Ann. Ist. Super. Sanità, vol. 30, n. 4 (1994), pp. 433-437

The management of information: storage and retrieval of data

Ferdinando CHIODO *(a)*, Antonio MENDITTO *(a)*, Fatoumata OUANE-KEITA *(b)* and Sergio CAROLI *(c)*

(a) Laboratorio di Biochimica Clinica, Istituto Superiore di Sanità, Rome, Italy
(b) International Register of Potentially Toxic Chemicals, United Nations Environment Programme, Geneva, Switzerland
(c) Laboratorio di Tossicologia Applicata, Istituto Superore di Sanità, Rome, Italy

Summary. - Internationally harmonized and cost-effective control of chemicals marketed worldwide greatly depend both on the generation of and easy access to reliable and comparable experimental information. Stored data are of use only if information can be retrieved quickly in an understandable form. Some models and theories of information retrieval (e.g. fuzzy set theory, probabilistic approach, artificial intelligence) are briefly discussed first, then followed by applications (such as indexing and clustering techniques). Finally the structure of databases is briefly reviewed.

Key words: information systems, toxicology, databases.

Riassunto *(Il trattamento dell'informazione: immagazzinamento e reperimento dei dati).* - Un controllo delle sostanze chimiche commerciali che sia efficace e compatibile con la loro diffusione internazionale è in funzione di una informazione attendibile e della sua accessibilità. L'utilizzazione dei dati è possibile solo se il loro reperimento è rapido e la presentazione comprensibile. Sono esposti brevemente dapprima alcuni modelli e teorie che stanno alla base del reperimento delle informazioni (ad es., gli insiemi sfumati, l'approccio probabilistico, l'intelligenza artificiale). Successivamente, sono trattate succintamente alcune applicazioni dei modelli (quali l'indicizzazione o le tecniche di raggruppamento) e la struttura delle basi da dati.

Parole chiave: sistemi di informazione, tossicologia, basi di dati.

Introduction

Internationally harmonized and cost-effective management of chemicals marketed worldwide greatly depend both on the generation of and easy access to reliable and comparable experimental information.

As it is well known, the management of data has been a challenge for human beings since the ancient times. Before the computer era, Aristotle's scheme of interlocking classifications of knowledge is an example of data management. The earliest attempts to design computers, such as Babbage's analytical machine, included plans to store and sort data. In the late nineteenth century, the Jacquard machine used holes punched in paper cards to represent instructions for controlling the action of a loom. These cards became the basis for card-sorting devices used to classify statistical data. After that many database programs were designed around the physical cards; the greatest information that could be encoded on a punch card became the length of a standard field [1].

Storage of data

Nowadays, devices for storing the coded information include magnetic tapes, disks, drums (or cylinders), laser discs, bubble memories, electron beam accessed memories (EBAMs), charge coupled devices (CCDs) and punched cards. Punched cards and paper tapes represent an older technology which today has been totally replaced.. Magnetic devices, the most popular type, employ the property of magnetic material that permits particles to be polarized in one of two directions, corresponding to either value of a binary code (digital computers process information in the form of binary codes). Laser discs (sometimes in the form of compact discs, CD) store digital information in the form of surface bumps and pits and offer great storage capacity and retrieval speed. Bubble memories (in which microscopic bubble patterns are formed on wafers of garnet crystals) and CCDs (which use the presence or absence of electronic charges to store information) are sometimes called "electronic disks" because they store information in

patterns that make it accessible cyclically. Both EBAMs and CCDs are technologies in which information is encoded as electric charges. Secondary-memory media typically used in personal computers include one or more magnetic floppy disks, each of which can store up to about 1.4 million characters of text information, and internally mounted hard disks that can store about 1 billion or more characters each.

Retrieval of data

As anyone who works in an office knows, having a large quantity of information might cause difficult access to any particular piece of information. Increased commercial use of newer storage technologies could have a major influence on database design. In the middle of the twentieth century, computers were provided with magnetic tapes for storing large quantities of data; but they had to be searched sequentially and therefore this process was very slow. Nonetheless, databases were increasingly used, first on mainframes, by large enterprises [2]. In the past three decades there has been a rapid increase both in the capacity of electronic devices to store information and in the cost of storage. In general, the development of software for data management has not kept the pace.

