ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 68-70

Comparison of Parallel and Apriori algorithms in Data Mining

D. Sasirega

Assistant professor, Department of Computer Science

KG College of Arts and Science, Coimbatore

d.sasirega@kgcas.com

Received 06 April 2017; Accepted 09 May 2017

ABSTRACT

Association rules are helpful and fascinating several data processing applications. Apriori algorithmic program is that the known association rule algorithmic program. This algorithmic program work with storage so as to access input and output the results. This paper shows the way to use this algorithmic program adapting the g storage system to the current drawback by the assistance of hints and parallel options.

Keywords: Apriori algorithmic

INTRODUCTION

Association rules are statements that helps to uncover relationships between on the face of it unrelated information during a data service} or different information repository.An association rule has 2 components, AN antecedent (if) and a resultant (then). AN ANtecedent is an item found within the information. A resultant is data that's found together with the antecedent. These rules ar created by analyzing information for frequent if/then patterns and victimization the standards support and confidence to spot the foremost necessary relationships. Support is data of however ofttimes things seem within the information. Confidence shows the quantity of times the if/then statements are found to be true. In data processing, association rules are used for analyzing and predicting client behavior. They play a vital role in searching.

Parallel algorithmic program data processing helps to search out meaning patterns or rules in giant datasets. It deals with statistics, high performance computing, analysis machine learning, and neural networks.

A main feature of knowledge mining tasks is that they're resource oriented and operate giant sets of knowledge. information sources ar measurement in gigabytes or terabytes . Data mining algorithms which will mine vast databases during a cheap quantity of your time.

several recursive enhancements projected in several serial algorithms, huge} size and spatial property of the many databases makes data processing tasks too slow and too big. it's thus a growing ought to develop economical parallel data processing algorithms which will run on a distributed system. during this project, many parallel data processing algorithms, supported the illustrious Apriori algorithmic program, was enforced and their performance evaluated on a parallel machine.

In data processing parallel algorithmic program or simultaneous algorithmic program, as opposition a standard serial (or serial) algorithmic program, is AN algorithmic program which may be dead a bit at a time on many various process devices, then combined along once more at the top to induce the proper result.

Parallel algorithms ar valuable to substantial enhancements in processing systems and additionally the increase of multi-core processors. In general, it's easier to construct a laptop computer with one fast processor than one with many slow processors with identical turnout. but processor speed is increased primarily by shrinking the equipment, and stylish processors ar pushing physical size and heat limits. These twin barriers have flipped the equation, making processing wise even for little systems.

The correspondence in academic degree rule can yield improved performance on many alternative forms of computers. as AN example, on a parallel laptop computer, the operations throughout a parallel rule are going to be per-formed at a similar time by utterly totally different processors. what's additional, even on a single-processor laptop computer the

Corresponding author: D. Sasirega |

correspondence in academic degree rule are going to be exploited by victimization multiple sensible units, pipelined sensible units, or pipelined memory systems. Thus, it is vital to create a distinction between the correspondence in academic degree rule and additionally the power of any express laptop computer to perform multiple operations in parallel.
process algorithms ar said to be a variation of Apriori. Writing parallel processing algorithms ar a non-trivial task. the foremost challenges associated with parallel processing embrace
- minimizing I/O
- minimizing synchronization and communication
- effective load equalization
- effective data layout (horizontal vs. vertical)
-preferring the foremost effective search procedure - minimizing/avoiding duplication of labor
Four Parallel Algorithms were used:
1. Count Distribution – parallelizing the task of measure the frequency of a pattern within a info
2. Candidate Distribution – parallelizing the task of generating longer patterns
3. Hybrid Count and Candidate Distribution – a hybrid rule that tries to combine the strengths of the on prime of algorithms
4. Sampling with Hybrid Count ANd Candidate Distribution – AN rule that tries to entirely use a sample of the data.

