ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

Design and Implementation of External Sorting as a Novel Tool for efficient processing of gas holders Details

[1] S. Hrushikesava Raju, [2] Dr.M.Nagabhushana Rao
Research Scholar, Rayalaseema University, Kurnool
Professor, Dept. of CSE, K L University, Vijayawada

Abstract:- **The importance of external sorting and its proposed methodology is discussed. The proposed methodology is refined in such a way that incurs few input, output costs, number of disk accesses, number of runs and other parameters compared to other existing methodologies. The proposed methodology is transformed as a tool with specific features in order to process huge gas details located to a particular area. The information stored on external devices requires external sorting technique which arranges the information in a particular manner according to the user option requirement. To do this, the features are assumed as facilities that may be used to process the information according to area selected. The results obtained using this novel tool is specified against existing literature methodologies. The comparison also be done among these approaches against this tool. This novel tool is applicable to the application in which gas holder's details are stored and this tool's features such as Esorting, results, and filtering to be applied over that data which output the reports to the user to use them efficiently**

Index Terms: **novel tool, external sorting, features, gas holder's, operations, E Sorting, Filtering**

I. INTRODUCTION

The real world applications consist of huge massive data. In order to process it, data preprocessing is required which transforms the given data set into refined data set. The refined data set means outliers or inconsistencies won't be present in the data set. The technique required called external sorting which sorts or arranges the elements in the huge data set in particular order. In case of numeric application or data set, arranging that given set is done in either ascending or descending order. In case of record oriented applications, arranging the records is done according to a particular attribute or element. There were many techniques used for external sorting. Each technique although finally performing external sorting and producing the result. When observed the methodology, they are not GUI based and are not user friendly and also their weak methodology consumes more time in terms of number of input, output operations, number of disk accesses, number of runs, and number of passes. Any external sorting technique is considered efficient when it consume least number of processing units in terms of number of input, output operations, number of disk accesses, number of runs, and number of passes. Here, the technique external sorting is taken as study. The few existing methodologies are studied and their drawbacks are listed in Introduction. Also, the need of proposed technique is discussed in Introduction and its working and appearance is discussed in Methodology. In this paper, the importance of transforming refined external sorting in to a tool is discussed along with the tool's specific features such as ESorting in case of arranging the huge number of records in a particular order, filtering in case of duplication of customer connections, and results in which final reports are stored.

II. RELATED WORK

The external sorting is useful in many places where huge data to be stored and that should be set right according to the requirement. There were few other methodologies used to implement external sorting which are such as greedy sorting, query based sorting, and sorting using lemmas. Each methodology has advantages and also overheads. But these overheads are complicate compared to proposed methodology. In the theoretical methodologies that exist such as Double Buffering, Key Sorting, Replacement Selection, B+-tree clustered and un-clustered indexing, traditional external sorting (Either K-way or Poly-phase Merge Sorting) pitfalls are notified and moreover the main drawback of these are not user friendly and are not GUI based featured

The following table lists the drawbacks of these above approaches:

Existing Methodology	Drawbacks
Double buffering	Additional buffer is maintained for each input and output buffer
Key Sorting	Each record associated the key cause expensive.
Replacement Selection	Involves many swapping from root with last node and then discards last node value, reconstructs heap until one element remains. Its Time complexity – O(n * log n)
Clustered B+ - tree	Sorts by traversing the leaf pages
Un-Clustered B+ - tree	Sorts by data records. Additional cost incurred for each page retrieved just once.
K-way external sorting / Poly-phase External Sorting	Takes more number of comparisons and more number of disk accesses when the redundant data exists.

Table 1: Disadvantages of existing external sorting approaches

In [1, 2, 3], the greedy sorting considers many disks. Those disks have data to be sorted stripped in to blocks. The disks should be synchronized while transferring the data. This approach is based on multiple disk model. This model also assumes certain lemmas as rules, and also has a predefined algorithm that outputs best runs. Merge sort is applied on those runs. This process is repeated until all runs are processed that results only one run, which contain all sorted data. The disadvantage of this approach is synchronizing all the disks, a lot of attention is required over disks working. It takes more resources into count in processing the external sorting. In [4], external sorting using lemmas starts sorting based on the three lemmas. This model based on a single disk although it consumes number of input and output operations in less in number. The lemmas description and working of this approach are briefed as follows:

Step1: Specific lemmas to be defined as rules to follow in doing the sorting

Step2: Provide the algorithm using buffers, priority queues, and make certain assumptions, Step3: Process the given data by following step2 and step1

Step4: Output the sorted data.

