IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Parallel PSNM Duplicate Record Detection with Map Reduce

Shubhangi Anandrao Dhane[1], Prof. Amrit Priyadarshi [2]

PG Scholar, Department of Computer Engineering, DGOI, COE, Bhigwan, Savitribai Phule Pune University, India[1]

Assistant Professor, Dept of Computer Engineering, DGOI, COE, Bhigwan, Savitribai Phule Pune University, India[2]

Abstract: Duplicate detection is a problem of serious substance in many applications, including customer relationship management, personal information management or data mining. Duplicate detection is method of detecting all cases of multiple illustration of same real world object. A representative example is customer relationship management, where a company loses money by sending multiple catalogs to the same person, who in turn is wound up lowering customer satisfaction. Another application is data mining, where correct input data is necessary to construct useful reports that form the basis of decision mechanisms.

Keywords: Duplicate Detection, Entity Resolution, Progressiveness, Pay-As-You-Go, Data cleaning, Map Reduce.

I. INTRODUCTION

Data mining depends on effective data collection and warehousing as well as computer processing. Most important property of a company is 'Data' but when data change or poor data entry, data errors such as duplicate detection occurs we want to make data cleaning for duplicate detection. However, duplicate detection processes expensive due to pure size of dataset [1].Duplicate detection is the procedure of identifying various representations same real-world objective in a information source[1]. The quality of duplicate detection, i.e., its effectiveness, scalability cannot be ignored because of the significant size of the database.

The duplicate detection problem has two aspects: First, the multiple representations are usually not the same but contain differences, such as misspellings, changed addresses, or missing values. This makes it difficult to detect these duplicates. Second, duplicate detection is a very expensive operation, as it requires the comparison of every possible pair of duplicates using the typically complex similarity calculate.The paper proposes the Parallel Duplicate Detection with Map reduce concept. The paper is structured in following way.

II. RELATED WORK

1.K. Elmagarmid, P. G. Ipeirotis, V. S.Verykios. [5].
Entity Resolution (ER) is original idea of hints, which can guide an ER algorithm to focus on resolving the more expected similar records first. Pay-As-You-Go ER and noticeably propose hints as a general technique for fast ER. several interesting problems remain to be solved, and different formal analysis of different types of hints and a general direction for constructing and updating the "best" hint for any given ER algorithm. There are three types of hints that are compatible with different ER algorithms as a sorted list of record pairs , a hierarchy of record partitions and an order list of records. Benefit of use of hints in ER is increase the number of same records recognized using a partial amount of work and improve ER quality.

2. S. E. Whang, D. Marmaros, and H. Garcia Molina. [6].
Non identical duplicate entries in database records detected by using this techniques. Paper work on both approaches for duplicate record detection. First is Distance-Based Approach which is used to calculate the distance between specific fields, using the proper distance metric for each field, and then compute the biased distance between the records. Second is Rule based Approach which is A special case of distance-based approaches, it uses rules to term whether two records are the identical or not. Rule-based approaches can be dignified as distance-based techniques, where the distance between two records represented in single bit binary number.

3. U.Draisbach,F.Naumann,S.Szott,& O.Wonneberg[7]
Duplicate Count Strategy which adapts the window size built on the different noticed duplicates. In this paper, There are three strategies: Key similarity strategy: Window size is based on the correspondence of the sorting keys: The window size is improved if sorting keys are similar and thus more related records can be expected. Record similarity strategy: Window size is based on the correspondence of the records: As a modification of the key similarity strategy, one regards as a replacement for the actual similarity of the records within the window. Duplicate count strategy: Window size is based on the number of well-known duplicates: If many duplicates are found within a window, it is possible if it create within an increased window.

4. U.Draisbach, F.Naumann [8].
Sorted Blocks which is a generalization of blocking and windowing methods. Blocking methods split records into separate subsets, Sorted Neighbourhood Method use sliding window over the arranged records and compare

DOI 10.17148/IJARCCE.2016.56147

records only within the window. Sorted Blocks in contrast to the Sorted Neighbourhood Method is the variable partition size instead of fixed size window. This methods can't assess strategies that group records with a high chance of being duplicates in the same partitions.

5. L.Kolb, A.Thor, E.Rahm.[9] Entity Resolution is also Called as Deduplication. It is used to make sure all entities related to the similar real world item. It is importance for data quality and data integration. Map reduces for parallel execution of SN blocking. Combination of blocking and parallel processing methods is used to implements efficient entity resolution for large datasets.

