Automatic Query Generation and Query Optimization for Declarative Crowdsourcing System

Miss.Priyanka D Sarode[1], Prof.Imran R Shaikh[2]
[1]*Computer Engineering Department, Savitri Bai Phule University, Pune, India,*
[2]*Computer Engineering Department, Savitri Bai Phule University, Pune, India,*
[1]priyankasarode238@gmail.com
[2]Imran.shaikh22@gmail.com

Abstract— In recent era firing optimized query on huge amount of data or on bulk data flooding from various resources is hot cake for researchers. In this search process crowd-sourcing attracted many industries as an effective tool for utilization of human intelligence for various works that computers and smart system cannot perform. Hence in this crowd sourcing various solutions for database operations are provided. Also recent crowd-sourcing systems have user friendly panels that helps newbie or user not having SQL knowledge to get desired result by firing query in his terms and his expected conditions. For user query, respective system analyzes the query, generate the execution plan, execute it, get the result, resolves errors. While this query can execute in many ways and its effective execution time will vary as per the execution plan. Hence to avoid this uncertainty in query execution there must be proper execution plan and query optimizer that evaluates the ways of execution OR execution plans from time and cost point of view and finally system should select potentially good plan for execution. In this paper we are analyzing the various approaches that help to resolve the user queries in crowd sourcing systems which supports cost based query optimization, optimizing multiple crowd sourcing operators and allow tradeoff in between cost from monetary point of view and latency.

Index Terms— Crowd sourcing, Query Optimization, User queries, Query execution plans.

1 INTRODUCTION

Query optimization is an operation of frequent relational database management systems. The query optimizer workout to regulate the most active way to evaluate a given query by considering the possible query plans. Practically, the query optimizer cannot be getting directly by users once queries are submitted to database server, and parsed by the parser; they are then moved to the query optimizer where optimization occurs. However, some database engines grant guiding the query optimizer with hints. Queries results are produced by accessing relative database data and evaluating it in a way that return the requested information. By the reason of database structures are complicated, in most cases, and especially for not-very-simple queries, the needed data for a query can be gathered from a database by bring it in different ways, through different data-structures, and in different orders. Each different way commonly requires different processing time. Processing times of the same query may have high variance, a second to minutes, hours, depending on the way selected. The plan of query optimization, which is an automated process, is to find the way to process a user query in small amount of time. Hence there must be system that helps to analyze the query, optimize it, find query execution plans and finally predict potential query plan for execution over crowd sourced data.

2 LITERATURE SURVEY

CrowdOp: Query Optimization for Declarative Crowd sourcing System [1]

In this paper declarative crowd sourcing is considered and refined query optimization system algorithm is discussed. In this paper declarative crowd sourcing is considered and refined query optimization system query algorithm is discussed.

Using the Crowd for Top-k and Group-by Queries [2]

In this paper, authors formally study the problem of evaluating such max/top-k and group-by queries using the crowd. Given two data elements, the answer to a type question is "true" if the elements have the same type and therefore belong to the same cluster. This paper proposes a Bayesian model to show the clustering approach. This is in contrast to their model where they assume that there is a fixed (but unknown) set of clusters partitioning the elements.
In this paper, 'Ranking based' and value-based error model is considered for optimization of query.

A Hybrid Machine-Crowd sourcing System for Matching Web Tables [3]

This paper proposed, Concept-based approach and Hybrid machine-crowd sourcing framework. This approach effectively addresses difficulties in web table matching.
Concept-based approach maps each column of a web table in a well-developed knowledge base, which represents it. And hybrid machine-crowd sourcing framework approach breaks human intelligence for different columns in web table. In this paper, authors made a simplification that the crowd was assumed to produce perfect answer.

CrowdScreen: Algorithms for Filtering Data with Humans [4]

In this paper, authors focused on fundamental building blocks, an algorithm to filter a set of data items. Authors use the term filter for each of the properties they wish to check.
Types of filter used:

1. "Image shows a scientist," and
2. "Image of people in which people looking towards the camera."

The optimal and heuristic algorithms efficiently find filtering strategies that result in significant cost savings relative to commonly-used strategies in crowdsourcing applications.

CDAS: A Crowdsourcing Data Analytics System [5]

In this paper, two types of models are proposed first one is PREDICTION MODEL (i.e.Economic Model in AMT, Voting-based Prediction) and the other is VERIFICATION MODEL (i.e. Probability-based Verification, Online Processing). These proposed model results show that their proposed model can provide high-quality answers while keeping the total cost low. The natural expertise of human workers to perform complex tasks that are very challenging for computers is allowed by Crowdsourcing techniques.This paper proposes quality-sensitive model.

Counting with the Crowd [6]

In this paper, technique is used to identify coordinated attacks from multiple workers.
In this paper author identified for images, a count-based approach to achieve accuracy. In order to Magnitude less HIT label-based approach is used. In this paper, authors find text-based counts; they also found that the label-based approach has better accuracy.

