(bica)

(Draft only, for comment)

CAMBRIDGE DATABASE SYSTEM

CDS 2000

Sarah Harrison
Julian Jacobs
Alan Macfarlane
Martin Porter

(Department of Social Anthropology, Cambridge)

INTRODUCTION TO THE SYSTEM

With the spread of large and complex databases, sometimes containing descriptions of thousands of images on a videodisc, as well as long texts, the demand for flexible and powerful information retrieval systems working on micro-computers is growing.

Of course it is always possible to search through visual or textual databases sequentially, or using a hand index. Or one can create a computer indexing system, using a commercial package, such as the DBase series. The following is an overview of an alternative information retrieval system, for use with large sets of material on optical, compact or computer discs. A preliminary account of what we hoped to do was given in B.I.C.A. No.5 (February 1987). We have now done what we set out to do and a brief introduction to the information retrieval side may be of interest to readers.

The system is based on the earlier 'Muscat' (Museum cataloguing System) written by Dr. Martin Porter. It has been developed specifically for use with the large set of visual and textual materials concerning the Naga peoples of Assam. But since it works on IBM-compatible microcomputers and is a powerful and general system, it will have other applications.

The descriptions of the visual images and sounds, as well as texts, are divided into 'records'. Each record can contain a number of fields, which contains a 'code' part and a piece of text giving information. Each field may contain a group of sub-fields. There is no fixed format to the records and fields can be of any length, provided that no single record has more than 64,000 characters. The number of records is not limited.

This is a system which allows the user to describe photographs, film sequences, music, museum objects, manuscript sources and printed books, making it possible to move from one to the other without difficulty. This is unusual, since it has not previously been possible or necessary to jump between such different media. Furthermore it includes 'relevance feedback and query expansion', features which enable the user and computer system to unite in setting up better enquiries, as explained briefly below.

INDEXING AND CAPTIONING

The division of the information can be made according to one's needs. One obvious way to divide information is into substantive and administrative fields.

The substantive part includes a caption, a text field, and various keyword fields. The keyword fields allow the user to record details of people, places, ethnic groups, dates, subjects and themes.

Index terms in the database are extracted from words in caption and keyword fields and can thus be searched. But index terms are not extracted from the pure text field. For this reason, important information in the text field should be identified and inserted into the keyword fields. The information does not have to be assigned to any pre-planned hierarchical ordering of categories.

Administrative information concerns the medium (photograph, object), the present location and significant details of acquisition; it could also include, in a library or museum, the shelf location of the item concerned. The information in these fields also enters the index of terms and can be searched for.

The captioning of visual images, including museum objects, is necessarily a very subjective matter. Although many attempts have been made to standardise this by providing a check-list of what should be noted, none of these can provide more than a preliminary set of categories. After a large amount of testing, we have decided on a relatively simple selection as follows.

In the captions to photographs, we have broadly described what is happening, if there is action, or what the nature of the subject matter appears to be. Any particularly striking details may be noted, for instance a particularly fine piece of ornamentation. We have tested this procedure and found that several different people looking at the same photograph independently will describe it in roughly the same way. Yet there can be little doubt that people from a different culture and with different interests would describe the photographs in different ways.

The captions can only thus be a first approximation, and users will have to add further details (often contained in other fields) after searching and analysing images. Captions are necessary, however, since database searches can only be made by presenting the user with a set of relevant answers identified by their short captions.

In the case of objects we have tried to include something on the size, materials, functions, colours, motifs of each object. But when dealing with a complex three dimensional object, it is obvious that one can only capture a little of its complex character in words. That is why we also have a picture.

Likewise in the case of moving film, a sequence lasting twenty seconds, involving several people, could generate several pages of textual description if one noted each gesture, posture, interaction, all the material objects present. We have merely simplified this in most cases to one line, for instance "group of men and boys catching fish". Again, it will be up to users to refine what can only be a preliminary index.

The same simplification is clearly necessary with texts. Often there is a paragraph which contains information on many different topics, for instance marriage payments, political alliances, economic transactions, the interrelations of chiefs and subjects. In the short caption one can merely take out what appear to be some of the more central themes.

SEARCHING AND INFORMATION RETRIEVAL

Free text and structured queries in general.

