ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

A Survey of Three Different Data Structures to Manage Mega Datasets in DBMS

[I]Minoo Khosravani Dehkordi, [II]Lawal Adamu, [III]Umapathy Eganathan
[I,II]Master Student, University of Putra Malaysia, Malaysia
[III]Research Scholar in CS, Vels University-Chennai, India

Abstract
Metadata is the most important concept in database management system as well as in order to manage mega dataset effectively and efficiently here implied three difference data structures. Managing of mega dataset is a crucial and effective approach in data base systems and further it exists in fundamental approaches are customer review, business transactions and other commerce activities. There are various approaches have been proposed. In this paper, the authors employ the Hash Table structure, Data Linked structure in order to manage datasets and also we have suggested combination of these two well-known data structure is the third structure to improve efficiency of data management. This tool allows the users can frequently do searching, inserting, and deleting operations rapidly with easiest way. After implementation process here used a mega text file of costumer review to evaluate performance of those structures. The experimental results shown that the data structure outperforms significantly in comparison with others in several aspects.

Keywords
Metadata, Database, Linked List, Hash Table, Data Structure

I. Introduction

A data set is a named collection of data that structured (formatted) in a specific construction and managed by specific modules that is based on the data set organization. Datasets are most commonly found in tables, spreadsheets and databases. Datasets are often created and managed to provide information of an organization's body. These are necessary for business continuity, accountability, and evidence based decisions such as costumer review and so on. A customer review [1] is the information of a product or service made by customers according to their experiments and obtained results of using a product or service. In the other words, customer review is coming directly from customers about the satisfaction or dissatisfaction they feel with a product or a service. Over time, these reviews form a large amount of data that sellers, competitors and producers can use them to extract very useful information in order to identify new product bundles, finding the root cause of manufacturing problems, to prevent customer attrition and acquire new customers, cross-sell to existing customers, and profile customers with more accuracy. Hence, analyzing and manipulating these data becomes an important issue in computing world.

The benefits promised by customer review have motivated researchers to suggest a lot of proposal in this area. The overall goal of them is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it falls into business data mining categories [2] and involves database, data management aspects, data pre-processing, and data structure. As discussed above, professionals are trained to analyze and interpret data, but the increases in data amount, data type, and analytical dimensions, they have gone beyond storage, transmission, and processing a huge dataset. These huge data need to be converted into information and knowledge in order to support decision making. For this reason, choosing correct data structure is the most important task.

In this study, this paper intend to examine three structures called the Hash Table structure, data linked structure in order to manage datasets. Also we have suggested combination of these two well-known data structure to improve efficiency of management. This tool allows the users to search, insert, and delete. The remaining sections of this paper are organized as follows: From the following some of the related study will briefly explain about the overview of the Customer Review, and in further way it described how to apply the mentioned structures to mange mega files of costumer review. At last the obtained numerical results will be presented as a comparison and finally ended with conclusions.

II. Related study

Customer review has a large family composed of various structure and algorithms and the scope is still expanding because researchers devote to improve the efficiency and accuracy of the existed algorithms. Most researches in customer feedback and business data mining area focus on improving efficiency and accuracy of single business decision and application. Fewer efforts are devoted into the discussion of applicability and fitness.

To manage the dataset of customer feedback, there are many techniques available in which some of them are as follows:

Association rule learning [19] that Agarwal, Imielinski proposed in 1993, are a well studied method to discover interesting relations between variables/features in large dataset. It is intended to identify strong rules that can be effective in uncovering unknown relationships, and provide some results [18].

Document clustering [20] is a process that automatically groups a set of documents into subgroups, called clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis.

Statistical Classification [22] is an important data mining technique that the researchers use it for many fields including customer feedback, business application, and so on. This technique involves extracting interesting patterns representing knowledge from large real-world databases.

Text summarization [21] is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document.

Link analysis is based on a branch of mathematics called graph theory and it has yielded promising results in solving some problems such as analyzing customer feedback, telephone call patterns, and so on.There are a huge amount of aspects and product

feature extraction tools in order to manage the costumer review [3, 4, 5, 6, 7, 8, 9, and 10], which we summarized and discussed some of them in the following:

The earliest attempt at aspect detection was based on the classic information extraction approach of using frequently occurring noun phrases presented by Hu and Liu [11]. Their work can be considered as the initiator work on aspect extraction from reviews. Popescu and Etzioni [12] developed an unsupervised information extraction system called OPINE. Yi et al. [13] developed a set of aspect candidate extraction heuristics for extracting an aspect from product reviews based on the observation that aspect terms are nouns, they extract only noun phrases from documents and apply two feature selection algorithms, mixture language model [14] and likelihood ratio [15].

