Information Services & Use 6 (1986) 161–168
North-Holland

Design of personal information retrieval systems *

Maristella Agosti
Facoltà di Scienze Statistiche, Università di Padova, Italy

and

Franco Spilotro
Enidata, Divisione Prodotti TEMA, Bologna, Italy

This report presents an operational definition of 'personal information retrieval systems', and describes design principles of current applications and existing systems, current trends, and research guidelines for general design principles.

1. Introduction

In recent years, the falling prices of micro- and minicomputers have led to their use by previously excluded categories. This has resulted in the diffusion of computers in the home as well as their use in new applications in the office, in research laboratories and in management. These new applications of computers has led to an increasing demand for the development of automated information storage and retrieval—that is, tools for developing applications on a different scale than those developed in the 60s and 70s. These new tools are required by users as a support for traditional information retrieval (IR) operations, as well as supporting new requirements for these types of applications.

Such new software tools may be called "personal information retrieval systems" (PIRS), but there is no unified view of the characteristics which are peculiar to them. Such a view could be used as the initial step for the design of PIRS. This paper is concerned with the presentation of this view and of an initial analysis of the design principles of PIRS.

The first part of the paper is devoted to an operational definition of PIRS. This definition derives from the applications environment, user categories, document

* Version of a paper presented at the Second International Conference on the Application of Micro-Computers in Information, Documentation and Libraries, Baden-Baden, FRG, 17–21 March 1986. Permission for publication here has been granted by the Deutsche Gesellschaft für Dokumentation.

0167-5265/86/$03.50

base, and the functional characteristics which are required from a PIR system. The functional characteristics of a PIRS are analysed using the pertinent subset of the reference scheme for PIRS which has been developed by the authors and which has been published elsewhere [2]. The second part of the paper covers an initial analysis for the identification of the basic design principles which will allow the development of personal information retrieval systems.

2. An operational definition of PIRS

It is difficult, if not impossible, to give a precise definition of PIRS, because there are neither quantitative nor qualitative parameters which can be used to delimit exactly a "personal" application of information retrieval and thus to distinguish a PIR system from other IR systems.

It is possible, however, to give an operational or dynamic definition of PIRS, and of the personal applications of IR, if different aspects are considered together. The next four subsections are devoted to the presentation of these aspects. It should be noted that these four aspects need to be considered together for defining an application or a system as "personal".

2.1. PIRS application environments

Some of the hardware and software capabilities required for personal computing are:

– *A friendly interface*. The operating system must provide easy user/hardware interaction.

– *Easy learning*. The user needs to be able to start working with a new software tool in a short time using only the documentation provided; the user will learn the more advanced capabilities of the tool if it is to be used in a systematic way.

– *Text processing versus data processing*. Powerful tools are needed for different types of text processing (e.g. editing, formating, mailing) or synthetic presentations of data as those produced with spread sheet tools and business graphics packages, while sophisticated or complex numerical processing is seldom necessary.

– *User support*. The user needs to be supported in his normal work environment.

– *Integration*. The results of different software tools should be interchangeable.

– *Portability*. Information, data and software tools may be available on different personal computers which can be placed in different locations (at home, in the office or in different offices, etc.).

– *Communication*. Different users should be able to communicate the results of their processing, even if they use different personal computers.

– *Access control*. The user wants to be completely autonomous in the use of automatic tools and he wants to manage independently his data bases which can store also confidential records (e.g. medical records).

Typical application environments of personal computing are the office, the home, the study of a professional. In most such applications information retrieval capabilities for the management of documents is necessary. The documents may be of a completely different nature depending upon the environment: in the office they are letters, memoranda, minutes; at home they may be letters, management notes, personal records; in the study of a professional they depend upon the type of profession, ranging from students' work to medical records. In the recent past, the need for personal information retrieval facilities has been urgently felt in the office environment because most office work involves the storage and retrieval of information and data and their selective dissemination to sub-groups of the office automation system users. But the request for PIR tools is spreading in other environments, and the market for these systems is steadily growing.

2.2. *PIRS user categories*

As can be deduced from the types of environments in which these systems may be introduced, the users can vary widely: from an office clerk to a lawyer. Thus it is difficult to suggest a traditional list of user categories. It is important to consider the dimensions of the classes of potential users for one of these systems, because the storage and retrieval criteria will also vary from user to user within the same group:

- a single user in home applications (or for some classes of professionals and managers)
- 4–5 users who share the same data base with the same personal computer
- 50–100 final users in an office environment with personal computers that are connected via a local network
- an unknown number of users when a personal data base takes on a certain importance and it is made available over a public communications network.

It can be seen that the number of users is not a parameter which can be used for defining an application as 'personal'. Consequently a user requirements study for a personal system appears to be very difficult. Such a study runs the risk of being too generalised if the types of users involved in the application are not precisely determined before designing the system.

2.3. *Document base managed by PIRS*

Traditional information retrieval capabilities are necessary for the management of small quantities of documents, and it is this management of small quantities of documents which is unusual in IR. But neither the number nor the type of documents which have to be managed by an IR application can be used as decision parameters for the adoption of an automatic personal IR system.

