Comp. Appl. Arch. 13, 1985.

INTERPRETING FUZZY DATA

Stephen George

Queen Mary College Computer Centre, University of London,
Mile End Road, London E1 4NS

Abstract

The continuing use of computers for collection and interrogation of data in archaeology is generating data-gathering forms of ever greater complexity. The normally chosen computer software systems require a more exact definition of allowable data values, while in practice the user's view of the data tends to be more inconcise. This paper demonstrates how it is possible to interrogate free text, for example the description or probable dating of a pottery sherd, using fuzzy searching techniques. CAFS (Contents Addressable FileStore) is specialised hardware/software, with a very fast search time, that is used on ICL mainframes, which have inbuilt text organisation and fuzzy operators. An Archaeological environment where this technique may be of use is described.

Introduction

There are many problems that can occur when moving infomation from manual systems to computer-based systems. Depending on when the decision to start using computers is taken, the problem falls into two categories:

Designing environments where a new archaeological dig is starting
Having use of computer resources after the start of the dig

Both of these environments have inherent difficulties in the required data analysis and the man-machine interface.

This paper mainly deals with the second problem, namely the transition between manual and computer based systems once some or all of the information has been collected. A particular problem with the kind of information collected in archaeology is that it is very likely that this data collection cannot be repeated. It should be said that more questions are asked than solutions are given.

Background

From 1979 to 1985 there has been an archaeological excavation in Gravina, Southern Italy. Part of the work of the excavation is to give practical experience to students. The field work comprises one month every summer on site and one month in Rome analysing the pottery, etc. The personnel come from the Archaeology Department, Lancaster University and the Classics Department, Queen Mary College, London University.

In 1983 a disk-based BBC microcomputer system with a lineprinter was borrowed with the intention of evaluating its usefulness in aiding the collection, validation and interrogation of site excavation data both then and in the future. Various other mainframe, mini, and microcomputers have also been available at both

University Computer Centres and Departments including an ICL 2988 mainframe, at Queen Mary College, which has specialised searching capabilities.

Transition

Among the many problems that can occur when moving from manual to computer based systems, this paper deals with four that have special emphasis to research oriented problems. These are:

integration:
the translation of all possible information onto a computer
facility:
availability of resources
knowledge:
a need to know what is wanted
data rigidity:
the need to put information into pre-defined boxes

There is a lot of interrelation between these topics. Is it, for example, feasible in terms of filestore availability to integrate all the data? What amount of keyword coding is required, if any?

Integration

Independent of the hardware and software facilities available, all information must be analysed with a view to mapping onto a computer system, thus allowing full data integrity. The generally recognised computer systems used tend to be based on some database structure, but even if the target system is based on classical file systems, this analysis is still invaluable. Actually to perform the analysis helps the archaeologists and the computer personnel to comprehend the structure and meaning of the data involved.

This analysis can best be organised by using two well known techniques: Entity-Relationship modelling (Chen 1976) and Third Normal Form Analysis (Date 1981). A good example of the use of Entity-Relationship Modelling, with particular reference to Archaeology is given by (Chapman 1984).

Before computer resources were made available at Gravina the recording system was based on disjunct books, diagrams and photographs. Other, admittedly more obscure, information was recorded on the figurative matchbox back.

Given the timescales available and following the two techniques above, intermediate pro formas were designed and with the future intention of integrating all the information as shown in Figure 1.

Various other aspects, for example soil analysis, should also be included in this view of the data.

These pro formas were designed to form the basis of the main interface between the fieldwork and the computer. Great care has to be taken when designing these transforms so that no data are lost nor are superfluous data generated.

Education

The level of computer literacy among archaeologists is increasing, especially with the advent of personal computers, but the old saw that a little learning is dangerous is true and an inadequately designed system can cause gross inconsistencies in data Designing a computer environment to map real-world situations requires competent knowledge of data analysis.

```
SITE BOOKS -----------> CONTEXT PF'S ------>          D
PLANS --------- ---------------------------->
CONTEXT PHOTOS -------- ------------------->          A
                      >GENERAL POTTERY PF'S>
POTTERY BOOKS -------->                               T
                      >IND. SHERD PF'S  --->
POTTERY DRAWINGS -------------------------->          A
POTTERY PHOTOS ---------------------------->
                      >GENERAL BONE PF'S  ->
BONE BOOK ------------>IND. BONE PF'S  ---->          A
                      >IND. TOOTH PF'S  --->
BONE DRAWINGS ------------- -------------->           R
BONE PHOTOS --------- --------- ----------->
SMALL FINDS BOOK ----->SMALL FINDS PF'S --->          E
SMALL FINDS DRAWINGS ---------------------->
SMALL FINDS PHOTOS ------------------------>          A

MANUAL  ---------> FORMALISATION  - ------> COMPUTER
SYSTEMS                                     SYSTEMS
```

Figure 1: Schematic pro forma for recording excavation data

Especially within a teaching environment, it is expected that all students should have the opportunity to perform all the tasks a contemporary archaeologist performs. This education should involve use of the computer resources, whether for simple data entry or interrogation. The ability of students to fill in the relevant pro formas accurately is greatly increased after subjection to the rigours of computer data entry. The insistence in recording on using certain scales of measurement, for example lengths always being recorded in centimetres, results in more consistent data.

By using both data field and intra-record validation, linked with user friendly data collection programs or packages, even higher levels of data consistency can be achieved.

Whenever further information becomes available, for example a photograph of a potsherd, then both manual and computer based updating procedures should be available.

Facility

When choosing the type of computer system required many different criteria are involved. There are basically three different kinds of resources to be taken into consideration, hardware, software and people. Possibly human resources are the most important. What use is a computer system with no level of expertise in using or administering it?

