Global Journal of Computer Science and Technology
Volume 11 Issue 13 Version 1.0 August 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Encoding and Decoding Techniques for Distributed Data Storage Systems

By A.Anbarasi, Dr.K.Vivekanandan

Bharathiar University,Coimbatore,India

Abstract - Dimensionality reduction is the conversion of high-dimensional data into a meaningful representation of reduced data. Preferably, the reduced representation has a dimensionality that corresponds to the essential dimensionality of the data. The essential dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [4]. Dimensionality reduction is important in many domains, since it facilitates classification, visualization, and compression of high-dimensional data, by helpful the curse of dimensionality and other undesired properties of high-dimensional spaces [5]. Dimension reduction can be beneficial not only for reasons of computational efficiency but also because it can improve the accuracy of the analysis. In this research area, it significantly reduces the storage spaces.

Keywords : Dimensionality reduction, high-dimension and storage.

GJCST Classification : E.2, E.3

Encoding and Decoding Techniques for Distributed Data Storage Systems

A.Anbarasi[α], Dr.K.Vivekanandan[Ω]

Abstract - Dimensionality reduction is the conversion of high-dimensional data into a meaningful representation of reduced data. Preferably, the reduced representation has a dimensionality that corresponds to the essential dimensionality of the data. The essential dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [4]. Dimensionality reduction is important in many domains, since it facilitates classification, visualization, and compression of high-dimensional data, by helpful the curse of dimensionality and other undesired properties of high-dimensional spaces [5]. Dimension reduction can be beneficial not only for reasons of computational efficiency but also because it can improve the accuracy of the analysis. In this research area, it significantly reduces the storage spaces.

Keywords : Dimensionality reduction, high-dimension and storage.

I. Introduction

The analysis and mining of large volumes of transaction data for making business decisions. Today it has become a key success factor than ever for vendors to understand their customers and their buying patterns. If they don't they will lose them. In order to gain competitive advantage it is necessary to understand the relationships that prevail across the data items among millions of transactions. The amount of data currently available for studying the buying pattern is extensive and increasing rapidly year by year. Therefore the need to devise reliable and scalable techniques to explore the millions of transactions for the customer buying pattern continues to be important. Above this, the increasing volume of data sets data demands for huge amounts of resources in storage space and computation time. As it is not feasible to have huge storage spaces to store the explosively growing data in a single location they are stored in distributed database and data warehouse located in different geographical location. Inherently data distributed over a network with limited bandwidth and computational resources motivated the development of distributed data mining (DDM).

Though mining process in DDM is carried out in distributed locations parallel and generates required results in the local areas it is necessary to analyze these local patterns to obtain the global data model. Hence the knowledge derived from local distributed location is moved to the central site and the local results are combined there to obtain the final result. This approach is less expensive but may produce ambiguous and incorrect global results. Even though communication is a bottleneck problem in a central data repository it guarantees accurate results of data analysis. To address the bottleneck problem in central learning strategy, this work proposes a dimension reduction method which uses the concept of sum of subset and scalable to very large databases. In this work the site which request data from different geographical locations is treated as central site.

II. Existing Work

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has as high a variance as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed.

Linear Discriminant Analysis (LDA) attempts to maximize the linear separability between data points belonging to different classes. In dissimilarity to most other dimensionality reduction techniques, LDA is a supervised technique. LDA finds a linear mapping M that maximizes the linear class separability in the low-dimensional representation of the data.

III. Problem Description

Data storage conversion algorithm transforms a transaction into a single dimension transaction with all attributes that appears in its original form. The encoded transactions are represented by a sequence of numbers, which is sum of subset approach. Any kind of combination of $2^1,2^2,2^3…2^n$,the sum of different values gives as unique, This is way motivated to do the my research work. By this way, the new transaction is smaller than the original form and hence the cost of Storage is reduced.To offer highly specialized solutions for small parts of the general problem.

Author [α] *: Research Scholar,Bharathiar University,Coimbatore,India. Telephone: 09751149851 E-mail : anbarasi2@gmail.com*
Author [Ω] *: Prof,Bharatiary University,School of Management and studies,Coimbatore India.*
E-mail : vivekbsmed@gmail.com

IV. Encoding and Decoding Techniques for Data Storage

A matrix is constructed with the given set of data items as shown in the Table-1. The order and dimensions of the matrix are user defined. The only constraint is that the number of columns should not exceed 14, as the value of 2 ^ 14 will exceed the range of an 'int'.

pizza	sauce	sugar	sweet bun	
soft drink	fruit wheat bread	honey	jam	Dry fruits
wheat bread	bun	burger	butter	chickn

Table 1: Display of item set

Of the entire set of data items, a transaction is always a subset of the data items. This subset of data items is encoded into a reduced database. If it reduced database minimized the memory area. The Table-2 explains the encoding process. For each data item in the transaction, the row 'i' and column 'j' is noted. The value 2 ^ j is calculated and added to the i'th value in the transaction value E. The process is repeated for each data item one by one and final 'n' digits from the Table-2 give the Encoded transaction. Data Item is reduced to 34, 50, and 10. The reduced form of transactions in table 2 is given in table 5.

