Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.290 – 298

RESEARCH ARTICLE

Optimizing Query Performance with OLAP to Discovering the Diagnosis of Diabetes

L. Cynthiya Juliyet
Department of Computer Science
Bishop Heber College, Tiruchirappalli
Tamilnadu,India

Mr.K.Mohamed Amanullah
Asst.professor Department of Computer Science
Bishop Heber College, Tiruchirappalli
Tamilnadu,India

Abstract

Decision support system to provide On-Line Analytical Processing (OLAP) techniques are used to provide analysis of data. The major confusion is that when to use Database Queries in Data Mining and when to use On-line Analytical Processing (OLAP).The various purposes according to the retrieved information are used to the different requirements. The purpose of retrieved information might be in various users' behavior prediction or for the purpose of Decision Support System (DSS) for effective decision making predict patients who might be diagnosed with diabetes. This paper also provides a focused answer using historical data of concerned patients, with an emphasis on their new requirements focus on the step and we propose then the use of AK-Mode algorithm.

Keywords: Data mining, Diabetic Approach, Clustering, AK-mode algorithm.

Introduction

The data warehouse is becoming more and more important in terms of data considered to making the decision through their capacity to participate heterogeneous data from manifold information sources in a common storage space, for querying and analysis. so most of the people avoids going through the large volume of database because of the lack of time in now a days world. The quality of services is important to deliver the healthcare Industry faces strong pressures and also lower cost. Generally, data produced is high, fragmented, imperfect, inexact, in the incorrect place, or difficult to make sense [14]. A critical issue facing the industry is the lacks of convenient and timely information. These information retrieval approach allows to fetch the large number of database within compact time and in an easy format of the way the number of people chooses these techniques as a source of information retrieval techniques includes the Data Mining, Database Queries and On-line Analytical Processing (OLAP).

The idea is to construct the Data warehouse strategy from OLAP requirements. According to [5] & [6] OLAP systems have rapidly gained momentum in both the academic and research communities, mainly due to their quick and multidimensional analysis capabilities. In order to facilitate this task propose the use of clustering as a data mining technique to group the different schemas resulting from the process of transforming the requirements.

OLAP

Mostly OLAP (On-line Analytical Processing) Consist of:

- Brief the data before it is possible to execute the queries.
- Summarization can be represented as cubes and sub cubes.
- Cube is pre-calculated and pre-aggregated data.
- It allows reporting data, visualizing data and interaction with views of data.
- It uses the star schema, snowflake plan which consists in fact table and dimension table.

Literature Review

The purpose of this literature review is to introduce and identify the limitations of automatic schema generation process by the other researchers. Our focus would be on the use of hierarchical clustering to automate the process of OLAP schema generation. The review is categorized over three major themes: (1) combining OLAP with data mining, (2) use of hierarchical clustering with OLAP, and (3) automation in OLAP schema generation.

RupaBagdi et al [1] developed a decision support system which combined the strengths of both OLAP and data mining. This system would predict the future state and generate useful information for effective decision-making. They also compared the result of the ID3 and C4.5 decision tree algorithms. The system could discover hidden patterns in the data and it also enhanced real-time indicators and discovered bottlenecks and it improved information visualization.

Markl et al. [2] suggested that OLAP performance can be improved by using the Multidimensional Hierarchical Clustering (MHC) technique. Clustering was introduced as a way to speed up query aggregation without additional storage cost for view materialization. The authors identified the problem with queries which either select a very small set of data or perform aggregations on a fairly large data set. The sole contribution of their work is an encoding scheme for hierarchical dimensions that enables clustering of data with respect to multiple and hierarchical dimensions. The major strength of the work lies in the comparison of their MHC technique with the traditional bitmap indexing approach on the real world data (7GB in size) and finding an increase in the performance up to the factor of 10.

VelidePhani Kumar et al [3] analyzed diabetes data using various data mining techniques which involved, Naive Bayes, J48(C4.5) JRip ,Neural networks, Decision trees, KNN,Fuzzy logic and Genetic Algorithms based on accuracy and time. They found that that out of various data mining techniques which were employed to analyze the diabetes data. J48 (C4.5) took least time.