Somprasertsri and Lalitrojwong's [21] proposed a supervised model for aspect detection by combining lexical and syntactic features with a maximum entropy technique.

Wei et al. [16] proposed a semantic-based product aspect extraction (SPE) method. The SPE technique employs the same pruning rules as proposed in [11] in the pruning step to produce frequent product aspects from the set of candidate aspects. Finally, Zhu et al. [10] developed an aspect-based unsupervised opinion polling system. In their work, a multi-aspect boots trapping method based on RlogF metric [17] and an ambiguity degree is proposed to learn aspect-related terms for each aspect to be used for aspect identification.

III. Methodology

In order to implement customer review analysis, the procedure used here is to sort of data structures in order to store and manage dataset. There are so many different data structures which can be used here. However, choosing right data structure is a critical task in this area, because the performance of the data structure directly affects the performance of the program. In this task the taken data may be from 5 million records of costumer review, so using usual simple structure like array or queue is not suitable.

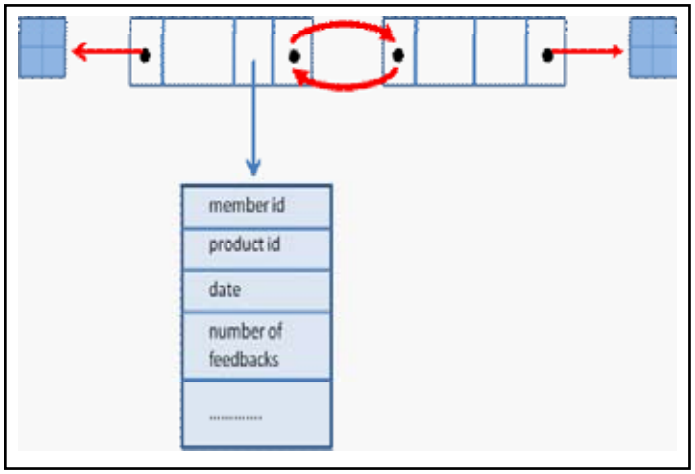
The major steps of this research are twofold. First of all need to design a tool to search, delete, and insert via using Hash table and Linked list separately. In the following it can be illustrated each of them.

A. Linked List

In computing, a linked list [23] is a data structure consisting of a group of nodes which together represent a sequence. Under the simplest form, each node is composed of a data and a reference (in other words, a link) to the next node in the sequence; more complex variants add additional links. This structure allows for efficient insertion or removal of elements from any position in the sequence. In general there are different types of linked list available in data structure. Namely single linked list, double linked list and so on. Single linked list can be connected with another single node. Double linked list consist front and rear. Following table indicate comparison of data structure an shows that linked list is more optimum than others:

	Linked list	Array	Dynamic array	Balanced tree	Random access list
Indexing	Θ(*n*)	Θ(1)	Θ(1)	Θ(log n)	Θ(log n)
Insert/delete at beginning	Θ(1)	N/A	Θ(*n*)	Θ(log n)	Θ(1)
Insert/delete at end	Θ(*n*) when last element is unknown; Θ(1) when last element is known	N/A	Θ(1) amortized	Θ(log *n*)	Θ(log *n*) updating
Insert/delete in middle	search time + Θ(1)[1][2][3]	N/A	Θ(*n*)	Θ(log *n*)	Θ(log *n*) updating
Wasted space (average)	Θ(*n*)	0	Θ(*n*)[4]	Θ(*n*)	Θ(*n*)

In this paper the proposed method is double linked list which each node contains the next-node link, data field, and the previous node link. Data field is containing information and a pointer that link to an array. This array is including member id, product id, date, number of helpful feedbacks, and number of feedbacks, rating, and so on. The primary figure of our proposed method is as follow:



In order to imply the linked list on datasets and reach the goal here authors have created two classes: 1- linked list Item class, and 2- linked list class. Linked list Item class indicates the each element of the list. This items contain, Previous, Next, Line number, and data. Linked list class executes all operations on dataset such as search, delete, insert, modify.

B. Hash Table

In computing, a hash table (hash map) is a data structure used to implement an associative array, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the correct value can be found. Ideally, the hash function will assign each key to a unique bucket, but this situation is rarely achievable in practice (usually some keys will hash to the same bucket). Instead, most hash table designs assume that hash collisions - different keys that are assigned by the hash function to the same bucket - will occur and must be accommodated in some way. In many situations, hash tables turn out to be more efficient than search trees or any other table lookup structure. For this reason, they are widely used in many kinds of computer software, particularly for associative arrays, database indexing, caches, and sets. In this paper here the authors suggested the Hash list function of Visual basic.

