Cuneiform Representation of Textual Data

D.M.TOTEV
Accounting Department
National University of Science and Technology
N U S T, P.O.Box 346, Bulawayo, Zimbabwe
Facsimile: +(263 9) 75953

Abstract: The fundamental concept of the Pythagorean Brotherhood "All things are numbers" could easily be implemented in processing textual data considered not only from a linguistic point of view but also as a quantity of numbers. This Pythagorean tenet combined with another primary idea of antiquity - the stylised representation of objects by cuneiform scripts allows significant data compression using a binary independent medium (paper, music, optical fibers, etc.). Standardised 8x13 mesh representation of unsigned integers (up to 10^{31} by a single character) illustrates the utilisation of these ideas in data communication and archive databases. A simple GWBASIC program emulates an Optical Character Recognition (OCR) procedure and demonstrates the feasibility of the proposed new technique.

Keywords: Multimedia

Computing Review Category: D.3

1. Introduction

The concept "All things are numbers" is derived from the Pythagorean tenets that "philosophy is the highest music" and that the highest form of philosophy is related to numbers only. The meaning of this idea could perhaps be reworded thus:"all things have form, all things are form; and all forms can be defined by numbers". Therefore, the form of the square corresponds to a "square number", whereas 12 is an "oblong number", and 6 a "triangular" one. In a similar way, "cubic" and "pyramidal" numbers could be defined and this is the chief task of the Philosopher, the Lover of Wisdom. An example of the power of numbers is the famous theorem of a right-angled triangle ($C^2 = A^2 + B^2$). Numbers are considered by Pythagoreans as patterns of dots which form the above mentioned characteristic figures. Although arabic numbers are used traditionally and they have nothing in common with the dot-patterns of Pythagoreans, numbers are still called "figures", i.e. shapes.

Cuneiform scripts were used throughout the Middle East for over 3000 years until about the 1st Century BC and originated from pictographic writing, the oldest writing known. Pictograms were substituted gradually by cuneiform symbols representing syllables, words and simple phrases. The Sumerian cuneiform character "ti" means the phrase "vital power"; the symbol of the late cuneiform script  stands for the word "mountain".

2. Analysis of the late cuneiform script

The later period of cuneiform scripts (about 1st millennium BC) could be better explained if the development of ancient alphabets is outlined in brief.
A scan of the ancient writing systems reveals that pictograms and the cuneiform script are direct, stylised representation of objects in the real world. Such signs (for example - mountain) stand for whole words or even simple phrases. The space occupied by a story, described in a pictographic (cuneiform) way is much less than if the same story is written using any of the later alphabets such as Phoenician, Old Hebrew, Etruscan, Classical Greek, etc. Obviously, at that time carving messages was quite a laborious job and naturally, the tendency was to reduce the graphics size on the medium (stone or clay tablets, wood, etc.) and increase the expressiveness of writing. Such considerations justify the conclusion that the first writing systems were very economic in terms of the graphics representation and record storage respectively.
The late cuneiform script was quite economic in respect of record storage too, but the symbols used varied in size and complexity. For example, the noun "mouth" and the verb "to speak" were depicted by the sign ; the verb "to eat" was represented by . Compared with the symbol standing for the noun "mountain" these two are much more complicated and bigger in size.
Obviously, the first thing that should be done, in case of utilising the storage advantage of the late cuneiform script, is to standardise the size of each symbol and to maintain one and the same graphics quality for all signs. Unfortunately, the latter is a crucial point that could be achieved only if the wedge structure of the cuneiform script is changed. The most appropriate substitute for the wedge "pixels" of the standardised cuneiform representation is the distinctive and neat Pythagorean dot-pattern. Since Pythagorean patterns of dots are in reality numbers everything would be expressed in a numerical manner using the standardised and converted cuneiform script. At last, after a few "sleepwalking" moves the proposed idea has converged slowly to the nucleus of the Pythagorean truth.

3. Binary model of the standardised cuneiform representation

Standardised 8x13 mesh (fig. 1, left) for designing alphanumerical characters could be used to illustrate the proposed procedure. This could be done a little bit easier if suitable ready-made software is available, e.g. Resource Workshop, Borland 1992, for Windows.

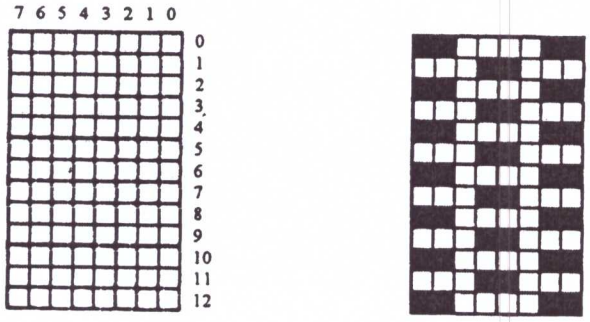

Figure 1

The essence of the standardised cuneiform approach is that each character of the new script stands for a positional notation of an unsigned integer number. The number is considered in a numerical system whose base (radix) is 256. In accordance with this notation, the corresponding boxes of the mesh are highlighted taking into account that the least