Among the questions that need to be asked include: What to buy? What is available? When is it available?

The system has to be extensible. Is the system large enough in both filestore requirements and availability of software facilities? Also should the extra-relational links be taken into consideration? There is a requirement that the various archaeological systems should, in some way be compatible with each other. Definitions of attributes should be of the same structure on different machines. Is time held as DD/MM/YY or YYYYMMDD?. Of course archaeological time creates its own problems, especially as the time intervals become less well-defined: summer 1580, 54AD, circa 1500BC, >40,000BP.

When choosing a particular information management system, given the financial instability of software and hardware manufacturers and distributors it is probably better to choose software that can run on a variety of systems.

Resilience of computer hardware and software can also be a problem. Therefore it is necessary that data transferral to an intermediate manual system should be feasible. This is particularly important when working in foreign places, where hardware support may be nonexistent.

Knowledge

Two aspects of what data are to be considered have to be resolved.

There must be a compromise between the requirement to map everything that is known of the data, in terms of entities, attributes and relationships and the practical limitations which demand that only a subset of this perceived environment is studied in detail.

The timeliness of the data must also be considered. New analytical techniques, for example use of Nuclear Magnetic Resonance on pottery sherds, and increased knowledge of the environment require that there should be a means by which the structural nature of the data system can reflect these new techniques.

Rigidity

Normally within commercial or purely scientific data systems it is possible to classify information into known boxes or ranges of values or domains, for example a yearly income or numeric output from a recording device. Within Archaeology, however, it may be impossible to give a precise analysis either quantitatively or qualitatively for some attributes.

Consider the following descriptions of pottery sherds:

Context no.	Description
778	1 burnt monochrome like A. Smalls c9th/c8th
772	mono toothed laddered and grid design c8th?
642	impasto - 2 sherds v.fine could be Hellenistic but difficult to say
634	mono c7th but one piece with red paint so probably bichrome early c7th? another c8th?
667	GR1 mono late c8th red decoration, outturned rim with ray // Monte Irsi 7 which is c7th
662	bichrome mostly c6th pendant rays

The fixed portion holds the data that can be represented exactly and easily. For example, measurements, where and when found, etc. The variable portion holds more vague or possibly early interpretations of the data. For example, about what date an artefact is or the type of design motif it possesses. The colouring of an artefact could be represented within the fixed portion by means of its range on Munsell colours, or as a description, for example, reddish brown. For the convenience of the end-user, at Gravina both techniques are used.

To formalise the information fully it requires skilled data input personnel and forms that can be extremely rigorous or overlong. An ability to hold certain data as natural text, but retaining some way of reliable and relatively quick interrogation, is advantageous.

Whether the system is supported by hardware or software, it should be able to perform fuzzy searching, to detect near matches in search criteria.

CAFS (Contents Addressable FileStore)

There is a capability on most ICL mainframe computers to attach a piece of hardware called CAFS. It allows large amounts of information to be read from disk, at a speed of about 1Mbyte/sec, and then filtered to return only those records which satisfy particular selection criteria (ICLCUA 1984).

This capability can be used with great success in archaeological work, not just for its inherent speed, but also for the fuzzy searching.

A simple case where an individual sherd pro forma is made up of a fixed part and varable part is shown in Figure 2.

```
                             RECORD
|--------------------|-----------------------------------
| FIXED PART         | VARIABLE PART
|--------------------|-----------------------------------

  Where found          Designs?
  When found           Colours?
  Weight               Dates?
  Pottery type         Comparanda?
                       (Any other worthwhile information)?
```

Figure 2: Specimen of mixed format record

The variable part of the information is then converted into self-identifying format as shown in Figure 3.

```
|-------------|-------------|-------------|
| TYPE        | LENGTH      | DATA        |
|-------------|-------------|-------------|
```

Figure 3: Self-identifying format for variable information

In a simple case, where words are given a type of 1 and punctuation is given a type of 2, the conversion can be carried out automatically. Giving, for pottery sherd 778 above:

[1][1][1] [1][5][burnt] [1][10][monochrome] [1][4][like] [1][1][A]
[2][1][.] [1][6][Smalls] [1][4][c9th] [1][4][c8th]

It is also possible to have many different types, for example dates and motifs, but there is then a requirement for manual conversion or a certain amount of semantic analysis.

Selection

After the conversion of the variable information, all of the data are now accessible using CAFS searching techniques. This can be performed either by a self-contained query language package or through a series of calls from a high level language program. With particular reference to textual data stored in Self-Identifying Format there are three fuzzy operators available:

```
!   Stem character: matches any ending
*   Omnibus character: matches any character
?   Phantom character: matches zero or one occurence
    of any character
```

Combination of these operators permits various selection criteria, for example, mono, ladder!, t**thed and ?8!. The last of these is of particular use for comparing dates. The subsets of the data selected by these techniques can also be combined using the operators AND, OR and NOT, which allow some quite involved selections. The CAFS system also has a built-in set of functions that includes facilities for totalling, finding maxima and equalities, etc.

Conclusions

Techniques such as those outlined above may allow simpler views of the archaeological data giving end-users a more natural view of the environment. With the CAFS hardware and software becoming available as recently as since 1983 it is yet to be seen whether it will have a significant impact on archaeological use of computer systems.

References

CHEN, P.P.S. 1976 The Entity-Relationship Model: towards a unified view of data. ACM TODS 1(i).

DATE, C.J. 1981 An Introduction to Database Systems (3rd. ed.). Addison-Wesley, New York.

CHAPMAN, J. 1984 Design of a Database for Archaeological Site Data. Computer Applications in Archaeology 12: 119-128.

ICLCUA. 1984 Exploiting CAFS-ISP, Working party report. ICL Computer Users Association.