Comparison of Time Taken and Compression Efficiency for Different Sizes of Databases

MRs. MEENAKSHI SHARMA[1] and Dr. Mrs Pushpa Suri[2]

[1] HCTM, Kaithal

Received: 20 August 2011 Accepted: 15 September 2011 Published: 30 September 2011

Abstract

Data compress for object oriented data warehousing. A data warehouse is an essential component to the decision support system. The traditional data warehouse provides only numeric and character data analysis. But as information technologies progress, complex data such as semi-structured and unstructured data become vastly used. Data Compression is of interest in business data warehousing, both because of the cost saving it offers and because of the large volume of data manipulated in many business application.[3],[5]. The entropy is used in many areas such as image processing, document images. But in our research we used the entropy in object oriented data warehousing. Creation of different sizes of databases in oracle. Employment of object oriented programming for compression using Datawarehousing.

Index terms— Data warehousing, Data compression, Object oriented, Entropy.

1 INTRODUCTION

ne of the hottest topics in the industry today is data warehousing and on-line analytical processing (OLAP). Although, data warehousing has been around in some form or another since the inception of data storage, people were never able to exploit the information that was wastefully sitting on a tape somewhere in a back room. Today, however, technology has advanced to a point to make access to this information an interactive reality. Organizations across the country and around the world are seeking expertise in this exploding field of data organization and manipulation. It is not a surpise, really, that business users want to get a better look at their data. Today, business opportunities measure in days, instead of months or years, and the more information empowering an entrepreneur or other business person, the better the chances of beating a competitor to the punch with a new product or service. The task of transitioning from a procedural mindset to an object-oriented paradigm can seem overwhelming; however, the transition does not require developers to step into another dimension or go to Mars in order to grasp a new way of doing things. In many ways, the object-oriented approach to development more closely mirrors the world we've been living in all along: We each know quite a bit about objects already. It is that knowledge we must discover and leverage in transitioning to object-oriented tools and methodologies.

A data warehouse is a mechanism for data storage and data retrieval. Data can be stored and retrieved with a multidimensional structure–hypercube or relational, a star schema structure or several other data storage techniques.

2 II.

3 DATA COMPRESSION

Data compression is of interest in business data warehousing, both because of the cost savings it offers and because of the large volume of data manipulated in many business applications. The types of local redundancy present in business data files include runs of zeros in numeric fields, sequences of blanks in alphanumeric fields, and fields which are present in some records and null in others. Run length encoding can be used to compress sequences of zeros or blanks. Null suppression may be accomplished through the use of presence bits. Another class of methods exploits cases in which only a limited set of attribute values exist. Dictionary substitution entails replacing alphanumeric representations of information such as bank account type, insurance policy type, sex, month, etc. by the few bits necessary to represent the limited number of possible attribute values.

The problem of compressing digital data can be decoupled into two subproblems: modeling and entropy coding. Whatever the given data may represent in the real world, in digital form it exists as a sequence of symbols, such as bits. The modeling problem is to choose a suitable symbolic representation for the data and to predict for each symbol of the representation the probability that it takes each of the allowable values for that symbol. The entropy-coding problem is to code each symbol as compactly as possible, given this knowledge of probabilities. (In the realm of lossy compression, there is a third subproblem: evaluating the relative importance of various kinds of errors.)

For example, suppose if it is required to transmit messages composed of the four letters a, b, c, and d. A straightforward scheme for coding these messages in bits would be to represent a by \00", b by \01", c by \10" and d by \11". However, suppose if it is known that for any letter of the message (independent of all other letters), a occurs with probability . This representation is more compact on average than the first one; indeed, it is the most compact representation possible (though not uniquely so). In this simple example, the modeling part of the problem is determining the probabilities for each symbol; the entropy-coding part of the problem is determining the representations in bits from those probabilities; the probabilities associated with the symbols play a fundamental role in entropy coding.

One well-known method of entropy coding is Huffman coding, which yields an optimal coding provided all symbol probabilities are integer powers of .5. Another method, yielding optimal compression performance for any set of probabilities, is arithmetic coding. In spite of the superior compression given by arithmetic coding, so far it has not been a dominant presence in real data-compression applications. This is most likely due to concerns over speed and complexity, as well as patent issues; a rapid, simple algorithm for arithmetic coding is therefore potentially very useful.