The disadvantage of this approach is it works strictly for a one disk model. Also, external factors considered such as buffers, priority queues, and additional entities consume more space although the input and output disk readings and writings are minimized. In [5,6,7], External sorting based on query evaluation is applied on the database. It results output after the query is imposed in SQL or equivalent Query Language. It is not suitable for arranging the items in particular order. It takes the query as input for the database which contains the information in terms of tables. Based on the criteria, the query is framed and obtains the result which is a subset of the given dataset from the given database itself. In this, three access methods are used such as indexing, iteration, and partitioning where indexing takes hash or tree indices. The operator evaluation brings output from a query. The kinds of operators involved are selection, projection, join, group by and order by. The functioning of this system can be shows flow of activities such as query parser, query optimizer, query plan evaluator in processing its functionality. The given query taken up by the query parser which translates it into Q-opt plan and that plan is made best in terms of cost, and time. This model is dependent on database language. In [8,9], NEXSORT is purely invented for XML documents to sort. It is applied over internet also in sorting XML files. This tool is developed to sort XML files rather than flat files. The time NEXSORT takes is less compared to sort a flat file. This assumes certain criteria to sort the XML document. It sorts based on hierarchical structure of the XML document which consists of intermediate elements and leaf elements. The disadvantage of this is it is only applicable to the file hierarchical based files but to other kind of text files

Existing Methodology	Drawbacks
Greedy Sorting	It is a multiple disk model requires synchronization of all the disks while sorting, and requires more resources for sorting.
Sorting using Lemmas	Consume more space to hold external entities and is strictly single disk model.
Query based	It is query based language. It

Sorting	depends on backend supporting language like SQL etc. Its drawback is result returned is a subset of given data set. It can’t produce result whose size is same as actual data size.
NEXSORT	It is purely applicable to XML hierarchical structured based files. It is not supporting sorting of other kind of files.

Table 2: Disadvantages of other existing external sorting approaches

All the above discussed existing external sorting approaches are not GUI based. The impact of this is end user is not comfortable to run and debug these programs written in a variety of programming languages. The end user has to learn the suitable environment required to run the appropriate approach. It leads headache to the end user. To avoid all these, the trend looks forward to design a GUI based External Sorting. This is nothing but a automated tool taken for the specific real time application “processing gas holders details”. Moreover the existing methodologies are confined and are flexible to implement any real time application and its features. But the proposed GUI based External Sorting is efficient and flexible to implement the features over the appropriate application called “processing gas holder’s details and automatic preserving of unique gas connection in duplicates found”. Hence, now the study goes over understanding of the novel tool methodology and its features. Not only for processing gas holders details but also for another application in which to process PAN Card details and securing unique card details by phishing multiple cards owned by an individual.

III. PROPOSED METHODOLOGY

In this, data preprocessing is taken as first step in case of more redundancy. Once redundancy is identified, data preprocessing is applied that produces the clean data (won’t contain any redundancy). This is the process carried out in first phase. The second phase uses the tool in which few features are provided as operations and filtering is performed whenever duplication of gas holder’s details are found in gas holder applications.

The background working of external sorting is as follows:

Pseudo_procedure externalsorting(dataset[][]):

N=number of tapes or pages

Step1: divide the data on to the input tape or page.

Step2: Using varying run size, by default 3*N on first input tape1, 3*(N-1) on second input tape2, 3*(n-2) on third input tape3, and so on until last input tape run size 3. Sort the runs data using merge sorting concept and place on output tape.

Step3: Take the varying run size concept as key, first runs of varying size from each output tape is taken, and merge sort applied among them, the resulted sorted data can be stored on first input tape. The second runs from each output tape is taken, and merge sorting is applied among them, the resulted sorted data can be stored on second input tape,… and so on until last runs of varying size are taken, and their sorted data can be written on last input tape. In all these scenarios, the each tape run size is going to vary from one pass to another pass.

Step4: Repeat Step3 for rest of data on the input and output tapes alternatively until all tapes data brought on to one output tape.

This procedure is made clear through an example in results column.

The above procedure takes less time than traditional method time consumed. The importance of this external sorting is came to be known when this can be converted into a tool. Use that developed tool on the specific application namely "processing gas holder details". This ideology is used in back ground side of this GUI based methodology. The GUI functionality is defined as follows:

```
pseudo_procedure GUI_Externalsorting(Dataset[][]):
Step1: Select one attribute in the list of record attributes.
Step2: Apply External sorting on that attribute, call Esorting feature
Step3: result produced is sorted set of records
```

Any pseudo procedure or algorithm when transformed into a flow chart gives fruitful message to the end user.