significant byte is on top and the least significant bits are in the most right column of the matrix. After that, the uniform symbol is ready to be printed out (fig. 1, right).
Each row of the mesh (with eight columns) represents one byte. Since there are thirteen rows, the biggest unsigned integer that could be worked out using the mesh is 256^{13} (about 10^{31}). In case of the biggest number, all boxes of the mesh should be highlighted. The smallest number (zero) is depicted by blank space equivalent to the space occupied by any other symbol of the devised new script. The name of this script is quite symbolic because the only resemblance between the original cuneiform script and the modified one is that both are object-oriented representations. The former directly depicts a subset of objects while the latter represents unsigned integers (up to 10^{31}) which could be assigned to any group of objects with up to 10^{31} items.
At this point, a binary independent medium should be introduced because it is closely related to the data compression matter. A character of the modified cuneiform script could be only stored on a diskette (HDD, FDD, etc.) or in the computer memory (RAM, ROM) as a sequence of the bytes that make it up (13 bytes). In such a case there is no advantage at all because these hardware units are binary oriented by design. Paper, optical fiber, etc. are image oriented media that allow significant compression to be achieved. The proposed 8x13 mesh secures data compression ratio of 1:31 because one standardised cuneiform character could represent up to thirty one decimal digits. It means that utilising the object-oriented feature of binary independent medium, each of the new characters occupies about 31 times less carrier space than the represented decimal number with 31 digits. It should be pointed out that implementing the new script for data processing, the compression ratio varies depending on the values of numbers expressed.

4. Interpretation procedure

A very important question, which determines the applicability of the proposed script, is about the interpretation of the new symbols. Fortunately, the answer of this question is quite encouraging and the simplicity of the converting procedure could be illustrated using a paper medium and the example of fig. 1. The character of fig. 1 (right) represents $13x10^{30}$ approximately. This sign (▥) could be digitalised using a hand scanner (Genius, model B105A). The size of the corresponding graphics file (TGA extension) is 2146 bytes (uncompressed). The scanned symbol could be visualised on the screen and printed out on paper (fig. 2) applying iPhoto Deluxe software (U-Lead Systems, Inc., 1992).
For the sake of space, the top six rows of the mesh are presented only.

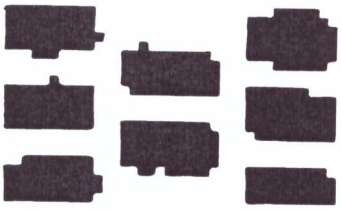

Figure 2

The graphics file could be easily processed with the following short GWBASIC program:

```
5 OPEN "R",#1,"A:SIGN.TGA",1
10 FIELD #1,1 AS B$
15 REM The heading of the file takes up the first 18 bytes.
20 B=18
```

```
25 FOR J=1 TO 58:P$=""
30 FOR I=1 TO 37:B=B+1
35 GET #1,B:IF ASC(B$)=0 THEN LPRINT CHR$(176);:GOTO 45
40 LPRINT " ";
45 NEXT I:LPRINT:NEXT J
50 CLOSE
55 END
```

The program emulates an OCR procedure and its output (the top six rows of the mesh only) is presented by fig. 3. This output shows that the interpretation of the graphics file results in quite regularly shaped structures representing adequately the original (please see fig. 2 as well).

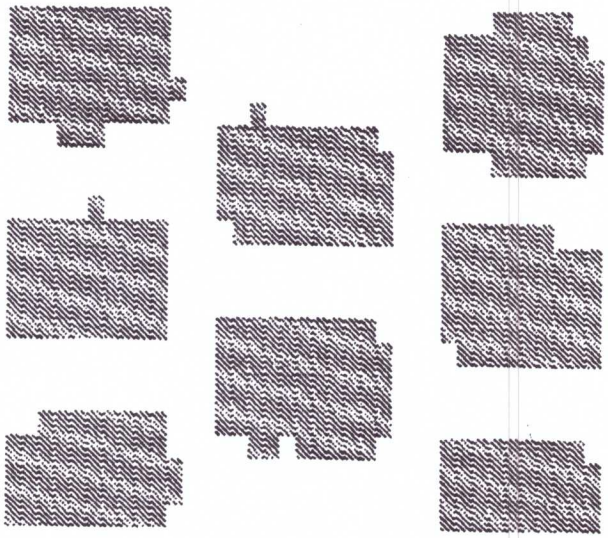

Figure 3

5. Applications

The reduction of the medium space could be even increased if numbers, represented by the new script, are attached to each word of a dictionary. For examplelet us print out the word "mountain" ten times in a traditionally linquistic way (enclosed in a single line box) mountain mountain mountain mountain mountain mountain mountain mountain mountain mountain and compare with ten of the proposed symbols (enclosed in a double line box)  .The above symbol stands for number 8450 that could be assigned to the word "mountain" in the Collins English Learner's Dictionary (for example).

The 8x13 mesh used provides 104 different table locations and the number of their permutations (104!) is more than 10^{100}. It means practically unbroken ciphers for data security could be devised on the base of the new script.

The new technique could be useful for data encryption, data communication (optical fibres), database archives (paper, microfiches).

6. References

1. I.Velikovsky, Worlds In Collision, Abacus 1972.
2. A.Koestler, The Sleepwalkers, Hutchinson & Co. (Publishers) Ltd,1959.