An algorithm which allows rapid encoding and decoding in a fashion akin to arithmetic coding is known as the Q-coder. The QM-coder is a subsequent variant. However, these algorithms being protected by patents, new algorithms with competitive performance continue to be of interest. The ELS algorithm is one such algorithm.

The ELS-coder works only with an alphabet of two symbols (0 and 1). One can certainly encode symbols from larger alphabets; but they must be converted to a two-symbol format first. The necessity for this conversion is a disadvantage, but the restriction to a two-symbol alphabet facilitates rapid coding and rapid probability estimation.

The ELS-coder decoding algorithm has already been described. The encoder must use its knowledge of the decoder's inner workings to create a data stream which will manipulate the decoder into producing the desired sequence of decoded symbols.

As a practical matter, the encoder need not actually consider the entire coded data stream at one time. One can partition the coded data stream at any time into three portions; from end to beginning of the data stream they are: preactive bytes, which as yet exert no in seuence over the current state of the decoder; active bytes, which affect the current state of the decoder and have more than one consistent value; and postactive bytes, which affect the current state of the decoder and have converged to a single consistent value. Each byte of the coded data stream goes from preactive to active to postactive; the earlier a byte's position in the stream, the earlier these transitions occur.

A byte is not actually moved to the external _le until it becomes postactive. Only the active portion of the data stream need be considered at any time. Since the internal buffer of the decoder contains two bytes, there are always at least two active bytes. The variable backlog counts the number of active bytes in excess of two. In theory backlog can take arbitrarily high values, but higher values become exponentially less likely. [13]. Sang et al in their paper "A novel approach to scene change detection using a cross entropy ," have shown that in huge video databases, an effective video indexing method is required. While manual indexing is the most effective approach to this goal, it is slow and expensive. Thus automatic indexing is desirable, and previously various indexing tools for video databases have been developed. For efficient video indexing and retrieval, the similarity measure is an important factor. This paper presents new similarity measures between frames and proposes a new algorithm to detect scene changes using a cross entropy defined between two histograms. Experimental results show that the proposed algorithm is fast and effective compared with several conventional algorithms to detect abrupt scene changes and gradual transitions including fade in/out and flash light scenes **??**12].

4 III.

5 RELATED WORK

IV.

6 OBJECTIVE

The objective of the present study is to 1. Develop data of compression for object oriented data warehousing.

7 Devise efficient compression algorithms in data

warehousing to enhance the efficiency of the data warehousing packages so that less CPU time and less Memory is consumed. Implement compressor and expander using entropy algorithm and test its effectiveness on different sized databases [1] [2]