The flowchart of this GUI based External sorting is as follows:

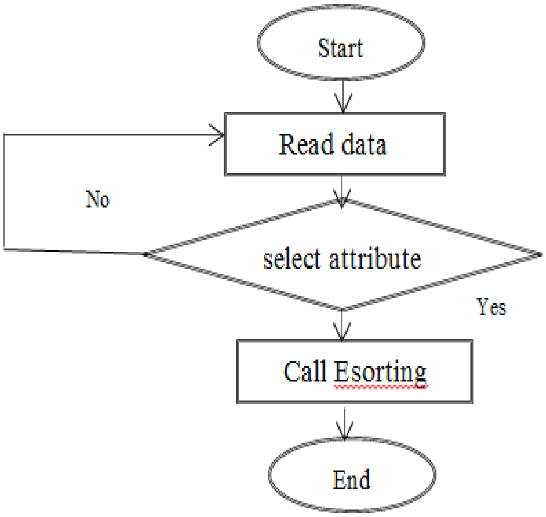


Fig.1 - GUI based Methodology: Novel External Sorting

The GUI based tool's exceptional features Esorting and filtering are defined as follows: ESorting feature Working

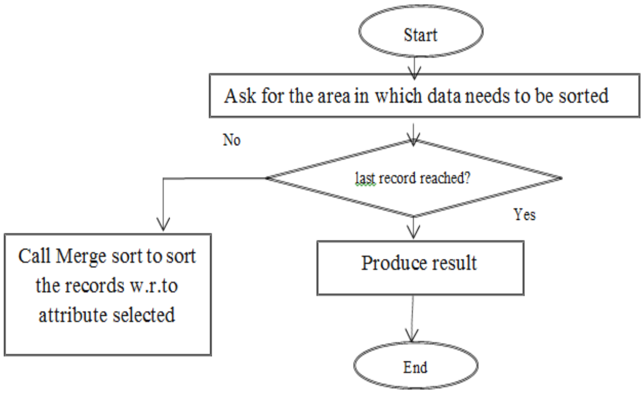

Start
Ask for the area in which data needs to be sorted
No
last record reached?
Yes
Call Merge sort to sort the records w.r.to attribute selected
Produce result
End

Fig.2 - Feature ESorting

Filtering feature Working is as follows:

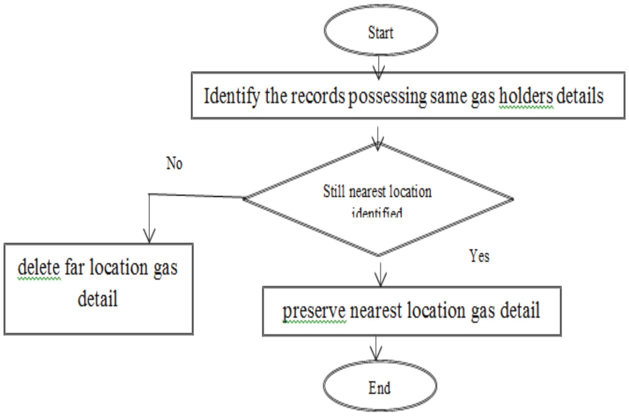

Start
Identify the records possessing same gas holders details
No
Still nearest location identified
Yes
delete far location gas detail
preserve nearest location gas detail
End

Fig.3 - Feature Filtering

In ESorting feature, The tool asks for the attribute or column name based on which records to be sorted. After the attribute is selected, the records to be sorted according that attribute in ascending order. The working of ESorting is depicted in Fig.2. In Filtering feature, The tool searches for duplicate connections in a particular area, listing of customers along with multiple connections, preserving nearest location gas connection of the customer using GPRS feature, and far location connections of that customers who have multiple connections are removed automatically, and updated unique list of customers in that selected area are outputted in the results feature. The working of filtering feature is depicted in Fig.3.

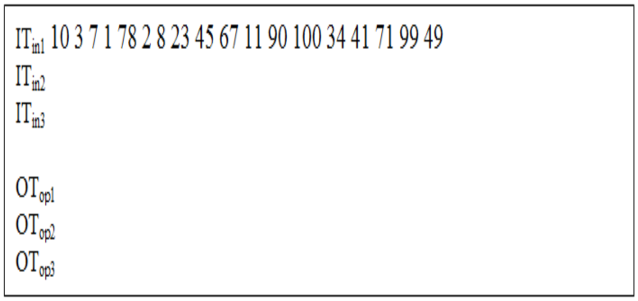
The result feature displays modified report after features updated.

IV. RESULTS

The proposed external sorting as background structure using varying run size (refined methodology) works as follows on the above data problem.

Case 1: Applications possessing numerical data:

Initial run construction pass: This takes all the data onto one input tape.