Databases for on-line bibliographic retrieval, built on mainframe computers, began to appear during the 1960's and the 1970's. To make it possible to create databases with customized menus and specialized forms, database management programs emerged with their own languages, called "procedural languages". As from the late 1960's, the programming ideas that software and data should be created in standard forms led database management programs to become by far more complex. In the subsequent decade programs for creating databases on personal computers appeared. Since then, databases have become essential to most organized activities in business, industry, education and governments. Billions of marginally accessible paper records became available on-line. Several database programs, among others those containing information on chemicals, are now available, many of which have their own procedural built-in languages [3].

Basic file organizations include sequential files, indexed sequential files, indexed files and direct access files. Databases are occasionally distinguished from information retrieval systems by the types of data handled and the kinds of operations provided. Information retrieval systems were established to provide bibliographic and numeric information as a service, whilst databases were developed to handle formatted data. Nowadays this boundary is becoming blurred [4]. Unstructured databases (also called full-text databases) incorporate information that is too unpredictable and rambling to fit easily into a highly-structured form like a relational database. These databases resemble large files of text, with occasional breaks for a new record, like a new chapter in a book. By requesting combinations of words that might appear in the text, a user can call forth the entire "chapter" on a paper support or on the screen. Some unstructured databases also contain structured fields, added to each record, to allow for faster, more controllable searching. Many on-line databases include a structured set of bibliographic fields, and then a long text passage in each record [5]. Low retrieval is a major drawback of the unrecorded method of data storage. To speed up the search process the files can be sorted into a serial order which makes possible to retrieve an entry data by some techniques, such as binary search, buckets, hashing, extensible hashing or B-tree. Generally, each record type or relation will be stored in a single file. Indexes are commonly implemented using a B-tree. A B-tree has the form of an upside-down tree (many leaves at the bottom and one root at the top, where leaves and roots are categories) and readily makes it possible to answer questions about items that were adjacent in the original sequence. Links among files may be handled either directly - through pointers to other records - or indirectly, through values that can be used to find appropriate records. Several file organizations have been developed to simplify retrieval on multiple keys, multikey indexes, multikey hashing, multipaging and combinatorial filing schemas. These organizations may be advantageous for specialized applications. In general, they are not usually competitive with those file organizations based on the use of separate indexes [6].

Man-machine interface

Devices that provide for the movement of information between the central processing unit of a computer system and the external world are called input/output (I/O) devices. Every computer functions by accepting inputs and producing outputs. Input is the control information (programs and commands) that directs computing activities; output is information produced because of computing activities. I/O devices may be characterized according to the information medium, the hardware technology, the speed of information transfer and the amount or capacity of information involved. Many devices support the movement of information between a storage medium and processor. Others support communication between the computer system and the external world of noncomputer devices.

Theories and models of information retrieval

There is no generally accepted theory or model of information retrieval. Various approaches do exist - e.g. fuzzy-set, utility and probabilistic models - but all of

them generate some problems. Information retrieval models attempt to relate representations of queries for information to representations of records in a database, usually in terms of bibliographic references to documents. This relationship is generally called "relevance". There are several views of what it means for a given document to be "relevant" to a query. According to one view, it means that the document is "useful" to the user; while according to another, it means that the document is "pertinent" [7].

Given the fuzzy nature of relevance, the *fuzzy-set theory* about data retrieval has been suggested. A "fuzzy-set" is a set whose membership is not precisely defined, in contrast to the dichotomous "membership" or "non-membership" of an element of a normal set. The fuzzy-set model allows for degrees of membership (e.g., the set of hazardous chemicals), whereas normal set model does not. It is also possible to add weights to the terms in a query, allowing some terms to be more important than others in describing the needs of the user. The fuzzy-set model considers the queries as Boolean logic expressions, that is, involving the connectives AND, OR and NOT [8].

The central theme of the *probabilistic approach* is that the relevance of a given term may be estimated by the probability or the relative frequency of occurrence of that term. A measure of the probability of the relevance of a given document to a particular query may be based on a vector representing that document. Moreover, one can add the principles of relevance feedback to this model; e.g. one can retrieve some documents, let the user decide on the relevance of those documents, and finally resort to the user's relevance opinion to revise the probability estimates, possibly using Bayesian statistics [9]. Probability theories provide a reasonable base for theoretical research in a wide range of topics as term clustering, frequency weighting, relevance weighting and ranking.

The probability approach is related to a *utility model*, based on the decision theories. There are utility values (positive and negative) associated with the retrieval of a document. In this case one decides whether indexing via a given document or retrieving a given document in response to a given query as alternatives in a decision, where the probabilities of relevance can be measured and the utilities for good and bad decisions can be calculated [7].