Data Item	Matches with		E after adding 2j to the existing value in ith column		
	ith Row of	jth Column of			
Jam	2	4	0	0+16	0
Wheat bread	1	1	0+2	16	0
Chicken	1	5	2+32	16	0
Soft drink	2	1	34	16+2	0
Dry fruits	2	5	34	18+32	0
Sugar	3	3	34	50	0+8
Pizza	3	1	34	50	8+2

Table 2 : Illustration of the process of encoding

Table 3 contains all the transactions in the encoded form. It will be these encoded values that will be transferred across the network between the client and the server.

34	50	10
58	16	02
18	16	18

Table 3 : Dimension Reduced database

Tables 4a, 4b, 4c indicate how the transaction value is decoded to obtain the original data items of the transaction. Each value in the reduced form is taken and dealt separately in each table to obtain the data items in that particular row. For eg. , in the Transaction 34-50-10, the value 34 is dealt in Table 4a, 50 in Table 4b and 10 in Table 4c. The decoding process of is shown in table 4a , 4b and 4c In the reduced transaction T1 34, 50, 10

i=1; d1 = 34; n=5

Table 4a :

d1	n	Tk	d1= d1 – 2n	d1 – 2n ≥ 0	
				C(1, n)	Tk = Tk || C(1, n)
34	5	Empty	34-32 =2	Chicken	Tk = Chicken
02	4	Tk= Chicken	02-16=-14		
02	3	Tk= Chicken	02-08=-06		
02	2	Tk= Chicken	02-06=-04		
02	1	Tk= Chicken	02-02=0	Wheat bread	Tk = Chicken, Wheat bread

i=2; d1 = 50; n=5

Table 4b :

d1	n	Tk	d1= d1 – 2n	d1 – 2n ≥ 0	
				C(3,n)	Tk = Tk || C(3, n)
10	3	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink	10-08 =02	Sugar	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink, Sugar
02	2	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink, Sugar	02-04= 02		
02	1	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink, Sugar	02-02= 00	Pizza	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink, Sugar, Pizza

i=3; d_1 = 10; n=3

Table 4c :

d1	n	Tk	d1= d1 – 2n	d1 – 2n ≥ 0	
				C(2, n)	Tk = Tk || C(2, n)
34	5	Tk = Chicken, Wheat bread,	50-32 =18	Dry fruits	Tk = Chicken, Wheat bread, Dry fruits
02	4	Tk = Chicken, Wheat bread, Dry fruits	18-16= 02	Jam	Tk = Chicken, Wheat bread, Dry fruits, Jam
02	3	Tk = Chicken, Wheat bread, Dry fruits, Jam	02-08=-06		
02	2	Tk = Chicken, Wheat bread, Dry fruits, Jam	02-06=-04		
02	1	Tk = Chicken, Wheat bread, Dry fruits, Jam	02-02=0	Wheat bread	Tk = Chicken, Wheat bread, Dry fruits, Jam, Soft drink

The Transaction T1 = Chicken, Wheat bread, Dry fruits, Jam, Soft drink, Sugar, Pizza

In Table 4a, the items chosen in the row 1 are found by decoding the number '34'. Since the given matrix has 5 columns –the values of 2^5, 2^4, 2^3, 2^2 and 2^1 are all subtracted from the value '34' one by one, cumulatively. Each time, the subtraction gives a positive value, the corresponding column's data item is chosen. The final list of chosen data items in this table indicates the original items from the transaction. The process is repeated on the other values of the reduced transaction form ie, 50 n 10 in Tables 4b and 4c respectively. Thus the remaining data items from the transaction are also decoded.

V. Conclusion

The above stated technology appears to be the most fitting and forceful method adaptable in the distributed data as well as in the distributed data mining process in terms of speed and competence when we measure it up to the old methods. Another useful characteristic that is covered under this new technology is that it could be updated constantly when it is essential since the data is maintained at remote sites. The huge quantity of data is not needed to be stored to the much location for this purpose hence the storage spaces are used most favorably. To make complete use of the novel technology, the customary client server distributed data mining scheme must be entirely replaced with it. Methodology expansions for merging the accumulated information from different spots are in advancement. The purpose with this paper was to provide an overview of the specific of approach that can be employed for dimension reduction when processing high dimension data.

References References Referencias

1. Wu-Shan Jiang, Ji-Hui Yu., Distributed Data Mining on the Grid, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.
2. U.P,Kulkarni, K.K. Tangod, S.R.Mangalwede, A.R.Yardi, " Exploring the capabilities of Mobile Agents in Distributed Data Mining, 10th International Database Engineering and Applications Symposium (IDEAS'06), 2006 IEEE.
3. J.E. Jackson. A User's Guide to Principal Components, New York: John Wiley and Sons, 1991.
4. K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press Professional, Inc., San Diego,CA, USA, 1990.
5. L.O. Jimenez and D.A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical,and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man and Cybernetics, 28(1):39–54, 1997.

This page is intentionally left blank