IT_{in1} 10 3 7 1 78 2 8 23 45 67 11 90 100 34 41 71 99 49

IT_{in2}

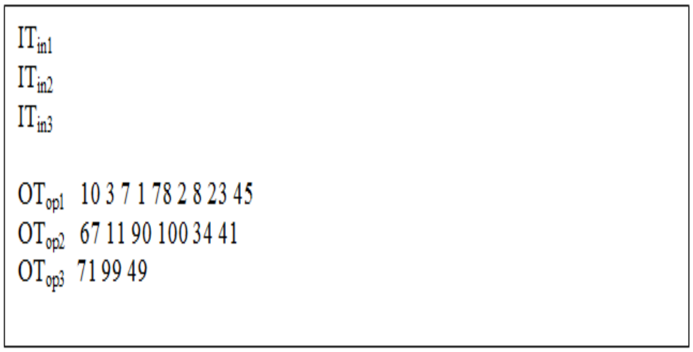
IT_{in3}

OT_{op1}

OT_{op2}

OT_{op3}

Step1: 3 tapes assumed. Hence, the first output tape run size is 3 * no. of tapes=3*3=9. The data to be read from the given set for the first output tape is 9. After removing redundancy from the given data set 10 3 3 7 1 1 1 78 2 2 2 2 3 3 3 10 8 23 45 67 11 90 100 23 34 41 71 99 49 34 45, the data set remains is 10 3 7 1 78 2 8 23 45 67 11 90 100 34 41 71 99 49. The size of this set is 15. For first iteration, First output tape takes run size 9, second input tape takes 6, and third input tape takes 3 according to refined methodology. For second time, storing of data from output tapes to input tapes takes varying run size.



IT_{in1}

IT_{in2}

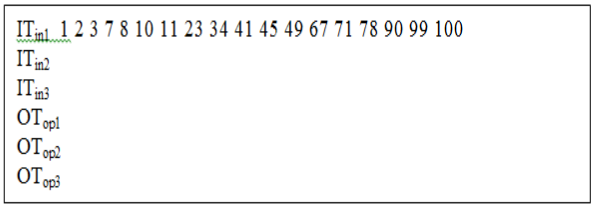
IT_{in3}

OT_{op1} 10 3 7 1 78 2 8 23 45

OT_{op2} 67 11 90 100 34 41

OT_{op3} 71 99 49

Step2: Data from the output pages / tapes can be read, apply merge sorting, and result written on input tapes alternatively. When the whole sorted data resulted on single input tape, by default take this sorted data on to first output tape for producing. This won’t take any separate pass. All output tapes first runs are sorted using merge sort and brought the result onto input tape. If the data on the first input tape is sorted using merge sorting. The sorted data set to be written on the first output tape.



IT_{in1} 1 2 3 7 8 10 11 23 34 41 45 49 67 71 78 90 99 100

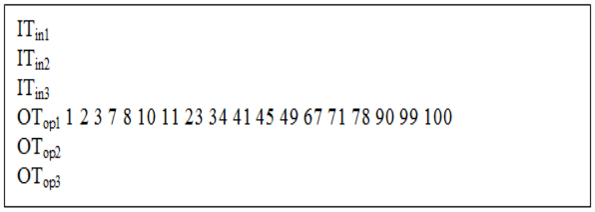
IT_{in2}

IT_{in3}

OT_{op1}

OT_{op2}

OT_{op3}



IT_{in1}

IT_{in2}

IT_{in3}

OT_{op1} 1 2 3 7 8 10 11 23 34 41 45 49 67 71 78 90 99 100

OT_{op2}

OT_{op3}

For the data without redundancy, two passes only are involved to sort the data on the external device. The following is the time consumed by the proposed background external sorting over traditional external sorting:

Factor	**Sorting using traditional sorting**	**Sorting using refined methodology**
No. of disk accesses / No. of Input and output costs	=2 * 18 * (log $_3$ 18/3 + 1) = 2*18* 3 = 108 accesses	=2 * 18 = 36 accesses
Number of runs	9	4
Number of passes	3	2

The following is the graph that denotes performance of the refined external sorting and traditional external sorting:

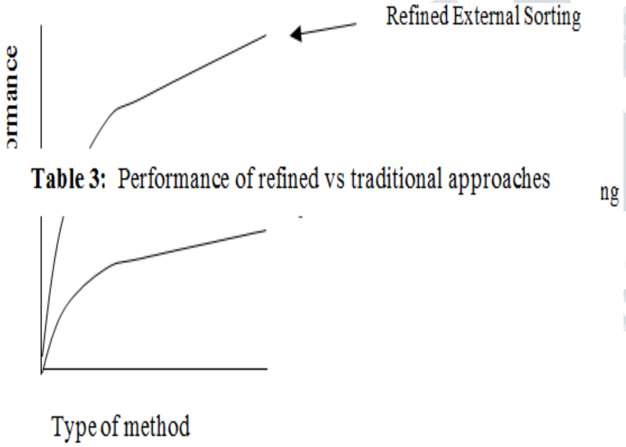


Table 3: Performance of refined vs traditional approaches

Fig.4: Performance of refined vs traditional approaches

The case 1 deals with processing only numeric type data or single attribute type.

Case 2: Applications possessing data in records:

By taking this methodology as GUI based approach in which how Esorting and filtering features are processed as follows through the screen shots: In case of unregistered customer, user has to click on registration link. Later, user details to be posted in the following format.