Another approach from probability theory is the *vector space model*. This model represents both the query and each document as points in the Euclidean space of all the possible index terms, and it assumes that each information item can be placed in a vector space. In fact, this model is based on abstract algebra. Thus, one uses clustering or similarity measures based on space distances between document clusters and the queries. Furthermore, the problem with all probability models is that they generally do not incorporate Boolean operators.

The basic elements of *cluster searching* may be useful for some applications in information retrieval. Clusters are groups of items - usually documents - that are identified as related according to some measure of similarity that considers the characteristics used to represent the items themselves. Croft reviewed the likelihood of errors across all clusters and examined coefficients associating a new document with a cluster using a generalization of the Dice's coefficient (Dice's coefficient is a measure of the similarity among items) [10].

The incorporation of *artificial intelligence (AI) methodologies* into the information retrieval domain is a further development. The fields of artificial intelligence cover engineering, computer science, philosophy, logic, psychology and linguistics, including natural language processing, knowledge representation and expert systems. The processing of natural language for information retrieval is a long-sought goal, while the retrieval of specific information without the need for full natural language processing is the goal of *passage retrieval*. The problems of relating syntactic structures, semantic structures and user concepts are very complex. The examination of AI techniques as applied to information retrieval systems has just started and their practical implementation is still in its infancy. Procedures for analyzing natural language so that it can be used in retrieval systems are limited to medical environments, which are inherently structured. Probably the process of understanding or analyzing natural language is not algorithmic. It is better that the users learn those activities that cannot be represented as algorithms rather than all intellectual activities are replaced [7, 11].

Applications

Model of information system must contain information elements and relationships among the elements. If the information elements are ordered, and each ordered element is assigned an attribute, then one has the essential elements of a relational model (see below). The formalization of this model permits mathematical analysis to be performed - such as determining the equivalence of representations - and it might be used to ensure efficient file design or to establish an efficient syntactic structure for the information system. To improve the effective retrieval of information items, aids as clustered files, ranking algorithms, term weighting and automatic feedback processing have been developed.

A useful index term can represent the content of a document and can also permit the documents to be distinguished to which it has been assigned from other documents. There is not, however, a general theory of document representation. Terms can be chosen statistically, linguistically, by mental experiments based

on expected utility, or by the subjective judgment of "aboutness" made by the indexer. Thus, the importance of a term as a representative of a particular document may be proportional to its occurrence in that document and inversely proportional to the number of documents in the file in which it occurs.

The same method used for the computation of discrimination values may also be used to cluster terms. Such clustering provides a thesaurus that can be used to expand searches and enhance recall. Such vocabularies are used in current on-line searching practice not as a recall device, but as a mean of limiting retrieval [12].

Access methods are techniques, for organizing and retrieving information, oriented towards ensuring search efficiency, whatever the subject or content of that information. The interest is focused on the representation of the items, in the organization of the files and in the specific techniques used to find the desired item. The techniques used to find an item depend on the file structure and, in turn, are dictated by the particular use of a file. Both truncated and derived search keys, e.g., are efficient and often effective. However, occasionally ambiguities may arise. In searching multiple attributes, the use of a combination of a limited number of indexes may be more efficient than the use of multiple indexes when the ratio of retrieval to update is high, when many records have similar keys, when queries are complex and when it is more important to minimize search time than storage [13].

Feedback in information retrieval is a process attempting to improve the query by using relevance or nonrelevance of information. Feedback may be done manually ("query revision") or automatically ("automatic feedback" or "relevance feedback"). Ranking algorithms attempt to measure the degree of a document to a query and to rank the document according to this measure. Ranking algorithms are of considerable help to inexperienced users, to users seeking a few highly relevant documents or to users who need "a first impression" [14].

Databases

A database is a collection of data usually containing text and numbers. Databases have been used to serve multiple applications. They are usually implemented on computers and are characterized by one feature of their design; data structures provide accurate, simplified models of the real world. The data are stored in formats that are independent of the specific application programs to allow for multiple uses and to support future changes. Redundancy is avoided to conserve storage resources and to maintain consistency among multiple uses of the same information. Upon request, a specific group of facts can be extracted from the full collection. Some electronic databases are linked to telephone lines so that they can be reached by dialing through a modem ("on-line databases") [3].

Every database is designed around a central core list of facts. These facts are grouped in discrete "records" that generally contain "fields", each with a different type of information. The records are retrieved from the electronic file by means of a key; a key, in turn, consists of a field, a combination of several fields or of a part of a field.