III.PROBLEM STATEMENT

Progressive Sorted Neighbourhood Method (PSNM) is based on the traditional sorted neighbourhood method increase the efficiency of duplicate detection with limited execution time. This methods detect only duplicate records in serially, is not removing that records.

Therefore we proposed new method named, PPSNM with Map Reduce that enable the efficient parallel execution of data-intensive tasks such as duplicate detection on large data sets. After that perform delete operation on copied records. The proposed system used for database record duplicate detection and database record deletion.

IV.PROPOSED SYSTEM

To overcome the problem of serial duplicate detection this work proposes an efficient and flexible detection scheme that supports both Progressive duplicate detection with map reduce and parallel duplicate detection. The proposed system is based on Map Reduce Algorithm.

A]Training Dataset :- In this Process user give the input data to the proposed system. Here training dataset loaded from company database or inserting from user .

B]Data Preprocessing:-Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

Data pre-processing is a proven method of resolving such issues. Data pre-processing prepare raw data for further processing.

C]Data Separation: In this process we separate the large amount of data ,i.e.large data cannot be fit in to main memory so it is divided into different parts each part is called as cluster.

D]Duplicate Detection: In this process we detect the duplicate records from cluster.

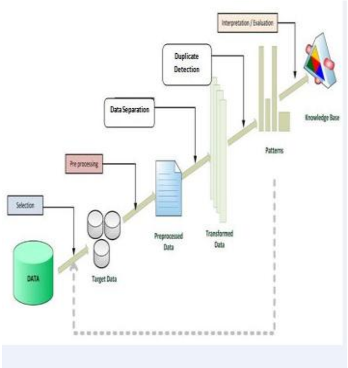


Fig.1 Proposed System Architecture

PSNM:-Progressive duplicate detection algorithms apply on selective input dataset(Cluster) that significantly increase the efficiency of finding duplicates if the execution time is limited. Duplicate detection is done on this phase .PSNM detect duplicate records sequentially. So Execution Time is higher than PSNM.
MAPREDUCE:- Map reduce algorithm apply on selective input dataset(Cluster) that significantly increase the efficiency of finding duplicates if the execution time is limited than PSNM. Duplicate detection is done on this phase. Map Reduce detect duplicate records Parallely. So Execution Time is less than PSNM.

V.ALGORITHM

The algorithm has the following steps.
Algorithm
Input: Training Data set
The following steps explain the implementations :
1. Start
2. PSNM algorithm apply to input data set
3. MAPREDUCE algorithm apply to input data set
4 Take result from PSNM and MAPREUCE
5. Compare the result.
6. Stop
Mathematical Model
Let S be a system that draws the output; such that S = I, Op, Om|Fs where
Set Theory
1. Let S be a system that draws the output.
S = {}be a system.
2. Obtain an input token for PSNM IP = {}
3. Obtain an input token for MAPREDUCE
IM = {}
4. PSNM processing on data set IP, MAPREDUCE processing on data set IM
5. Output of PSNM Op = {}
6. Output of MAPREDUCE Om = {}
7. S= {IP,IM,Op,Om}
Set Theory
1. Let S be a system that draws the output.
S = {}be a system.
2. Input given to PSNM algorithm.
IN ={A,B,C,D,E}
A,B,C,D,E =PARTITIONS of large data set, hear assume only 5 partition

3. Obtain an input token for MAPREDUCE.
IS = {A1,A2,A3,A4,A5}
A1,A2,A3,A4,A5 =PARTITIONS of large data set, hear assume only 5 partition
4. PSNM Processing on data set IP,
MAPREDUCE processing on data set IM
5. Output of PSNM.
Op = {Opo}
Op= Output of PSNM;
6. Output of MAPREDUCE
Om = { Omo }
Om= Output of MAPREDUCE.
7.Compare Output of PSNM and Output of Map Reduce
Selection Sort: Comparisons:

$$(N-1)+(N-2)+(N-3)+\ldots+1 \qquad = \frac{(N-1)*N}{2}$$

Duplicate Record Detection:
Here represent the set of ordered record pairs (with a record drawn from each file A and B for each pair) as

AXB = {(a, b); a ∈ A, b ∈ B

Each record pair is assigned to either class M or U. Record pairs belonging to the M class are identified as matching whilst record pairs belonging to the U class are identified as non matching. Therefore