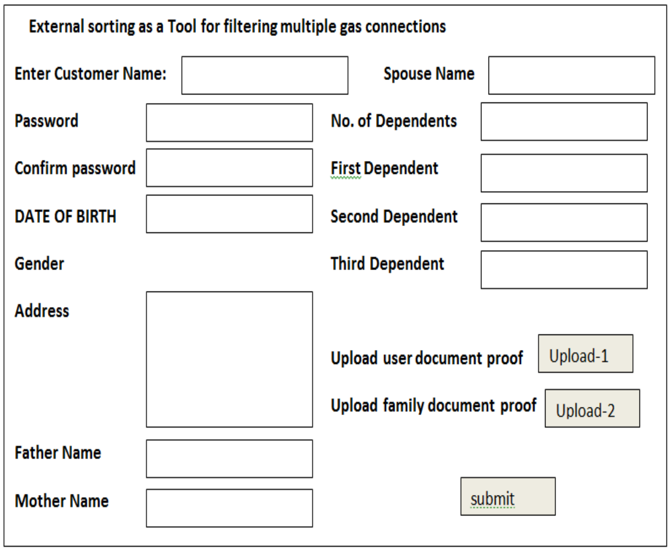
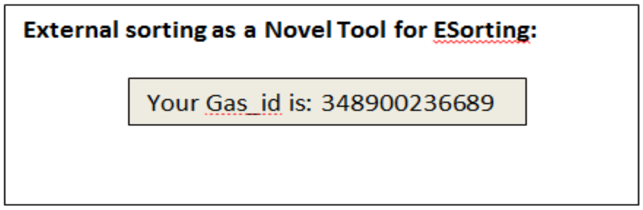

External sorting as a Tool for filtering multiple gas connections

Fig. 5: registration

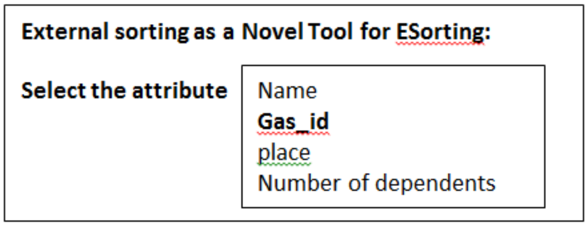
The random number generator having a particular size is used for displaying the unique gas identifier


External sorting as a Novel Tool for ESorting:

Your Gas_id is: 348900236689

Fig.6: gas id generated randomly

Based on size of gas id, the random number generator will generate unique number to the customer. The case 2 deals with nonnumeric type applications involving record based files that consist of data in terms of attributes or fields. The advantage of this case 2 type is sorting can also be performed using two or more attributes. Processing of ESorting feature: This feature allows the external sorting to be implemented on the record set based on the selected attribut. For selecting the attribute, the following snapshot is considred as an example**.**



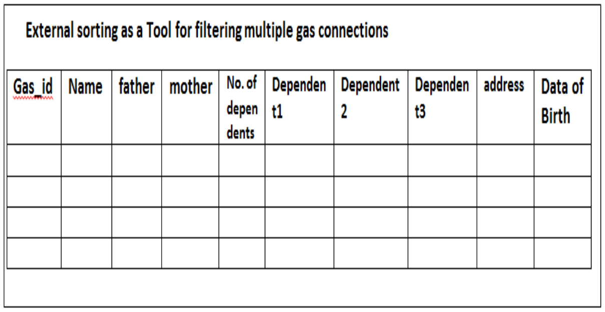
External sorting as a Novel Tool for ESorting:

Select the attribute

Name
Gas_id
place
Number of dependents

Fig.7: gas id random generation

The Next screen is obtained as follows if the gas_id is the selected attribute



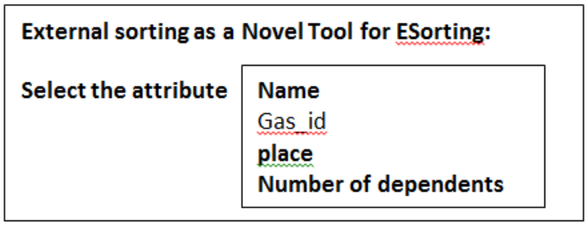
External sorting as a Tool for filtering multiple gas connections

Gas_id	Name	father	mother	No. of dependents	Dependent1	Dependent 2	Dependent3	address	Data of Birth

Fig. 8: displaying of results

The external sorting based on attribute is done including the remaining column set of that attribute. The advantage of this ESorting is processing external sorting based on not only single attribute but also on two or more attributes (multiple attributes). For example, gas holders details need to be sorted using attributes Name, place, and No. of dependents. Here, three attributes are considered in processing the gas holders details in one area.