Ben Messaoud et al. [11], propose OpAC(Operator for Aggregation by Clustering) which is considered as a new operator for multidimensional on line analysis. It consists in using the agglomerative hierarchical clustering to achieve a semantic aggregation on the attributes of a data cube dimension. The authors propose taking advantage from both the OLAP and the Data Mining to get at the end an analysis process that provides the exploration, explication and prediction capabilities.

K. Rajesh et al.[4], carried out a research to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes. The training dataset used for data mining classification was the Pima Indians Diabetes Database they applied Different classification techniques and found out that c4.5 classification algorithm was the best algorithm to classify the data set.

Chen et al. [12], suggest a scalable DW andOLAP-based engine for analyzing web log records. The proposed framework supports the typical OLAP operation and DM operations such as extended multilevel and multidimensional association rules. The OLAP server is used as a computation engine to support DM operations.

Hann et al. [25], proposed the generation of tool specific schemata for OLAP from conceptual graphical models. Their work described the design and implementation of the generation component in the context of their own data warehouse environment. The principle issues of designing and implementing such an automatic schema generation component and the possible solutions have been discussed by the authors. Further topics are the use of graph grammars for specifying and parsing graphical multidimensional schema descriptions and the integration of the generation process into a metadata cantered modeling tool environment.

Implementation Methodology

In this section, based on our implementation discuss in detail about the steps involved in the implementation of the proposed model using the traditional version of k-mode, this level is ignored. So, to overcome this problem proposes "AK-Mode" that extends Simple Matching (SM) dissimilarity measure by adding the ontology; by this way improve the efficiency of this measure.

Data set

The data set used for the purpose of this study is Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases data sets are available this below sites.

https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Data cleaning

Data cleaning, as indicated, is also closely related to data mining, with the objective of suggesting possible inconsistencies.

Hierarchical clustering of data

Hierarchal Clustering Explorer (HCE) tool for generating the hierarchical clusters of data. This tool takes input data file and allows the hierarchical clustering of given data based on

different clustering parameters. At this point, user can select the parameters to perform specific type of hierarchical clustering on the data.

Clustering

Cluster analysis [8] divides the data points into groups of points that are "close" to each other. It starts with every data point being a cluster and repeatedly aggregates the most similar (least dissimilar) groups together until there is just one big group. The number of groups can be chosen subsequently.

THE AK-MODE ALGORITHM

In this section, we propose AK-Mode which , is an extension of the k-mode algorithm. The Data Mining (DM) is to summarize the data in novel ways that are both understandable and useful to the data owner" [10] analyze the observational data set to find unsuspected relationships. Numerous techniques and algorithms are used, in the following we give some of them: clustering, classification, prediction, etc. In our case recommend the use of clustering because it is the process of separating a given items into sets of similar foundations, so that items within a cluster have high comparison to one another, but are very unrelated to items in other clusters [5].

Algorithm: AK-Mode Optimization

1: for r = 0 → R do

2: u[0][r] ← 0, J[0][r] ← ∅

3: end for

4: for k = 1 → n do

5: u[k][0] ← 0

6: for r = 1 → R do

7: max ← 0

8: for Mk,j∈ Mk do

9: b ← u[k − 1][r − Rk,j] + ak · Uk(Mk,j) 10: if b > max then

11: max ← b

12: J[k][r] ← J[k − 1][r − Rk,j] ∪ {Mk,j} 13: end if

14: end for

15: u[k][r] ← max

16: end for

17: end for

18: return J[n][R]

Proposed Method

The association data provides the new format of diabetes diagnosis. This model syndicates the concepts and OLAP mining the idea of disclosure model. In our work extracted from the OLAP requirements proposed the AK-Mode which, is an extension of k-mode used to cluster the schemas. The goal is to behind the working and gets a set of clusters. Everyone contains set of plans belonging to facilitate the construction of data in same domain. As viewpoints are suggest the use of "union-based algorithm" instead of "frequency- based algorithm" to advance the update of the "Mode" ensure the fusion of different schemas existing in one cluster to get the corresponding data mart schema at every time propose the matching and mapping techniques.