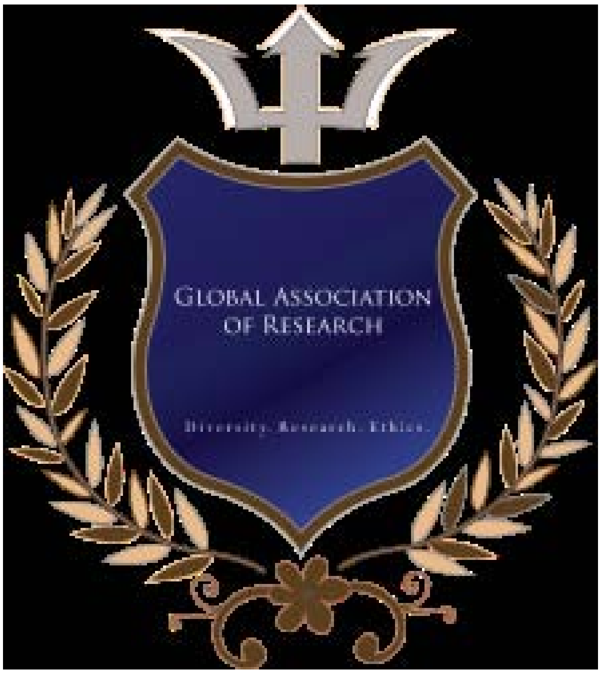
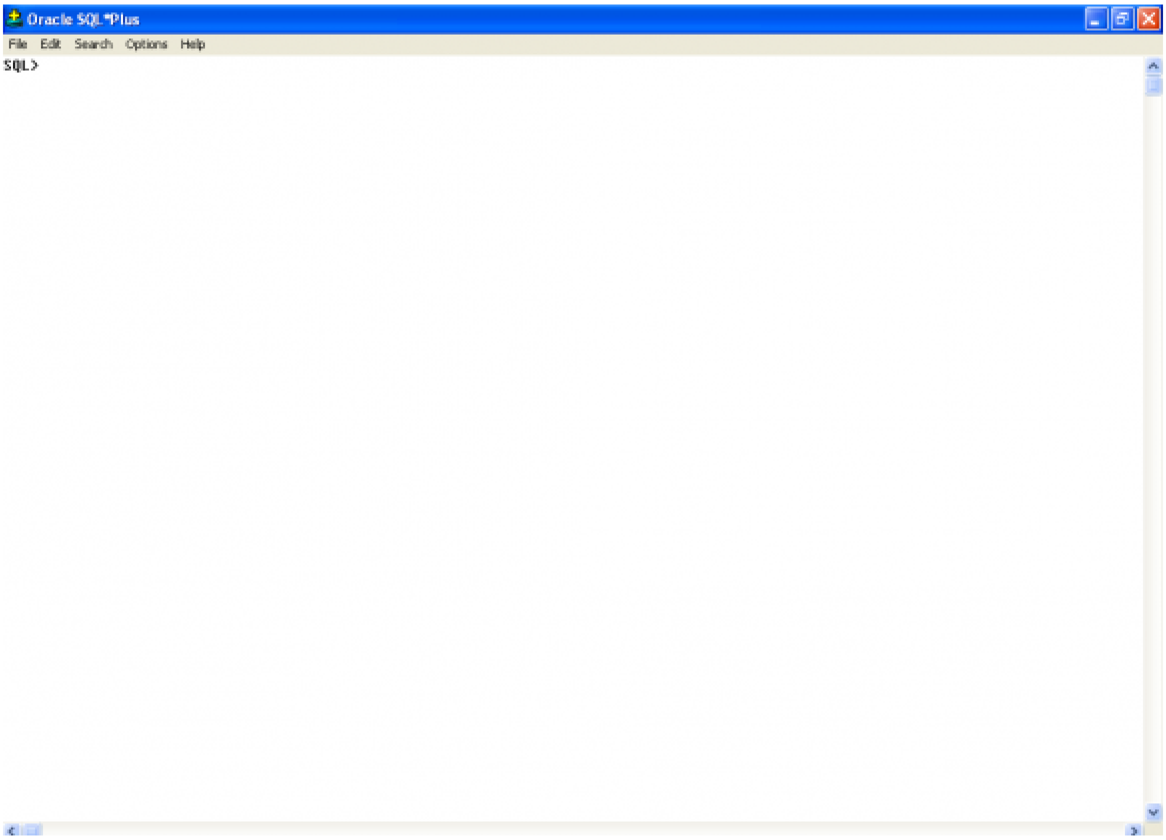


Figure 1:

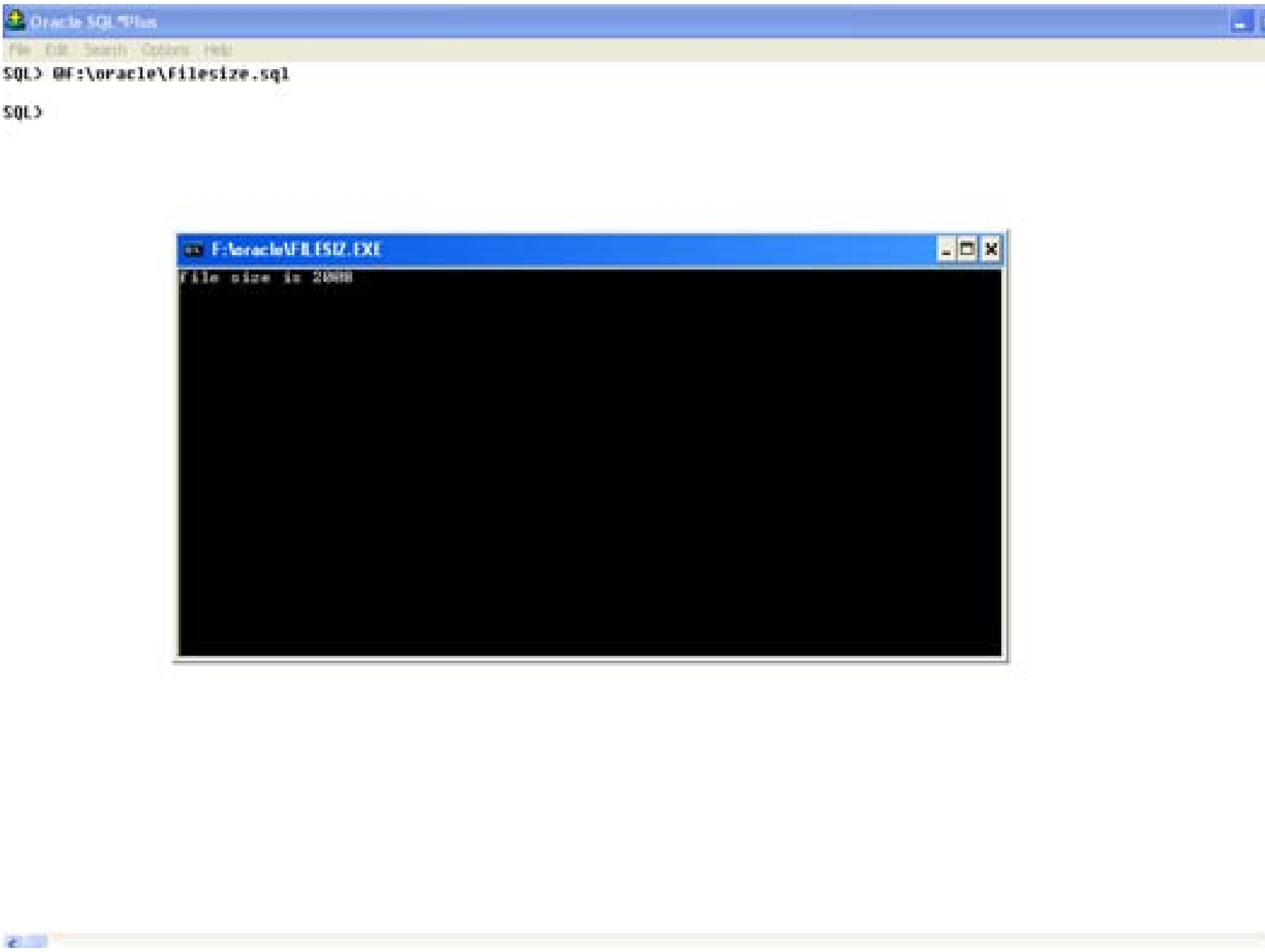
[1] © 2011 Global Journals Inc. (US)

[2] Comparison of Time Taken and Compression Efficiency for Different Sizes of Databases



```
SQL>
```

Figure 2:



```
SQL> @F:\oracle\Filesize.sql

SQL>
```

```
File size is 2008
```

Figure 3:

.1 CONCLUSION

In this paper we have discuss the data compression and how the data is compresses in oracle 10g using object oriented language. Data Compression is of interest in business data warehousing, both because of the cost saving it offers and because of the large volume of data manipulated in many business application. The entropy is used in many areas such as image processing, document images. But in our research we used the entropy in object oriented data warehousing. Creation of different sizes of databases in oracle. Employment of object oriented programming for compression using data warehousing. Further compression of database .csv files using C++. Comparison of time taken and compression efficiency for different sizes of databases

[Hong] , Wei-Chou Chen; Tzung-Pei Hong .

[Xie] , Hua Xie .

[Lin ()] 'A composite data model in object-oriented data warehousing'. Wei-Yang Lin . TOOLS 31. Proceedings 1999. 1999. p. . (Technology of Object-Oriented Languages and Systems)

[Ambhore ()] 'A Implementation of Object Oriented Database Security'. P B Ambhore . *5th ACIS International Conference on Software Engineering Research*, August 20-22, 2007. 2007. IEEE Computer Society. Institute of Electrical and Electronic Engineers

[Park ()] 'A novel approach to scene change detection using a cross entropy'. Rae-Hong Park . *Proceedings. 2000 International Conference on*, (2000 International Conference on) 2000. 2000. 3 p. .

[Liu et al. (2003)] 'An entropy based segmentation algorithm for computergenerated document images'. L Liu , Y Dong , X Song , G Fan . *Proceedings. 2003 International Conference on*, (2003 International Conference on) 2003. 2003. Sept. 2003. 1 p. .