The following is the snapshot of this scenario:

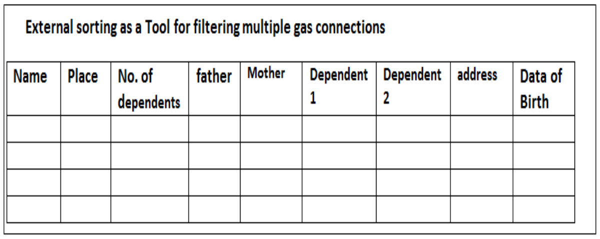


External sorting as a Novel Tool for ESorting:

Select the attribute

Name
Gas_id
place
Number of dependents

Fig. 9: selecting multiple attributes to perform ESorting

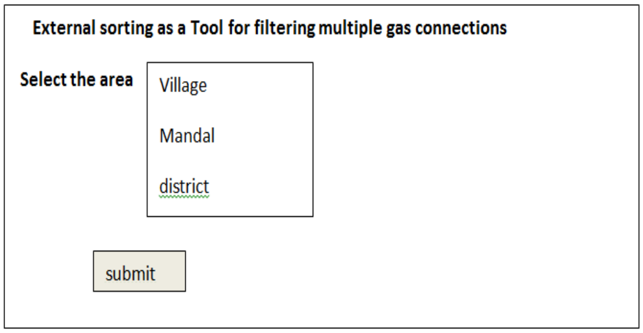


External sorting as a Tool for filtering multiple gas connections

Name	Place	No. of dependents	father	Mother	Dependent 1	Dependent 2	address	Data of Birth

Fig. 10: External sorting based on attributes Name, place, and Number of dependents

Processing of Filtering feature: To filter multiple gas connections, the area to be selected. After selection of the area, gas holder details are listed. By observing who are the customers having multiple connections, the filter feature to be clicked in order to remove far connection gas details.



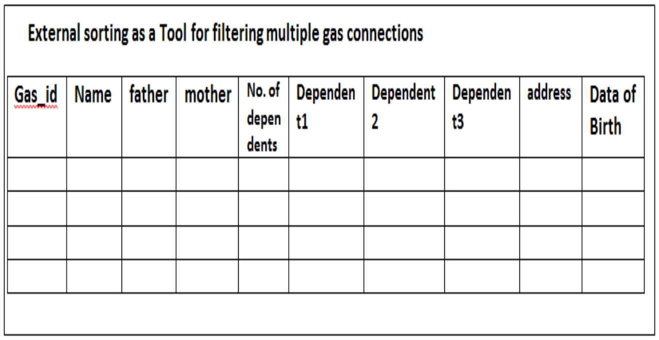
External sorting as a Tool for filtering multiple gas connections

Select the area

Village
Mandal
district

Fig. 11: selecting which area in which to perform filtering

The Next screen is obtained as follows:

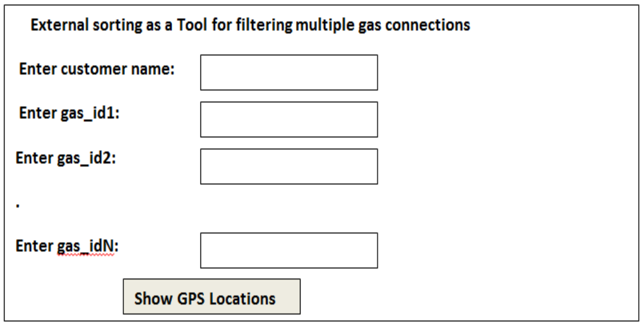


External sorting as a Tool for filtering multiple gas connections

Gas_id	Name	father	mother	No. of dependents	Dependent1	Dependent 2	Dependent3	address	Data of Birth

Fig. 12: displaying of results

From this, the admin analyze the details. Based on the address, the nearest location gad id is fetched using GPS. The far location connections are deleted in the filtering operation. The next screen is as follows: It asks to perform filtering operation.


External sorting as a Tool for filtering multiple gas connections

Enter customer name:

Enter gas_id1:

Enter gas_id2:

.

Enter gas_idN:

Show GPS Locations

Fig.13: filtering

The next screen is : It shows GPS locations of the customer having multiple gas connections with respect to the customer address, and dependents. The nearest location of gas agency connection is preserved. The rest of the connections are taken automatically in the fields mentioned in ninth screen

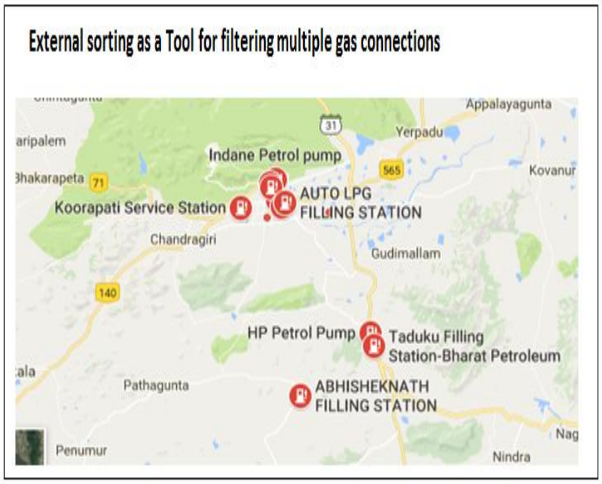
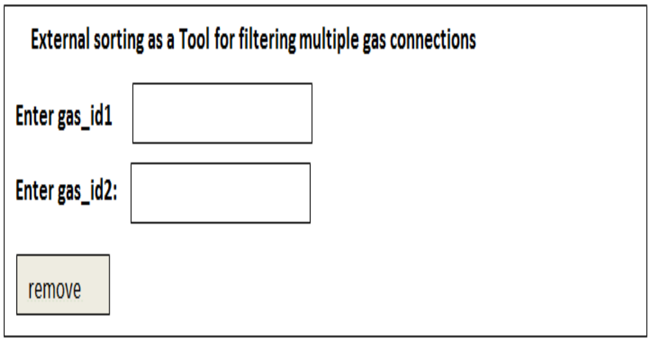