Human powered Sorts and Joins [7]

In this paper, authors compare items for sorting and joining data, two of the most common operations in DBMSs. MTurk platform is used Qurk, runs on top of crowdsourcing.

Deco: Declarative Crowdsourcing [8]

In this paper authors describe, Deco's data model, query language, and our prototype.
In this, Crowdsourcing and Databases Crowdsourcing Algorithms are used that are offered practical and principled approach for accessing crowd data and also integrating it with conventional data.
Query Optimization over Crowdsourced Data [9]
In this paper, Deco's cost-based query optimizer, building on Deco's data model, query language, and query execution engine is proposed. Objective of Deco's is to find the best query plan to answer a query.
It describes Deco's cost-based query optimizer. The Primary goal Deco's is to find the best query plan to answer a query.

Learning from Crowds [10]

In this paper authors were proposed probabilistic approach. This approach is used for supervised learning. This used to evaluate different experts and also gives an estimate of the actual hidden labels. Output indicates that the proposed method is superior to the commonly used majority voting baseline. Two key assumptions: (1) performance of each annotator does not depend on the feature vector for a given instance and (2) conditional on the truth the experts are independent, that is, they make their errors independently.

Finding with the Crowd [11]

This paper formally define the problem using the metrics of cost and time, and design optimal algorithms that span the skyline of cost and time, i.e., Authors provide designers the ability to control the cost vs. time tradeoff. In this paper, we studied the fundamental CROWDFIND of problem, relevant in many crowdsourcing applications. Authors developed a solution that lies on the skyline of cost and latency for two settings: when humans answer correctly, and when they may make errors. They made the simplifying assumption that all workers are equally capable, identifying spam workers and learning accuracies of workers over time while solving CROWDFIND problems are also interesting extensions.

Max Algorithms in Crowdsourcing Environments [12]

In this paper, authors investigated methods for retrieving the maximum item from a set in a crowdsourcing environment.

They developed parameterized families of algorithms to retrieve the maximum item and proposed strategies to tune these algorithms under various human error and cost models. Also they evaluate under many metrics, both analytically and via simulations, the tradeoff between three quantities: (1) quality, (2) monetary cost, and (3) execution time.
Algorithm Used:

- PARAMETERIZED FAMILIES OF MAX ALGORITHMS
  1. Plurality Rule
  2. Bubble Max Algorithms
  3. Tournament Max Algorithms

Model: 1.Human Error Models

CrowdER: Crowdsourcing Entity Resolution [13]
This paper represents studied the problem of crowdsourcing entity resolution. Authors described how machine-only approaches often fall short on quality, while brute force people only approaches are too slow and expensive.
Thus, they proposed a hybrid human-machine workflow to address this problem. In the context of this hybrid approach, In particular, the results indicated that (1) The two-tiered approach generated fewer cluster-based HITs than existing algorithms;
(2) Hybrid human-machine workflow significantly reduced the number of HITs compared to human-based techniques, and achieved higher quality than the state-of-the-art machine based techniques; and
(3) The cluster-based HITs can provide lower latency than a pair-based approach.
In this paper authors, proposed techniques that are as follow:
ENTITY RESOLUTION TECHNIQUES
Machine based Techniques
Hybrid Human Machine Workflow
HIT Generation Techniques:
Pairbased HIT Generation
Clusterbased HIT Generation

A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data [14]
In this paper, the techniques used are as follow:
QUERY PROCESSING ON DIRTY DATA
Sampling Error
Data Error
SampleClean Framework
In this paper, authors propose SampleClean, a novel framework which only requires users to clean a sample of data, and utilizes the cleaned sample to obtain unbiased query results with confidence intervals.
They also identify three types of data errors (i.e., value error, condition error and duplication error) that may affect query results, and develop NormalizedSC and RawSC to estimate query results for the data with these errors.

Question Selection for Crowd Entity Resolution [15]
This paper examines the problem of enhancing Entity Resolution (ER) with the help of crowdsourcing.
Algorithm:
brute-force" algorithm
For deriving the best question that has the highest expected accuracy.
2. GCER algorithm to produce an approximate result within polynomial time.
3. Half algorithm

3 PROPOSED SYSTEM

From the mentioned literature survey it is clear that there are existing systems that work on query optimization where datasets or databases are no so complicated. There are systems that work on the query execution plans though datasets have some problematic values. Though there are smart query optimizers, they are unable to deal in declarative crowd sourcing area. In this environment when user fire some query then existing system are unable to work on it form time estimation point of view. Also existing systems are unable to select cost effective query plan. Hence there must be such system that properly analyze the user query in crowd sourcing environment , also proposed system should introduce smart query optimizer that find proper query plans and finally evaluate it properly from monetary cost point of view and execution time point of view.