Electronic databases use one or more query languages to help users call forth particular records. The language used to define the logical structure of the database is a data definition language (DDL) and can be used to specify data items and their interrelationships. A language used to access the data in the database is a data manipulation language (DML) and can be used to do insertions, deletions, updates and retrieval. When a DML is a self-contained language, it is often called a query language (QL).

The architecture of a database can be described in terms of three separate levels. The general logical description of the entire database is the conceptual level, commonly called a schema. Subsets of the schema that contain only the data needed for particular applications are called subschemas and describe the external level. The description of the physical storage structures used to store the database on a specific computer system is the internal level [4]. Logical and physical descriptions refer to the way data appear to the application programmer or end user and to the way the data appear to hardware and to data management software, respectively [3].

Four generations of data models for databases can be identified:

a) *primitive data models*, which are based on file structures;

b) *classical data models*, which include the hierarchical, relational and network models.

The *hierarchical* databases are named so because in each group of records one field has been designed as the master field and the others are subordinate to it. In hierarchical databases, therefore, every set of relationships among records is restricted to a hierarchy (or a tree). Examples are the organization charts of a corporation or the biological classification systems. To find any individual record, the database program must tediously work down through the chain of categories. In the *network* model, relationships among record types are not restricted to a hierarchy, but can form a network or graph. In this model multiple connections can be established between files and the interrelationships among record types are represented in terms of sets. The network model is somewhat more flexible than the hierarchical one and one-to-many relationships can be represented easily. *Relational* databases are analogous to tables in which the row represents facts and the columns represent attributes.

This kind of database is called "relational" because such tables are based on relational mathematics. The hierarchy of fields is abolished and all fields can be utilized as keys to retrieve information. Multiple-file relational databases are used when information is too complex to fit into a single flat-file table [4, 6];

c) *semantic data models* (direct extension of classical models, mathematical models, irreducible models and semantic hierarchy models), which enable more complex relationships to be described than it is possible with classical data models (but they may be more difficult to use);

4) *logic models*. The use of logic as a formalism for databases is that it is possible to use a single formalism to describe the schema and to express queries, integrity constraints and procedures. The use of logical inference rules allows new fact to be inferred ("deductive databases").

A more recent concern is with the implications of growing rates of hazardous substances. A full illustration and a detailed list of major databases on chemicals can be found elsewhere in this issue.

Submitted on invitation.
Accepted on 8 September 1994.

REFERENCES

1. MARTIN, J. 1983. *Managing the data-base environment*. Prentice-Hall, Englewood Cliffs (NJ).

2. GOLDSTEIN, C.M. 1984. Computer-based information storage technologies. *Annu. Rev. Inf. Sci. Technol.* **19**: 65-96.

3. HUFFENBERGER, M. A. & WIGINGTON, R. L. 1979. Database management systems. *Annu. Rev. Inf. Sci. Technol.* **14**: 153-190.

4. EASTMAN, C.M. 1985. Database management systems. *Annu. Rev. Inf. Sci. Technol.* **20**: 91-115.

5. TEOREY, T.J. & FRY, J.P. 1982. *Design of database structures*. Prentice-Hall, Englewood Cliffs (NJ).

6. LESK, M. 1984. Computer software for information management. *Sci. Am.* **251**(3): 114-125.

7. BOYCE, B.R. & KRAFT, D.H. 1985. Principles and theories in information science. *Annu. Rev. Inf. Sci. Technol.* **20**: 153-178.

8. ROBERTSON, S.E. 1978. On the nature of fuzz: a diatribe. *J. Am. Soc. Inf. Sci.* **29**(6): 304-307.

9. VAN RIJSBERGEN, C.J., HARPER, D.J. & PORTER, M.F. 1981. The selection of good search terms. *Inf. Process. Manag.* **17**(2): 77-91.

10. CROFT, W.B. 1977. Clustering large files of documents using the single-link method. *J. Am. Soc. Inf. Sci.* **28**(6): 341-344.

11. MCGILL, M.J. & HUITFELDT, J. 1979. Experimental techniques of information retrieval. *Annu. Rev. Inf. Sci. Technol.* **14**: 93-127.

12. SIEVERT, M.E. & BOYCE, B.R. 1983. Hedge trimming and the resurrection of the controlled vocabulary in online searching. *Online Rev.* 7(6): 489-494.

13. SHNEIDERMAN, B. 1977. Reduced combined indexes for efficient multiple attribute retrieval. *Inf. Systems* **2**(4): 149-154.

14. NOREAULT, T., KOLL, M. & MCGILL, M.J. 1977. Automatic ranked output from Boolean searches in SIRE. *J. Am. Soc. Inf. Sci.* **28**(6): 333-339.