[De and Sil (2010)] 'ANFIS tuned no-reference quality prediction of distorted/decompressed images featuring wavelet entropy'. I De , J Sil . *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, 8-10 Oct. 2010. p. .

[Swift ()] *Building Advanced Data Warehouse, NCR Corporation*, R Swift . 1996. California.

[Tu and Tran (2002)] 'Context-based entropy coding of block transform coefficients for image compression'. C Tu , T D Tran . *IEEE Transactions on* Nov 2002. 11 (11) p. . (Image Processing)

[Park ()] *Data Warehouse Designing on Relational Database Systems*, M Park . 1996. Stanford: Informix Co.

[Eder et al. ()] J Eder , H Frank , W Liebhart . *Optimization of Object-Oriented Queries by Inverse Methods. Proceedings of East/West Database Workshop*, (Austria) 1994.

[Jegou and Guillemot (2005)] 'Entropy coding with variable length re-writing systems'. H Jegou , C Guillemot . *ISIT 2005. Proceedings. International Symposium on*, 2005. Sept. 2005. p. . (Information Theory)

[Ortega (2002)] 'Entropy-and complexityconstrained classified quantizer design for distributed image classification'. A Ortega . *Multimedia Signal Processing* 2002. 11 Dec. 2002. 9 p. . (IEEE Workshop on)

[Institute of Electrical and Electronic Engineers Forum on Strategic Technology ()] 'Institute of Electrical and Electronic Engineers'. *Forum on Strategic Technology* 2007. IEEE Computer Society.

[Scales et al. (1995)] 'Lossless Compression Using Conditional Entropy-Constrained Subband Quantization'. A Scales , W Roark , F Kossentini , M J T Smith . *Data Compression Conference, 1995. DCC '95. Proceedings*, Mar 1995. 498 p. .

[Molina ()] *Maintenance in Data Warehousing Environment*, G Molina . 1995. San Jose Co., California.

[Marlin ()] E Marlin . *ODBMS vs. Relational Object-Oriented Programming*, (SAGE, London) 1992.

[Bertino ()] 'Method precomputation in objectoriented databases'. E Bertino . *Proceedings of ACM-SIGOIS and IEEE-TC-OA International Conference on Organizational Computing Systems*, (ACM-SIGOIS and IEEE-TC-OA International Conference on Organizational Computing Systems) 1991.

[Gong et al. (1999)] 'On entropyconstrained residual vector quantization design'. Y Gong , M K H Fan , C.-M Huang . *Data Compression Conference*, 1999. Mar 1999. 526 p. . (Proceedings. DCC '99)

[Relational database schema integration by overlay and redundancy elimination methods, in International Jae Jin Koh (1993)] 'Relational database schema integration by overlay and redundancy elimination methods, in International'. *Jae Jin Koh* 3-6 October, 2007. Nov 1993. p. .

[Roussopoulos ()] N Roussopoulos . *Data Warehouses and Materialized Views*, (Greece) 1997. Leander Press.

[Michael ()] *The next generation DBMS*, S Michael . 1991. New York: Pearson Education.

[Shieh and Lin ()] 'The Novel Model of Object-Oriented Data Warehouses'. J C Shieh , H W Lin . *Workshop on Databases and Software Engineering*, 2006.

[Wei-Chou and Tzung-Pei ; Lin Wen-Yang (2009)] 'Three maintenance algorithms for compressed object-oriented data warehousing'. Chen Wei-Chou , Hong Tzung-Pei ; Lin Wen-Yang . *5. Boqiang Huang; Yuanyuan Wang; Jianhua Chen*, April 2009. 56 p. . (IEEE Transactions on)

[Chen and Reif ()] 'Using difficulty of prediction to decrease computation: fast sort, priority queue and convex hull on entropy bounded inputs'. S Chen , J H Reif . *Proceedings., 34th Annual Symposium on*, Sang Hyun, Kim (ed.) (34th Annual Symposium on) 1993.

[Lin ()] 'Using the compressed data model in objectoriented data warehousing'. Wei-Chou Chen; Tzung-Pei Hong; Wen-Yang Lin . *IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, 1999. 1999. 5 p. .

[De and Sil (2010)] 'Wavelet entropy based no-reference quality prediction of distorted/decompressed images'. I De , J Sil . *2nd International Conference on*, 2010. April 2010. 3 p. .