Second step is the implementation is combination of Hash Table and Linked list. It means each element of Hash table includes a

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

key and linked list. In the other words, the complete review of a customer into an element of link list. Then, by using an array consist of all member IDs assigned each cell of member ID array to its corresponding review link list. Method of the reading file is line by line. The Evaluation of each structure will present in the following and then it will be shown the particular structure outperforms significantly in comparison with others in several aspects.

IV. Evaluations and Results

In this section, the authors investigated the performance of three mentioned structures. They were implemented in Visual Basic under Windows 8 and were run on an Intel(R) Core(TM) i5-3337U, 1.80 GHz PC, with 6 GB system memory. The files used 4096 MB of costumer review which has 305403 records. Each structure is performed 10 runs. Following table depicts the average of runs for each of the mentioned structures:

Table 1: Results of HT, LL, Com of HT & LL

Reading Time	Structure	Operation Time
45.80 s	Hash Function	32.40 s
	Linked List	45.95 s
	Combination of Hash Function and Linked List	47.12 s

Following figures show the run average of three structures:

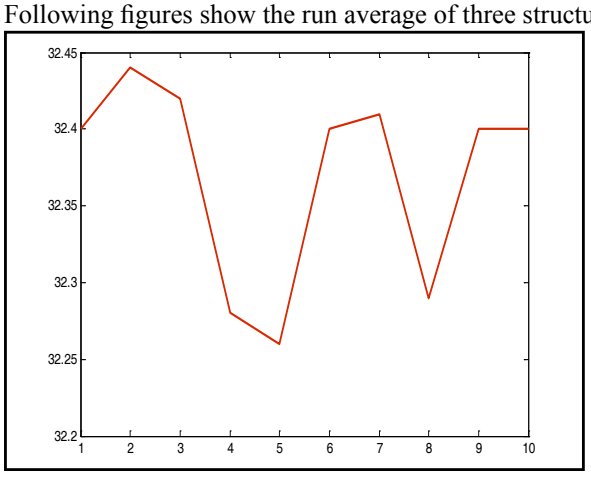


Fig. 3: The average of linked list runs.

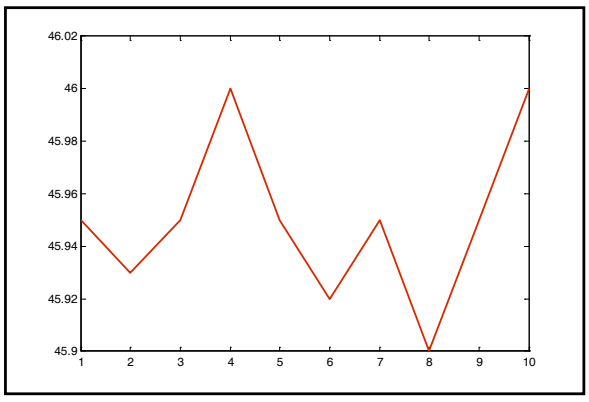


Fig. 4: The average history of Hash table.

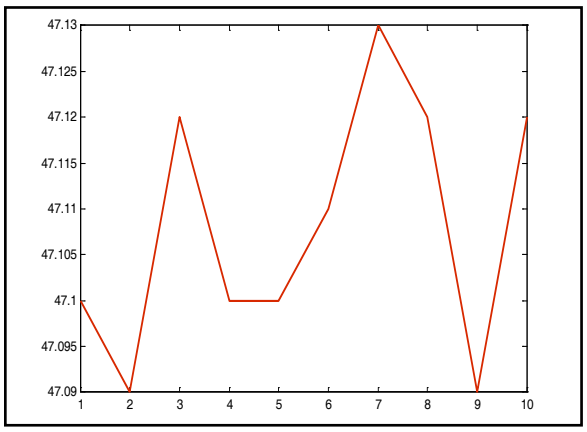


Fig.5 : The convergence history of combination of Hash Table and Linked list

An obvious observation from Table 1 is that Linked List is faster than others. It can perform all operations such as add, delete, modify in 32.40s. This is because the nature of linked list that has efficient insertion or removal elements technique.

V. Conclusion and future work

In this paper the study examined three structures called the Hash Table structure, data linked structure in order to manage datasets. Also the authors have suggested combination of these two well-known data structure to improve efficiency of management. The goal was to obtain the best and faster structure for managing huge data sets. An obvious observation from Table 1 is that Linked List is faster than others. As a future work, we are planning to extend linked list to make it faster and efficient for managing huge data set in some field like costumer review.