External sorting as a Tool for filtering multiple gas connections

Fig.13: GPS locations are displayed w.r.to multiple gas details

IV. CONCLUSION

Accordingly, the additional gas connections are specified by seeing the map, from which customer address and agency centers addresses are identified. The far different agencies are taken in to the account, only nearest agency center connection is preserved


External sorting as a Tool for filtering multiple gas connections

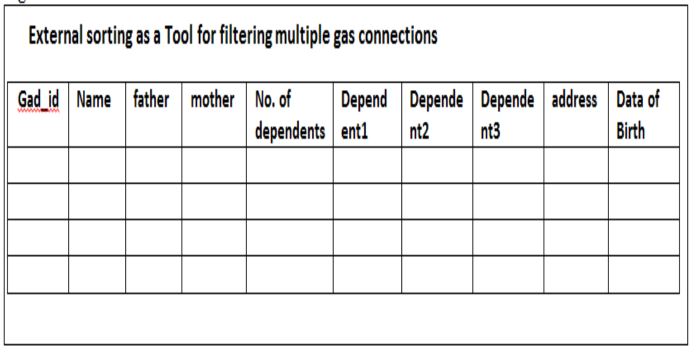
Enter gas_id1

Enter gas_id2:

remove

Fig.14: additional gas details are automatic loading in order to remove

The next screen shows unique gas connections of the customers in which eight screen having users having multiple connections is refined and twelveth screen is outputted having unique gas connections


External sorting as a Tool for filtering multiple gas connections

Gad_id	Name	father	mother	No. of dependents	Dependent1	Dependent2	Dependent3	address	Data of Birth

Fig. 15: unique gas details are displayed under particular area

The above features when performed achieves expected results in less time because of using refined external sorting using varying run size. The expected benefits achieved are user friendly, and look and feel.

In case of results feature: All these outputted screen snapshots are stored. The admin and users of this tool came to know that kind of sequence of operations performed at what time and their consequences are noted. Generally, the ESorting final reports and filtering final reports to be stored in the results feature.

This Novel Tool is GUI based environment which is easy to use by the end user. The background methodology used is refined external sorting in which first records are taken in output page or tapes, sort those records using varying run size concept and store in input tapes or pages using varying run size. This process is repeated until all the records are brought on to one page or tape. When sorting, it considers remaining column set in that record, preserves that data. The refined sorting produces the records having the selected attribute in sorted order along with their original column set.

V. CONCLUSION

The proposed methodology GUI based External Sorting as a Tool compared to specific existing methodologies is GUI based. The proposed GUI based methodology processes the given data using its operations such as Input, preprocess, Esorting, filtering and result. The Novel Tool performs Esorting and filtering operations in less time. In this, the fearure preprocess is selected in case of redundancy present in the data set and Esorting is applied which allows to sort based on more than one attribute and result feature produces sorted data in noted less time. The appearance of this GUI based Novel tool is user friendly and experiences best look and feel. The GUI based external sorting novel tool another feature "filtering" removes all unnecessary gas details of a particular area and preserves nearest location gas detail of the customers.

REFERENCES

[1] Sergi Elizalde, Peter Wrinler,"A Greedy Sort Algorithm", Rutgers Experimental Mathematical Seminar,

[2] Kevin Wayne,"Greedy Algorithms I", Pearson-Addision Wesley,2013

[3] "Chapter 16: Greedy Algorithms", https://www.cs.rochester.edu/~gildea/csc282 /slides/C16-greedy.pdf

[4] Fang Cheng Leu,Yin Te Tsai, Chuan Yi Tang, " An efficient external sorting algorithm", May 2000, Information Processing Letters 75 (2000) 159–163.

[5] Jignesh M. Patel, "External Sorting", Spring 2017, CS 564: Database Management Systems; (c), 2013

[6] G.Graefe, "External Sorting and Query Evaluation", ACM Computing Surveys 25(2), Faloutsos.

[7] "Query Processing: The basics", http://www3.cs.stonybrook.edu/~sas/courses/cse305 /lectures/ch10.pdf

[8] Adam Silberstein, Jun Yang, "NEXSORT:Sorting XML in External Memory", http://db.cs.duke.edu/papers/2004-ICDE-sy-nexsort.pdf, 2004

[9] Silberstein and J. Yang, "NeXSort: Sorting XML in external memory", Technical report, Duke University, July 2003.http://www.cs.duke.edu/dbgroup/papers/2003-sy-nexsort.pdf.