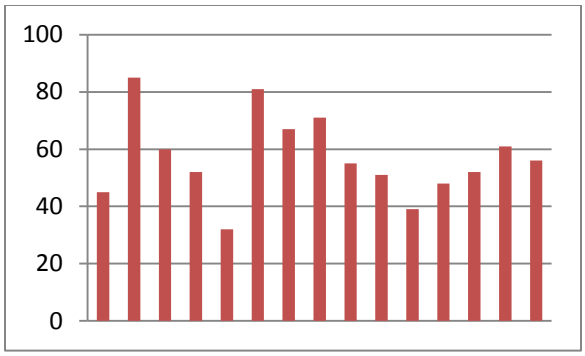
Sample data set

A knowledge discovery sample dataset is created to mine for two-year. The total dataset contains 768 instances. The following table shows the samples of the original dataset. It appearances the nine attributes out of which diabetes probability is the class attribute. The other seven attributes are used for decision making by C4.5 algorithm.

S.NO	ATTRIBUTES
1	ID
2	Sex
3	No. of. Times Pregnant
4	Plasma Glucose (mg/dL)
5	Diastolic B.P.(mm Hg)
6	Skin Fold thickness (mm
7	2-Hr Serum Insulin(mu U/ml)
8	BMI (Kg/m2)
9	Diabetes Probability
10	Diabetes Type

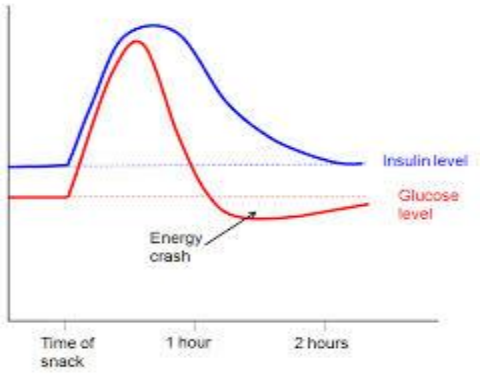
Result and Discussion

The report provides an analysis of more comprehensive and easier decision making process through the allocation of doctors to under-represented geographic areas. It allows improving the quality of doctors in the areas of representation.

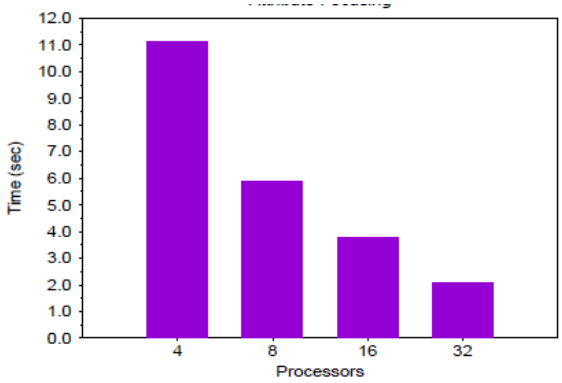


The Result of prediction to identify patients

However, by combining, we can improve the current operations and to detect patterns more accurately in a time. The above table shows the status of the patients in the diabetes record whether; the status is low, medium or high. Using this record the treatment method or ideas will be given to the patients.



Time for Consolidation



Doctors can predict patients who may be diagnosed with diabetes. The result can enhance the previous processes and expose more suitable patterns, for example, by analyzing patients demographics. Table demonstrations the result of prediction of a patient who was diagnosed as diabetic with high probability. The system was able to display this result in just 10 ms.

Conclusion

This paper has presented a clinical DSS based on OLAP with data mining to identify whether a patient can be diagnosed with diabetes with probability high, low or medium. This is powerful system because (1) it discovers hidden patterns in the data, (2) it enhances real-time indicators and discovers bottlenecks and (3) it improves information visualization. It is obvious from the result that the prototype system overcomes the physical plan design and execution requirement in the data warehousing environment.

References

[1] RupaBagdi, Prof. PramodPatil, "Diagnosis of Diabetes Using OLAP and Data Mining Integration" in International Journal of Computer Science & Communication Networks,Vol 2(3), 314-322.