In order to find a particular record and its attached visual or textual information, there are two main methods of searching. These are 'free text' and 'structured' (Boolean) searches. The two can be combined in this system. Structured queries (of the 'and' 'or' 'not' variety) are fairly standard in databases. However, they have certain inherent weaknesses. The number of answers retrieved is usually too large or too small; users often require an expert to compose boolean expressions of any complexity for them; the retrieved set of answers us usually not ranked in any way, and so it is necessary to inspect the entire list in the search for relevance.

The powerful feature of this system lies in the fact that it works in a way that makes it possible to inter-act with the computer. Thus it is possible to use human insight alongside computational power to improve the quality of the questions and hence the answers.

Some general features of 'probabalistic' searching and 'relevance feedback'.

For instance, one may ask a specific question, to which the best matching answer is given, then the next best answer and so on in order of declining relevance. The user is asked whether each answer is what he or she was looking for or not. Those marked as 'relevant' are then stored by the computer. The computer then presents to the user a list of the terms which appear to have been most significant in those answers marked as 'relevant'. This list will include other, associated, terms in the answers which the user had not realised were of importance. The user is then asked to add in whichever of these new terms might be used in re-phrasing the question in a more precise form. Then a better and more powerful query is re-run, bringing out further new answers and revealing further unexpected connections.

Expanding queries and making associations in searching.

In effect the computer is helping the user to find associations which were not originally anticipated. This system is therefore a powerful tool for expanding queries and for making links between hitherto unconnected facts. The software for the system has been developed to deal with materials in museums, libraries, archives and elsewhere. It will deal with databases of any size, including visual and non-visual materials, and works on a range of desk-top microcomputers, using less than 3OOk of RAM within which to run.

DATABASES

As with most structured databases systems, the material needs to be broken up into meaningful units. This can be conceived of as follows:

System --- database 1 --- database 2 --- database 3 and so on.

That is to say, it is possible, by choosing the appropriate names for the databases, to have several different ones held in one computer, any one of which can be made active as needed.

FILES OF DATA

Each specific database may include a number of separate files. For instance, our Naga videodisc includes files of indexes to moving films, photographs, objects and other materials. Thus one has the structure:

Database --- file 1 --- file 2 --- file 3 and others.

These files can be added to the database one at a time, as they are ready. There is no limit to the number of files in the database. The only constraint is the over-all size of the database.

Once inside the database these files lose their identity. The database contains records, but not files. Files are just the unit by which records are added into the database.

RECORDS AND THEIR STRUCTURE

Moving down one level, each file consists of a number of records. There is no limit to the number of records in a file. Individual records can contain up to 64,000 characters. Since it is the records which tend to be shown on the computer screen, it is sensible to keep them to roughly what will fit on one or two screens, in other words a paragraph of text. Thus one has:

files --- record 1 ---- record 2 ---- record 3 and onwards.
Each record is a separate entity; it is the most important unit in data organisation.

There are, in fact, three types of records. The R-records, are those which are indexed and are either complete in themselves, or cross-refer to images or texts. The T-records are text records, which are reached by means of an index. The A-records are 'control' records which can be used to set up the user interface, pages of help and for other purposes.

INFORMATION FIELDS

Fields within records, maximum number and length.

Each record in turn consists of a number of fields. Records are likely to have information within them which appears to fall into discrete fields, often in answer to the well-known questions "Who, what, when, where, how and why". These fields are indicated by a code or tag. There can be up to 255 separate codes per record. Since the codes can be repeated and used in many combinations, in effect the number of fields is unlimited. Thus we have the structure:

record --- field 1 --- field 2 --- field 3

Code and data parts of the fields.

Each of these fields in our conventions has two parts. A code part at the start, is indicated by an asterisk (*), to indicate that a new field in the record is being defined. This is followed by a letter or number or combination of these, which indicates what type of field the computer is to expect. For instance, we have decided that *t means a 'text' field, while *k means a 'keyword' field.

The information part of a record field; maximum field size.

The second part of the field consists of the actual information or data. Thus '*k fishes' would indicate a keyword field with the information or text word 'fishes'. Apart from the general upper limit of 64,000 characters per record, the information in the field can be of any length.

String and integer information in the fields.

