ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

New Approach for Range Aggregate Queries in Big Data Environments

Chetan Badhe [1], Mahendra Gawande[2], Ganesh Arbad[3],

Dept. of Computer Engineering, G. S. M. College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT*:* Range aggregate queries are nothing but to find certain aggregate function on all kinds of query ranges. Existing approaches to range aggregate queries does not provide accurate results in big data environments. So In this paper, we propose new approach for RAQ, that is Hadoop technology used to access range aggregate queries in big data environments. New approach for RAQ first divides big data into independent partitions with a balanced partitioning algorithm, and then generates a local estimation structure for individual partition and provides respected query result. When a range-aggregate query request arrives, new approach for RAQ obtains the result directly by summarizing local estimates from all partitions. Experimental results demonstrate that new approach for RAQ provides range-aggregate query results more earlier by using YARN rather than Hive.

KEYWORDS: Balanced partition, big data, hadoop, range-aggregate query, Background knowledge, TTP (trusted third party).

I. INTRODUCTION

BiG data is huge or complex type of data that traditional data processing applications are inadequate. Main challenges for big data environment include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy. Big data environment provides a new opportunity to various social aspects and preferences of individual everyday behaviors and explore fundamental questions about the complex world. Now days, it is important to provide efficient methods for big data analysis because amount of database is being increasing rapidly. We give an application example of big data analysis.

Background of Research

Our goal with the Feature Generation module is to provide a knowledge-intensive and computationally efficient coarse-grained analysis of *historical prices* which can be analyzed further in a second layer of reasoning. The domain knowledge implemented in the module is thus limited to methods and techniques in technical analysis. The technical analysis literature includes a wealth of different stock analysis techniques, some of which involve complicated and intricate price patterns subjective in both detection and interpretation. These methods would be both computationally expensive to detect and evaluate, and have consequently been disregarded. We thus apply Occam's razor to the choice of methods in technical analysis, focusing on the most popular indicators that can be efficiently operationalized and are intuitive in interpretation.

II. LITERATURE SURVEY

In this chapter we review overall literature of system. The [1] propose using 3D subspace-clustering algorithms to mine rules that are related to high stock price returns. The 3D subspace-clustering approach groups stocks that have similar fundamentals (financial ratios) and high price returns across years. The highlighted region in Figure 1a is a 3D subspace cluster containing stocks *s*2, *s*3, *s*4 that have similar fundamentals reflected in financial ratios *r*2, *r*3, *r*4 for years 1–3, 5–6, and 8–10. From Figure 1b, we can see that stocks *s*2, *s*3, *s*4 have high price returns.

Decision tree for stock trading is built based on technical analysis indicators. Indicators are set of mathematical formula that is calculated from stock prices data. These indicators are ordered to create certain trading rule, this rule become testing node for the decision tree. Based on the trading rule it can be calculated which price of stock would be suitable for a particular class. Those classes represent the decision that would be taken in the system The res ult will also be plotted to a chart for an easier trend analysis.Some systems are built based on financial market technical analysis

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

indicators (Exponential Moving Average, Moving Average Convergence Divergence, Relative Strength Index, Money Flow Index, and parabolic Stop and Reverse) [2] method is arrange indicators set into decision tree based on stock trading rules and classification which describe the rules and create buy, hold, and sell classes which represented decisions in trading. Decision classes then are analyzed for their profitability, geometric mean return, and cumulative wealth index. Furthermore sensitivity analysis is added into profitability analysis to obtain more positive value trading in decision making. The research purpose is to enhance decision making in technical .stock trading.
System [3] describe an intelligent stock trading system by combining support vector machine (SVM) algorithm and box theory of stock. The box theory believes a successful stock buying/selling generally occurs when the price effectively breaks out the original oscillation box into another new box. In the system, support vector machine algorithm is utilized to make forecasts of the top and bottom of the oscillation box. Then a trading strategy based on the box theory is constructed to make trading decisions. The different stock movement patterns, i.e, bull, bear and fluctuant market, are used to test the feasibility of the system.
The resulting model is intended to be used as a decision support tool or as an autonomous artificial trader if extended with an interface to the stock exchange. Machine learning approach is also very effective for stock market analysis. System [4] define a stock price prediction model will be created using concepts and techniques in technical analysis and machine learning. The resulting prediction model should be employed as an artificial trader that can be used to select stocks to trade on any given stock exchange.

In understanding the difficulties facing investors during the investment decision process; consider the case of common stocks in financial markets that produce on average significantly large return over the years than the saving account. However, a worthy range of investors avoid realizing these large returns due to the fact that "to pursue high returns investors must assume large risks. Model [5] formulated as a multi-criteria optimization model (maximizing the net profit and minimizing the maximum drawdown) to be solved for the contribution ratio of each trading decision model component in the trading decision pool. It has been demonstrated that the proposed strategy of combining different trading decision models results in noticeable increase of profit as well as considerable decrease in the maximum drawdown.