[2]. V. Markl, F. Ramasak and R. Bayer, "Improving OLAP performance by multidimensional hierarchical clustering," in Proc. of the 1999 Int'l Symposium on Database Engineering and Applications (IDEAS), 1999, p. 165.

[3] VelidePhani Kumar, Lakshmi Velide, "A data mining approach for prediction and treatment of diabetes disease" in international journal of science inventions today Volume 3, Issue 1, January-February 2014.

[4] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[5] A. Omari, M. B. Lamine, and S. Conrad, "On UsingClustering And Classification During The Design Phase To Build Well-Structured Retail Websites", IADISEuropean Conference on Data Mining 2008, Amsterdam, The Netherlands, 2008, pp. 51 59.

[6]. A. Cuzzocrea, D. Sacca and P. Serafino, A hierarchy driven compression technique for advanced OLAP visualization of multidimensional data cubes, in Proc. of 8th Int'l Conf. on Data Warehousing and Knowledge Discovery (DaWak), (Springer Verlag 2006), pp. 106-119.

[7]. V. Peralta, A. Marotta and R. Ruggia, "Towards the automation of data warehouse design," Technical Report TR-03-09, InCo, Universidad de la República, Montevideo, Uruguay, June 2003.

[8] Everitt B. (1980). Cluster Analysis (second edition). Halsted, New York.

[9]. S. Chaudhuri and U. Dayal, An overview of data warehousing and OLAP technology, ACM SIGMOD Record, Vol. 26 (1997), pp. 65-74.

[10] D. Hand, H. Mannila and P. Smyth, "Principles of Data Mining", MIT Press, Cambridge, MA, 2001.

[11] R. Ben Messaoud, S. Rabaséda, O. Boussaid, and F.Bentayeb, "OpAC: A New OLAP Operator Based on a Data Mining Method", ixth International Baltic Conference on Databases and Information Systems (DB&IS 04), Riga, Latvia, 2004.

[12] Q. Chen, U. Dayal, and M. Hsu, "An OLAP-based Scalable Web Access Analysis Engine", In Proceeding of CASCON'97: Meeting of Minds, Toronto, Canada, 1997.

[13] Panos, V., and Timos, S., A Survey on Logical Models for OLAP Databases. ACM Sigmod Record, *28*(4), 64-69, Dec. 1999.

[14] Robert, S.C., Joseph, A.V. and David, B., Microsoft Data Warehousing: Building Distributed Decision Support Systems, London: Idea Group Publishing, 1999.

[15] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, *4*(2), 9-17, March/April, 2002.

[16] Hedger, S.R., The Data Gold Rush, Byte, *20*(10), 83-88, 1995.

[17] Bill, G. F., Huigang, L. and Kem, P. K., Data Mining for the Health System Pharmacist. Hospital Pharmacy, *38*(9), 845- 850, 2003.

[18] Usama F., Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDBM ‟97), Olympia, WA., 2-11, 1997.

[19] Raymond P.D., Knowledge Management as a Precursor Achieving Successful Information Systems in Complex Environments. Proceedings of SEARCC Conference 2004, 127-134, Kuala Lumpur, Malaysia.

[20] Ralph, K. and Margy, R., The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling (2nd ed.), Canada: John Wiley & Sons, Inc, 2002.

[21] Torben, B.P. and Christian, S.J., Multidimensional Database Technology, IEEE Computer, *34*(12), 40-46, 2001, December.

[22] Usama, M. F., Data Mining and Knowledge Discovery: Making Sense Out of Data, IEEE Expert, 20-25, 1996, October.

[23] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, *4*(2), 9-17, March/April, 2002.

[24] David K. and Daniel O'Leary, Intelligent Executive Information Systems. IEEE Expert, *11*(6), 30-35, Dec. 1996.

[25] Hann, J., Kamber, M., Data Mining Concepts and Techniques, San Diego, USA: Morgan Kaufmann Publishers, pp. 294- 296

[26] RupaBagdi et al, International Journal of Computer Science & Communication Networks,Vol 2(3), 314-322