The information within a field can consist of either strings (that is a sequence of letters of the alphabet, numbers, punctuation marks etc., which are treated as a string of characters), or as integers. Integers are numbers which can be used for numerical calculations. The input specification defines each field as one or the other. If letters of the alphabet are typed into the integer field, the computer will indicate an error.

Group fields and their use.

The normal field contains only one type of information. But it is often the case that one will be dealing with material where some of the information applies to the whole record, while there are some sub-parts which have specific information relevant to that part only.

For instance, when a sequence of photographs have been taken rapidly of a particular event, say a dance, or there are several shots of movie film made in quick succession from different angles, it is unsatisfactory to separate them entirely as different records. On the other hand, each photograph or shot may need a special description, as well as the general description for the whole sequence. This can be represented thus:

record --- shot 1 --- shot 2 --- shot 3

In this type of record, the general heading is put at the top, and this will apply to all the records. But each sub-field may also have both a specific frame number and caption.

Sub-fields within fields.

Any field may contain within itself further sub-fields, in other words fields within fields. For instance, one can deal with the fore/surname problem by defining a structure which had a general name field (*name) which continued the two sub-fields (*forename *surname). In practice we have done this extensively only in relation to the production, collection and acquisition of artefacts. Each of these fields has to contain some other information, for instance the date, person, or place of collection of the object.

SUMMARY

The general structure of what we have described so far may be summarised in a diagram as follows:

```
 system
     \
      ....database 1   database 2   database 3.....

          \
           \
record 1  record 2  record 3.....

    \
     \
      field 1  field 2  field 3....
       (also 'group' fields)

          \
           \
            sub-field 1  sub-field 2  sub-field 3

              \
               \
                code part(*) and data part
```

SIZE, SPEED AND OPENNESS

Thus we have a system that within one size constraint, that of a maximum of 64000 characters per record (about l5 pages of typed information on standard A4 paper) is otherwise very flexible. It can be used to index most kinds of material.

It is relatively fast. Currently, searching a thirty megabyte database containing roughly four thousand pages of indexes and texts, split into some twenty thousand separate records, we find the following search times. If one asks for all the records indexed by a specific date, they are retrieved in less than a second; likewise all the records with a certain person or place mentioned will be found within two seconds. If one asked a structured (boolean) query, which asked for all the records containing the intersection of a person's name, a place name and a date, the records would be found again within two or three seconds or less.

'Free text' retrieval can take longer, because the records are ranked in order of the probability of their matching the query. Thus a query with three terms, each occurring about 15O times, will produce the best hundred answers in order of likelihood in less than three seconds. With ten terms, each with several hundred occurrences in the database, the query might take up ten seconds.

The speed is increased considerably by being able to combine structured and free text searching. The machine will only take a second or two to find all the records mentioning a place name, and only a few seconds to find and order the records which match the terms in the free text part of the query. Since the system contains a sophisticated suffix-stripping or 'stemming' algorithm, it is possible to type in a word like 'marry' and get all the variants (marriage, married, marrying etc.).

The system is an open one. The images on the videodisc in our system are fixed, but the indexes to them and all the subsidiary texts are held on a read/write medium (a hard disc). It is possible to delete records, change records, or continuously to add further records and texts to the database. It is also possible to extend the size of the database if it is too small.

Finally, in this application, CDS 2OOO is linked to a videodisc player. Having found the record describing an image, it is possible to ask to be shown the image, whether still or moving film, on the computer screen.

NOTES AND ACKNOWLEDGEMENTS

The funding for this project came from the Nuffield Foundation, the Leverhulme Trust, the Economic and Social Research Council, Cambridge University and King's College, Cambridge, to all of whom we are most grateful.

The software system was developed by Dr. Martin Porter. A fuller description of the system upon which it is based is contained in **The Muscat Manual** (Cambridge Computer Laboratory, 3rd edn.,July 1988), and **Introduction to Muscat** (Cambridge Computer Laboratory, Feb. 1988).

The development of the system and its linking to a particular set of historical and anthropological data and to a videodisc took place within the context of the Cambridge Experimental Videodisc Project at the Department of Social Anthropology, Free School Lane, Cambridge CB2 3RF.

Further information, including how to obtain copies of full manuals, data, videodiscs and software, can be obtained from this
address.