Previous research showed promising results on the possibility of correctly predicting the price direction of a stock or market index. We mention some of that work here. [6] proposed 5-days-ahead and 10-days-ahead predictive models are built using the random forests algorithm. The models are built on the historical data of the CROBEX index and on a few companies listed at the Zagreb Stock Exchange from various sectors. Several technical indicators, popular in quantitative analysis of stock markets, are selected as model inputs.
Some systems classify bi-clustering patterns into two, based on whether the pattern is defined on a single cluster or multiple clusters. If a bicluster pattern is defined on a single bicluster, we call the pattern a local pattern. Otherwise, we call the pattern a global pattern. [7] describe address this less studied, yet important problem and formally analyze several bi-clustering algorithms in terms of the bicluster patterns they attempt to discover. We systematically formulate the requirements for well-known patterns and show the constraints imposed by bi-clustering algorithms that determine their capacity to identify such patterns. We also give experimental results from a carefully designed test bed to evaluate the power of the employed search strategies.
The system [8] show the detail explanation of k-means classification. The [9] subsets of genes that have similar behavior under subsets of conditions, so we say that they coexpress, but behave independently under other subsets of conditions. Discovering such coexpressions can be helpful to uncover genomic knowledge such as gene networks or gene interactions. That is why, it is of utmost importance to make a simultaneous clustering of genes and conditions to identify clusters of genes that are coexpressed under clusters of conditions. This type of clustering is called bi-clustering . It also focused on bi-clustering of gene expression data. The rest of the paper is organized as follows: First, we introduce some definitions related to bi-clustering of microarray data. Then, we present in section 3 some evaluation functions and bi-clustering algorithms. Next, we show how to validate biclusters via bi-clustering tools on microarrays datasets.
System [10] define novel approach for customer segmentation which is the basic issue for an effective CRM (Customer Relationship Management). Firstly, the chi-square statistical analysis is applied to choose set of attributes and K-means algorithm is employed to quantize the value of each attribute. Then DBSCAN algorithm based on density is introduced to classify the customers into three groups (the first, the second and the third class). Finally bi-clustering based on improved Apriori algorithm is used in the three groups to obtain more detailed information. Experimental

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

results on the dataset of an airline company show that the bi-clustering could segment the customers more accurately and meticulously.

III. PPROPOSED SYSTEM

FastRAQ—a new approximate answering approach that acquires accurate estimations quickly for range-aggregate queries in big data environments. Fast RAQ first divides big data into independent partitions with a balanced partitioning algorithm, and then generates a local estimation sketch for each partition. When a range aggregate query request arrives, Fast RAQ obtains the result directly by summarizing local estimates from all partitions. The balanced partitioning algorithm works with a stratified sampling model. It divides all data into different groups with regard to their attribute values of interest, and further separates each group into multiple partitions according to the current data distributions and the number of available servers.

Methodology & Algorithm:

Grouping (Queries) - The stratified sampling is a method to subdivide the numerical value space into independent intervals with a batch of logarithm functions, and each interval stands for a group. When the number of logarithm functions is fixed, an arbitrary natural integer N can be mapped into a unique group g from the database values.

Partitioning(R,VP) : In big data environments, a partition is a unit for load balancing and local range-aggregate queries. FastRAQ uses the vectors set VP < Ve*i*…..… Ve*m*; Vr*i*……. Vrm> M to build partitions for all the incoming records, where M indicates the number of groups. In each partition, a dynamic sample is calculated from the current loaded records. Currently, FastRAQ uses a mean value of aggregation-column as the sample, which is Sample ¼ SUM=Counter, where SUM is sum of values from aggregation-column, and Counter is the number of records in the current partition. A detailed balanced partition algorithm

IV. ALGORITHMS

Input : *Q ;*
Q : select sum(Aggcolumn) otherColname where
li1<ColNamei<li2 opr lj1<ColNamej<lj2.
Output: *S;*
S: range-aggregate query result.
1: Deliver the request Q to all partitions;
2: **for** each partition i in partitions **do**
3: Compute the cardinality estimator of range li1 < ColNamei < li2 from the local histogram,
and let CEi be the estimator of the ith dimensions;
4: Compute the cardinality estimator of range lj1 < ColNamej < lj2 from the local histogram,
and let CEj be the estimator of the jth dimensions;
5: Merge the estimators CEi and CEj by the logical operator Opr, and compute the merged cardinality
estimator CEmerged;
6: Counti ← ħ (CEmerged); // ħ is a function of cardinality estimation.
7: Compute the sample for AggColumn, and let Samplei be the sample;
8: SUMi ← Counti ×Samplei ; //SUMi is a local range-aggregate query result;
9: **end for**
10: Set the approximate answering of FastRAQ as S.
Let S ← where M is the number of partitions;
11: **return** S;

V. RESULT

The framework is based on hadoop 2.0 rendition and MongoDB database. For dataset we utilize semi-structure xml information which comprise of a few quality. At present we contrast windows RAQ framework and Linux stage. The proposed framework gives the assessed results and we contrast and diverse existing frameworks. On the hypothetical results demonstrates the tasteful level exactness. Here table 1 demonstrates the proposed exactness also time many-sided quality how it is superior to anything existing methodologies.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