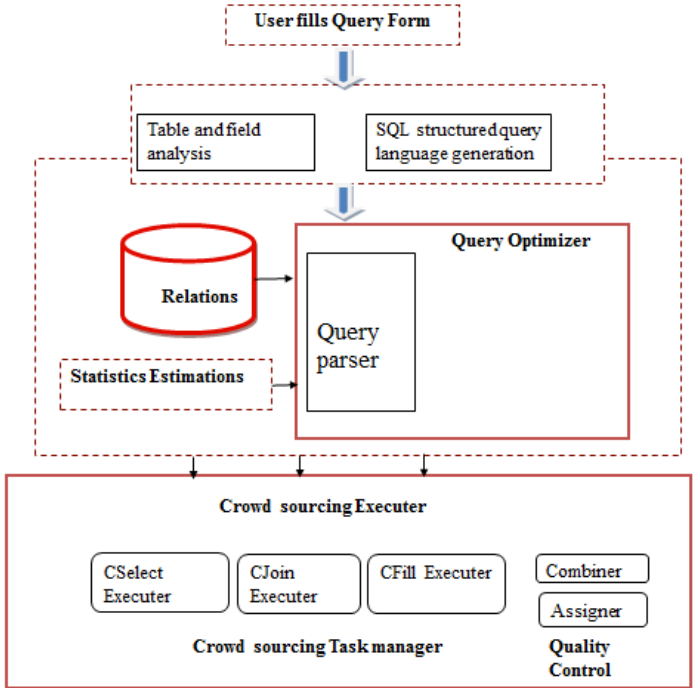

Figure 1: Block diagram of proposed system.

Hence in the proposed system user will first fill the form for the required attributes and conditions. The query generator module will automatically generate the query and this SQL query is issued by a crowd-sourcing environment for execution. The executer will first call QUERY OPTIMIZER. This optimizer parses the query and produces a best cost and time efficient query plan. The query plan is then executed by CROWDSOURCING EXECUTOR to generate human intelligence tasks (or HITs) and transfer these HITs on crowd sourcing platforms. Based on the HIT answers collected from the crowd, executer executes the query and returns the generated results to the user.

4 CONCLUSION

In crowd sourcing environment to hide query execution complexity and to encapsulate the execution phases there must be system that executes the user query with effective execution plans. System should recognize the best query execution plans using proposed algorithm in optimizer from cost and execution time point of view. This system should be user friendly so that newbie can fire his queries without knowing proper query language.

ACKNOWLEDGMENT

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thankful to Prof. S.R.Durugkar, Head of Department, Computer Engineering., S.N.D. College of Engineering and Research Center, Nashik. We also thank the anonymous reviewers for their comments.

REFERENCES

[1] CrowdOp: Query Optimization for Declarative Crowdsourcing Systems Ju Fan, Meihui Zhang, Stanley Kok, Meiyu Lu, and Beng Chin Ooi.

[2]] S. B. Davidson, S. Khanna, T. Milo, and S. Roy, "Using the crowd for top-k and group-by queries," in Proc. 16th Int. Conf. Database Theory, 2013, pp. 225–236.

[3] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang, "A hybrid machine-crowdsourcing system for matching web tables," in Proc.IEEE 30th Int. Conf. Data Eng., 2014, pp. 976–987.

[4] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB: Answering queries with crowdsourcing," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 61–72.

[5] [5] 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.

[6] 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.

[7] A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller, "Human-powered sorts and joins," Proc. VLDB Endowment, vol. 5, no. 1, pp. 13–24, 2011.

[8] A. G. Parameswaran, H. Park, H. Garcia-Molina, N. Polyzotis, and J. Widom, "Deco: Declarative crowdsourcing," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1203–1212.

[9] H. Park and J. Widom, "Query optimization over crowdsourced data," Proc. VLDB Endowment, vol. 6, no. 10, pp. 781–792, 2013.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," J. Mach. Learn. Res., vol. 11, pp. 1297–1322, 2010.

[11] A. D. Sharma, A. Parameswaran, H. Garcia-Molina, and A. Halevy, "Crowd-powered find algorithms," in Proc. IEEE 30th Int. Conf. Dta Eng., 2014, pp. 964–975.

[12] P. Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis, "Max algorithms in crowdsourcing environments," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 989–998.

[13] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," Proc. VLDB Endowment, vol. 5, no. 11, pp. 1483–1494, 2012.

[14] J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng, "Leveraging transitive relations for crowdsourced joins," in Proc. SIGMOD Int. Conf. Manage. Data, 2013, pp. 229–240.

[15] S. E. Whang, P. Lofgren, and H. Garcia-Molina, "Question selection for crowd entity resolution," Proc. VLDB Endowment, vol. 6, no. 6, pp. 349–360, 2013.

AUTHORS

Ms.Priyanka D Sarode received the B.E. degree in Information Technology from SND College of engineering & research centre, Yeola in 2013. She is currently pursuing her Masters degree in Computer Engineering from S.N.D. College of Engineering and Research Centre, Savitribai Phule Pune University Former UOP.This paper is published as a part of the research work done for the degree of Masters.

I. R. Shaikh professor at S.N.D. College of Engineering and Research Centre, Department of Computer Engineering, Savitribai Phule Pune University.