I. APRIORI algorithmic program

Apriori is also a classic rule for frequent item set mining and association rule learning over transactional databases. It issue by characteristic the frequent individual things inside the data and continuance them to larger and greater item sets as long as those item sets appear sufficiently sometimes inside the data. The frequent item sets verifyd by Apriori is accustomed confirm association rules that highlight general trends inside the database: this has applications in domains like market basket analysis.
Apriori employs a bottom-up, breadth-first search that enumerates every single frequent pattern in AN passing data. It starts by finding all frequent patterns of size one, that's then accustomed notice all frequent patterns of size a combine of etc. it's accustomed downward closure property of pattern support (all subsets of a frequent pattern ought to themselves be frequent) to prune the search space. thus entirely frequent patterns of size k square measure accustomed generate patterns of size k+1.
• Apriori is meant to manage on databases containing transactions (for example, collections of things bought by customers, or details of a web web site frequentation).
• The rule makes a trial to hunt out subsets that square measure common to a minimum of a minimum vary C (the cutoff, or confidence threshold) of the itemsets.
• Apriori uses a "bottom up" approach, where frequent subsets square measure extended one item at a time (a step known as candidate generation, and groups of candidates square measure tested against the data.
• The rule terminates once no a lot of victorious extensions square measure found.
• Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.
• desires several iterations of the data
• Uses a standardized minimum support threshold
• Difficulties to hunt out rarely occuring events
• Other ways (other than appriori) can address this by using a non-uniform minimum support thresold
• Some competitive totally different approaches target partition and sampling
Phases of information discovery:
1) Knowledge selection
2) Knowledge cleansing
3) Knowledge enrichment (integration with additional resources)
4) Knowledge transformation or secret writing
5) Data processing
6) Coverage and show (visualization) of the discovered data
Use of Apriori algorithm:
• Initial data: transactional information D and user-defined numeric minimun support threshold min_sup
• Algortihm uses data from previous iteration half to provide frequent itemsets
• this is often reflected inside the Latin origin of the name which suggests ”from what comes before” making frequent sets:
- Let’s define:
- Ck as a candidate itemset of size k
- Lk as a frequent itemset of size k

• Main steps of iteration are: o notice frequent set Lk-1
1) Be a part of step: Ck is generated by association Lk-1 with itself (cartesian product Lk-1 x Lk-1)
2) Prune step (apriori property): Any (k − 1) size itemset that is not frequent cannot be a group of a frequent k size item set, thus need to be removed
3) Frequent set Lk has been achieved making frequent sets (2)
• Algorithmic program uses breadth-first search and a hash tree structure to make candidate itemsets efficiently
• Then occurance frequency for each candidate itemset is counted

• Those candidate item sets that have higher frequency than minimum support threshold square measure qualified to be frequent itemsets
Apriori rule out pseudocode
L1= ;
for (k= 2; Lk-1 !=Ø; k++) do begin
Ck= candidates generated from Lk-1 (that is: product Lk-1 x Lk-1 and eliminating any
k-1 size itemset that is not frequent);
for every dealings t in info do increment the count of all candidates in
Ck that square measure contained in t
Lk = candidates in Ck with min_sup end
come Èk Lk;
Keep the basis phrase

IV RESULTS AND DISCUSSIONS

The goal of this experiment was to visualize the performance of parallelizing Apriori on C.P.U. for large data sets. Iris data was accustomed train the system that's "perhaps the foremost effective proverbial data to be found inside the pattern recognition literature" . The data set contains three classes of fifty instances each, where each class refers to a kind of iris plant. five attributes ar gift throughout this data that ar foliage Length, foliage breadth, floral leaf Length, floral leaf breadth, and additionally the class label attribute which could contain three values: "Iris-setosa", "Iris-virginica" and "Iris-versicolour". These informationsets were preprocessed before running the the algorithmic rule by building new data structures so they will slot in memory. However, given the tiny range of records within the Iris information, the experiment wouldn't mirror solid results, so all the fifty records were cloned and haphazardly appended one thousand times on to a novel larger Iris info of fifty,000 total records. The experiment was implemented on Associate in Nursing Intel 8-core machine and additionally the obtained results were recorded were compared to performance statistics of state-of-the art Apriori. By applying the on prime of mentioned parallel procedures, the C.P.U. program was ready to upset the state-of-the art Apriori. Results showed that the planned methodology is as fast as state-of-the art methodology but is healthier regarding BUS utilization where seventy fifth bus utilization was incurred by state-of-the art apriori and thirty fourth bus utilization was incurred by our planned methodology. This greatly reduced total power consumption of the algorithmic program.

CONCLUSION

during this paper parallel Apriori rule was implemented by applying a novel approach exploitation intermediate organization for computing and caching k+1 superset to be utilized in future iterations to facilitate computation of the superset in parallel and increase degree of similarity. The implementation of the parallel technique showed emulous results with state-of-the art apriori in speed but outperformed it in bus utilization by 400th the amount of your time of the rule which could build the rule economical for energy consumption than this best Apriori.

REFERENCES:

1. A survey of evolutionary algorithms for data mining and knowledge discovery,Alex a. freitas, Springer Berlin Heidelberg, Online ISBN 978-3-642-18965-4
2. www.computerscijournal.org/.../a-comparative-study-of-classification-techniques-in-datamining.
3. https://pdfs.semanticscholar.org/69c3/c9624d1cef93d45345523100823503feeaa8.pdf by MP Nancy - Cited by 20 - Related articles
4. www.ijcttjournal.org/Volume13/number-2/IJCTT-V13P117.pdf
5. *wireilla.com/papers/ijcsa/V5N5/5515ijcsa01.pdf*