This endeavor is completing the portion figuring on PIG advancement with help of HADOOP stage what not composing PC projects is done with the help of the Java vernacular. After that wander need execution of computations for get ready of aggregate inquiries and applying a figuring on them. The endeavor has named center point i.e. master center point and data centers (slave Node). In this endeavor as an issue of first significance the section of the data set is done which is uncontrolled as the Hadoop system does not control data. So by use of the balanced fragment estimation, data is controlled and pieces are made in the in the first place period of yield. By then to map the data and utilizing the time histogram is made. This is the second yield of the endeavor, which gives preference while organizing contain with the customer request. Essential purpose behind the endeavor is dealing with data successfully for the aggregate limits which are ended on one or more area on the gigantic data.

VI. CONCLUSION

VII. We have surveyed RAQ dispensation techniques in social databases. We providing a classification for RAQ methods based on several sizes such as the accepted query model, data admission, implementation level, and reinforced ranking functions. We deliberated the details of numerous algorithms to exemplify the different tests and data organization glitches they speech. We also deliberated linked RAQ indulgence methods in the XML part, as well as procedures used for counting XML rudiments. Finally, we providing a theoretical contextual of the ranking and RAQ processing problems from voting theory.Efficient dispensation of RAQ queries that contract with different bases of uncertainty and hairiness, in both data and inquiries, is a challenging task. Scheming uncertainty replicas to meet the wants of practical requests, as well as extending interpersonal processing to conform through different probabilistic copies, are two important subjects with many unexplained problems. Abusing the semantics of RAQ inquiries to identify optimization odds in these settings is another important question.

VIII. FUTURE WORK

Fast RAQ is a new approximate replying approach that acquires correct estimations quickly for range-aggregate queries in big data milieus. Fast RAQ first gulfs big data into independent dividers with a balanced partitioning process, and then generates a local estimation sketch for each divider. When a range aggregate query appeal arrives, Fast RAQ obtains the consequence directly by brief local estimates form all partitions.

ACKNOWLEDGMENT

We are profoundly grateful to **Prof. Shrinivas Halhalli** Project Coordinator **and Prof. Ratna Raja Kumar** Project Guide, for their expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. We are also grateful to **Prof. Shrinivas D.** for his support and guidance that have helped us to expand our horizons of thought and expression. We would like to express our deepest appreciation towards**,** Principal **P. B. Sayyad** G. S. Moze College of Engineering, Pune and **Prof. Ratna Raja Kumar,** Head of the Department, Computer Engineering Department whose invaluable guidance supported us in completing this project. At last we must express our sincere heartfelt gratitude to all staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

REFERENCES

[1] Kelvin Sim and Vivekanand Gopalkrishnan proposed 3D Subspace Clustering for Value Investing by the IEEE Computer Society in 2014.
[2] F.X. Satriyo D. Nugroho Teguh Bharata Adj, Silmi Fauziat proposed the DECISION SUPPORT SYSTEM FOR STOCK TRADING USING MULTIPLE INDICATORS DECISION TREE in I st International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) year 2014.
[3] Qinghua Wen, Zehong Yang, Yixu Song, Peifa Jia proposed Intelligent Stock Trading System based on SVM Algorithm and Oscillation Box Prediction Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009.
[4] Jan Ivar Larsen [4] proposed Predicting Stock Prices Using Technical Analysis and Machine Learning Norwegian University of Science and Technology Department of Computer and Information Science year June 2010.
[5] Tarek M. Gadallah and M. Nashat Fors proposed A New Approach for Combining Multi-Criteria Trading Decision Models, Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE), March 3 – 5, 2015.
[6] T. Manojlovic and I. Stajduhar proposed Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange MIPRO 2015, 25-29 May 2015, Opatija, Croatia.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

[7] Doruk Bozdag, Ashwin S. Kumar and Umit V. Catalyurek proposed Comparative Analysis of Bi-clustering Algorithms National Cancer Institute Grant R01CA141090; by the U.S. DOE SciDAC Institute Grant DE-FC02-06ER2775; by the U.S. National Science Foundation under Grants CNS- 0643969, OCI-0904809, OCI-0904802 and CNS-0403342.
[8] Kardi Teknomo proposed K-Means Clustering approach Teknomo, Kardi. K-Means Clustering Tutorials. http:\\people.revoledu.com\kardi\ tutorial\kMean\ July 2007.
[9] Haifa Ben Saber and Mourad Elloumi proposed A NEW SURVEY ON BI-CLUSTERING OF MICROARRAY DATA Natarajan Meghanathan et al. (Eds) : NeTCoM, CSIT, GRAPH-HOC, SPTM - 2014
[10] Xiaohui Hu and Haolan Zhang A New Customer Segmentation Framework Based on Bi-clustering Analysis JOURNAL OF SOFTWARE, VOL. 9, NO. 6, JUNE 2014.