[10] Mark Allen Weiss, "Chapter7:Data Structures and Algorithm Analysis in C++".

[11] Mark Allen Weiss , Chapter7: Data Structures and Algorithm Analysis in Java.

[12] Alfred V. Aho, John E. HopCroft and Jelfrey D. Ullman, "Sorting, Data Structures and Algorithms", Addison –Wesley, 1983.

[13] Micheline Kamber and Jiawei Han, "Data Preprocessing: Data Mining Principles and Techniques"

[14] Margaret H Dunham, Data Mining Introductory and Advanced Topics, Pearson Education, 2e, 2006.

[15] Sam Anahory and Dennis Murry, "Data Warehousing in the Real World", Pearson Education, 2003.

[16] D. E. Knuth (1985), Sorting and Searching, The Art of Computer Programming, Vol. 3, Addison –Wesley, Reading, MA, (1985).

[17] Alok Aggarwal and Jeffrey Scott Vitter, "Input and Output Complexity of Sorting and related problems", Algorithms and Data Structures, AV88.pdf.

[18] Ian H. Witten, Eibe Frank, Morgan Kaufmann, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition (Morgan Kaufmann Series in Data Management Systems), , 2005.

[19] Zhi – Hua Zhou, "Introduction to Data Mining", part3: Data Preprocessing, , Dept. of CSE, Nanjing University, Spring 2012, Pt03.pdf.

[20] Chiara Rebso, "Introduction to Data Mining: Data Preprocessing", KDD- LAB, ISTI – CNR, Pisa, Italy, data.ppt.pdf.

[21] "Data Mining: Data Preparation", www.cs.nyu.edu/courses/ spring08/G22.3033-003 /2prep.ppt.

[22] "Data Preprocessing", grid.cs.gsu.edu/~cscyqz/courses/ dm/slides/ch02.ppt

[23] Hrushikesava Raju S. , Nagabhushana Rao M.," Improvement of Time Complexity on Pattern Matching using One -Time Look Indexing and Data Preprocessing", International Journal of computer sciences and Engineering, Vol.4(11), pp.82-86, 2016, E-ISSN:2347-2693.

[24] Hrushikesava Raju S.,Swarna Latha T. ,"Dynamic Pattern Matching: Efficient Pattern Matching using Data Preprocessing with help of One time look indexing method", International Journal of Advanced Research in Computer Engineering and Technology,Vol.2(2),592-599,2013,ISSN:2278-1323.

[25] Hrushikesava Raju S. , Nagabhushana Rao M., "Pattern Matching Using Data Preproc-Essing With The Help Of One Time Look Indexing Method", International Journal of Pharmacy and Technology, Vol.8(3),pp.18395-18407,2016, ISSN:0975-766X.

[26] Hrushikesava Raju S. , Nagabhushana Rao M.,"A Review on Specific Data Structures Using Data Preprocessing and Refinement of Existing Algorithms in Order to Improve Time Complexities",International Journal of computer sciences and Engineering, Vol.4(9),pp.146-151,2016.

[27] HweeHwa Pang , Michael Carey J. , Miron Livny ," Memory-Adaptive External Sorting", pp.1-12. Proceedings of the 19th VLDB Conference Dublin, Ireland, 1993.

[28] John Yiannis, Justin Zobel, "Compression Techniques for Fast External Sorting", Proc. British National Conference on Databases, A.James (ed),Coventry, UK, pp.115-130, July2003.

[29] Young Sik Lee, Luis Cavazos Quero, Youngjae Lee,Jin-Soo Kim, Seungryoul Maeng, "Accelerating External Sorting via On-the-fly Data Merge in Active SSDs", pp.1-5,2014.

[30] Cormen Th. H., Leiserson Ch. E., Rivest R. L., and Stein C.,"Introduction to Algorithms", 2nd edition, MIT Press ,2001.

[31] Knuth D. E., "The Art of Computer Programming: Volume 3 (Sorting and Searching)",2nd edition, Addison-Wesley ,1998.

[32] Garcia Molina H., Ullman J. D., and Widom J., “Database Systems: The complete Book”,International edition, Prentice Hall , 2002.

[33] Ramakrishnan R.,Gehrke J.,”Database Management Systems”, Second Edition,pp.301-315,1997.

About Authors:

Mr. S. HrushiKesava Raju, working as a Professor in the Dept. of CSE, SIETK, Narayanavanam Road, Puttur. He is pursuing Ph.D from Rayalaseema University. His other areas of interest are Data Mining, Data Structures, and Networks.

Dr. M.Nagabhushana Rao, working as Professor in the Dept. of CSE, K.L.University, Vijayawada,A.P. He had completed Ph.D from S.V. University in the area of Data mining. He is presently guiding many research scholars in various disciplines.