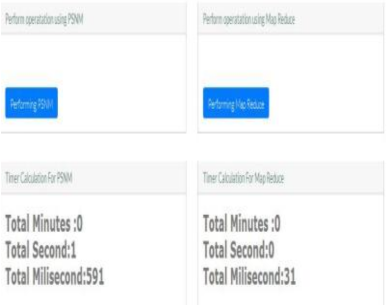
M = {(a, b); a = b, a ∈ A, b ∈ B}and
U = {(a, b); a != b, a ∈ A, b ∈ B}

Set M is called as Duplicate Record set.
Set U is called as Non-Duplicate Record set

The records corresponding to members of A and B are denoted by □ (a) and□ (b) respectively.

VI.RESULT ANALYSIS

Progressive Sorted Neighbourhood Method used for Detecting Duplicate Records in minimum amount of time as compare with simple Sorted Neighbourhood Method. The main drawback of PSNM is Time Complexity Because it detecting duplicate records serially. The performance evaluation of the proposed PPSNM Method is based on certain performance metrics. The performance metrics used in the paper is Map reduce Concept. This is Calculation of time required for finding duplicate detection using PSNM and Map Reduce Algorithm. After comparing , it is found that the performance of PPSNM(PSNM with Map reduce) is superior. The dataset used can be dynamically added and used as per user convenience. figure shows that time & space complexity.


Perform operatation using PSNM
Performing PSNM
Perform operatation using Map Reduce
Performing Map Reduce
Timer Calculation For PSNM
Timer Calculation For Map Reduce

Total Minutes :0
Total Second:1
Total Milisecond:591

Total Minutes :0
Total Second:0
Total Milisecond:31

Fig. 2 Calculating Time For Duplicate Detection

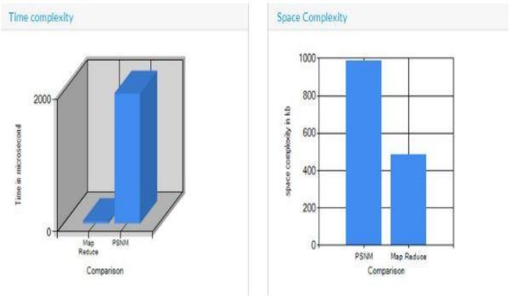


Fig.3 Comparison between PSNM & Map Reduce

VII.CONCLUSION

PPSNM and its utilization for duplicate record detection, and duplicate record deletion. On one hand, the extraction of PPSNM is faster than PSNM due to the Map Reduce concept. On the other hand, the improvement in detection effectiveness is consistently observed in two applications. This is achieved by indexing the PPSNM with Map Reduce.

ACKNOWLEDGEMENT

I would like to thank all people who help me in different way. Especially, I am thankful to my guide and P.G. coordinator **Prof. Amrit Priyadarshi** for his continuous support and guidance in my work. Also, I would like to thank H.O.D. of Computer Engineering Department, **Prof. S. S. Bere** for motivating me. Lastly, I thank to IJARCCE who have given opportunity to publish my paper.

REFERENCES

[1] Thorsten Papenbrock, Arvid Heise, and Felix Naumann,' Progressive Duplicate Detection' IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 25, no. 5, 2014.

[2] S. Yan, D. Lee, M. yen Kan, and C. L. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," in International Conference on Digital Libraries (ICDL), 2007.

[3] M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining and Knowledge Discovery, vol. 2, no. 1, 1998.

[4] X.Dong, A.Halevy, and J.Madhavan, "Reference reconciliation in complexinformation spaces," in Proceedings of the International Conference on Management of Data (SIGMOD), 2005.

[5] S.E.Whang, D.Marmaros, and H.Garcia-Molina, "Pay-as-you-go entity resolution" IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 25, no. 5, 2012.

[6] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicat record detection: A survey," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 19, no. 1, 2007.

[7] U.Draisbach, F.Naumann, S.Szott, and O. Wonneberg, "Adaptive windows for duplicate detection," in Proceedings of the International Conference on Data Engineering (ICDE), 2012.

[8] U.Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection." in International Conference on Data and Knowledge Engineering (ICDKE), 2011.

[9] L. Kolb, A. Thor, and E. Rahm, "Parallel sorted neighbourhood blockingwith mapreduce," in Proceedings of the Conference Datenbank system in Büro, Technik und Wissenschaft(BTW

DOI 10.17148/IJARCCE.2016.56147