and J. Widom, "Deco: Declarative crowdsoucing", in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage, 2012, pp. 1203-1212.
[8] http://www.mtruck.com/mtruck.
[9] Ju Fan, Meihui Zhang, Stanley Kok, Meiyu Lu, and Beng Chin Ooi, "CrowdOp: Query Optimization for Declarative Crowdsourcing Systems", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no.8, pp. 2078- 2092, August 2015.
[10] A. Marcus, E. Wu. S. Madden, and R.C.Miller, "Crowdsourced databases: Query processing with people", in Proc. 5th Biennial Conf. Innovative Data Syst. Res., 2011, pp. 211- 214.
[11] X.Liu, M. Lu, B.C. Ooi, Y. Shen, S. Wu, and M. Zhang, "CDAS: A crowdsourcing data analytics system", Proc. VLDB Endowment, vol. 5, no. 10, pp. 1040-1051, 2012.
[12] A.D.Sharma, A. Parameswaran, H. Garcia- Molina, and A. Halevy, "Crowd-powered find algorithms", in Proc. IEEE 30th Int. Conf. Data Eng., 2014, pp. 964-975.
[13] Man-Ching Yuen, Irwin King, and Kwong- Sak Leung, "A Survey of Crowdsourcing Systems", in Proc. IEEE Int. Conf. Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011, pp. 766-773.
[14] The free encyclopedia, http://en.wikipedia.org
[15] Yahoo! answers, http://answers.yahoo.com/
[16] The sheep market, http://www.thesheepmarket.com/.