VI. Acknowledgment

Authors says heartfelt thanks to Dr.Jothi Sophia Principal, CSI JACON, Madurai, India for the sponsorship and continuous encouragement.

References

[1] *Ayoub Bagheri, Mohamad Saraee, Franciska de Jong, "Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews", Knowledge-Based Systems 52 (2013) 201–213.*

[2] *Jia-Lang Seng, T.C. Chen b, "An analytic approach to select data mining for business decision", Expert Systems with Applications 37 (2010) 8042–8057.*

[3] *A. Hogenboom, F. Boon, F. Frasincar, A statistical approach to star rating classification of sentiment, Management Intelligent Systems (2012) 251–260.*

[4] *S. Moghaddam, M. Ester, ILDA: interdependent LDA model for learning latent spects and their ratings from online product reviews, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Publishing, ACM, 2011, pp. 665–674.*

[5] *G. Qiu, B. Liu, J. Bu, C. Chen, "Opinion word expansion and target extraction through double propagation", Computational Linguistics 37 (1) (2011) 9-27.*

[6] *T.T. Thet, J.C. Na, C.S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards", Journal of Information Science 36 (6) (2010) 823–848.*

[7] C.P. Wei, Y.M. Chen, C.S. Yang, C.C. Yang, "Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews", Information Systems and E-Business Management 8 (2) (2010) 149–167.

[8] Z. Zhai, B. Liu, H. Xu, P. Jia, "Constrained LDA for grouping product features in opinion mining, in: Proceedings of 15th Pacific–Asia Conference, Advances in Knowledge Discovery and Data Mining", 2011, pp 448–459.

[9] T.J. Zhan, C.H. Li, "Semantic dependent word pairs generative model for finegrained product feature mining", Advances in Knowledge Discovery and Data Mining (2011) 460–475.

[10] J. Zhu, H. Wang, M. Zhu, B.K. Tsou, M. Ma,"Aspect-based opinion polling from customer reviews", IEEE Transactions on Affective Computing 2 (1) (2011) 37–49.

[11] M. Hu, B. Liu, "Mining opinion features in customer reviews", in: Proceedings of 19th National Conference on Artificial Intelligence, Publishing, AAAI Press, 2004, pp. 755–760.

[12] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, Publishing, Association for Computational Linguistics, 2005, pp. 339–346.

[13] J. Yi, T. Nasukawa, R. Bunescu, W. Niblack, Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, in: Proceedings of Third IEEE International Conference on Data Mining, Publishing, 2003, pp. 427–434.

[14] C. Zhai, J. Lafferty, Model-based feedback in the language modeling approach to information retrieval, in: Proceedings of 10th International Conference on Information and Knowledge Management, Publishing, 2001, pp. 403–410.

[15] T. Dunning,"Accurate methods for the statistics of surprise and coincidence", Computational Linguistics 19 (1) (1993) 61–74.

[16] G. Somprasertsri, P. Lalitrojwong,"Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features", in: Proceedings of IEEE International Conference on Information Reuse and Integration, Publishing, 2008, pp. 250–255.

[17] E. Riloff, R. Jones, Learning dictionaries for information extraction by multilevel bootstrapping, in: Proceedings of 16th National Conference on Artificial Intelligence, Publishing, John Wiley & Sons LTD, 1999, pp. 474–479.

[18] Y.-L. Chen, K. Tang, R.-J. Shen, Y.-H. Hu,"Market basket analysis in a multiple store environment", Decision Support Systems 40 (2005) 339–354.

[19] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in largedatabases, Proceedingsof the 1993 ACM SIGMOD International Conference on Management of Data, USA, Washington, DC, 1993, pp. 207–216.

[20] Dr. Sankar Rajagopal Enterprise DW/BI Consultant Tata Consultancy Services, Newark, DE, USA. Customer data clustering using data mining technique. Int J Database Manage Syst 2011; 3(4):1. http://dx.doi.org/10.5121/ijdms.2011.

[21] Jiaming Zhan, Han Tong Loh, Ying Liu," Gather customer concerns from online product reviews – A text summarization approach", Expert Systems with Applications 36 (2009) 2107–2115.

[22] Runyu Jing, Jing Sun, Yuelong Wang, Menglong Li, Xuemei Pu, "PML: A parallel machine learning toolbox for data classification and regression", Expert Systems with Applications 37 (2010) 8042–8057.

[23] M. V. Wilkes, "Lists and why they are useful," in Proceeds of the ACM National Conference, Philadelphia, 1964.