It would be useful to discover a minimum number of documents below which it would not be worthwhile developing an automatic personal IR application. However, this magic number does not exist because the measure of convenience is

not ruled by the market nor by other economic laws. In fact the decision on the introduction of a PIRS seems to depend primarily upon the users' subjective evaluation criteria.

The documents can be of very different types: letters, memoranda, minutes, bibliographic references, addresses, graphics. Some different considerations may help in deciding upon the adoption of a PIRS. These considerations are based on an evaluation of the quality requirements which are necessary for a useful functioning of a PIR application. Many parameters for the design of a real PIR application can be derived from them:

- One major problem of a personal IR application may be the creation of a useful base of information; tools for the automatic loading of pertinent information from external resources may be of great value.
- The insertion of new information should be as automatic as possible and the time required for this operation should be considered as a design parameter.
- The life cycle of the managed information should not be too short: if the stored information is rapidly changing, or becomes obsolete in a short time, the updating may cause the failure of the application.
- Capability of interfacing different media: telex, microfilm, microfiches archives, laser disks.
- The capability of reproducing in the PIR system classification criteria (UDC, Dewey) and cataloguing rules (ISBD, RICA [regole italiane di catalogazione per autore]) that already exist.

Consequently, a PIR application is more likely to be successful if it manages data which do not require frequent updates or, if frequent updates are required, it should be possible to make them using automatic procedures or by copying information from other data bases in a selective way.

2.4. Functional characteristics

The user requirements for the functional characteristics of PIR systems are different from those of traditional IR systems because of the different types of user categories which have been shown in section 2.2.

Thus a tool needs to be developed for presenting the characteristics of these systems. Since such a tool was not available, it was necessary to develop one. The authors have developed a reference scheme for the presentation of the functional characteristics of traditional IR systems [1]. Two similar schemes are the "Feature catalogue of relational concepts, languages and systems" of the ANSI/X3/SPARC DBS–SG Relational Database Task Group (RTG) [4] and the functional criteria for describing the facilities required by the user of a text filing and retrieval system [3].

That initial reference scheme has been redesigned to include the more pertinent aspects of personal information processing. The result is the Reference Scheme for Personal IR Systems [2]. A subset of this scheme is being used here to identify and introduce the functional characteristics of PIRS. The sectors of this subset of the scheme are used in the following as headings which introduce the functional

characteristics of PIRS. It was necessary to precede the presentation of the functional characteristics with two sections, one on the definition of information structures and the other on the data base administration/redefinition facilities supported by the system.

2.4.1. Definition of information structures

Give details on the information structures which are available in the system, that is the data and the paradata structures, and then to give the generation rules of the information model that is supported by the system; provide information on the possibilities of defining an empty form or a schema-like structure for the definition of the information, which is going to be managed by the system; if the system does not support the definition of a personal structure for the definition of the information to be managed by the application, it is necessary that the user has the possibility of choosing some or all types of attributes of a diversified group which is supported by the system. That is it is necessary to have the possibility of defining, in some partial way, the structure of the documents that are going to be managed by the PIRS. The group of attributes types must include: title, author, date, abstract, descriptor or keyword, identification number, document status. For each attribute it is necessary to know: 1) explicitly the characteristics of the domain; 2) if the system has the capability of managing more than one attribute of the same type in one document (e.g. the author type).

One of the chosen attributes must be definable as a primary key and some others as secondary keys, that is it has to be possible to define the set of attributes that are inverted by the system; inform the user whether management of the keys can be dynamic and related to the evolution of the application: definition of new keys during the life of the application and the redefinition of the data base (see section 2.4.2); inform whether it is possible to construct a dictionary of synonyms of the values of the attributes and whether the system is able actively to use this dictionary in the management of the application.

2.4.2. Data base administration / redefinition

Facilities for the management of updated versions (or instances) of the data base should be available. That is the facilities that are available if it becomes necessary to change the data base definition during the lifetime of the application. Details on the regeneration and on the reorganisation procedures should be useful together with details on physical characteristics and constraints on developing real applications. Information on the maximum number of documents which are interactively manageable by the system should be given, together with the way this maximal number is related to the average dimension of the type of definable documents and on how it is related to other internal factors of the system. These data should be given for the different types of hardware configurations which can support the system.

2.4.3. Functional characteristics

2.4.3.1. Facilities for altering an instance of the data base.

– The insert facilities give the entry choices: local terminal, optical character

reader (OCR), remote terminal. The facility for the definition of insert operations is characterised.

– The modify facilities inform on the attributes whose values can be changed because of previous input errors or because of changes in their values.

– The delete facilities which inform on the methods of deleting documents from the data base and on the security facility for avoiding unauthorized access.

2.4.3.2. Query language. It is the tool which implements the interface with the user. It is based on a set of operations which are permitted by the information (data and paradata) structures supported by the system (see section 2.4.1)—that is, the selection mechanisms for the retrieval of information: the permitted Boolean operators, the conditionals (e.g. '=', ...) which can be used, and possibly the other types of operators (adjacency, ...) and the possibility of giving an imprecise value of the attributes that are objects of the query (e.g. truncation: *comput** instead of *computer, computers, computing, computation*, ...). It can include: other facilities for defining queries, the query iteration mechanism for sets of results, the possibility of storing parameterized queries which can be subsequently re-used.

2.4.3.3. Presentation of results. The types of results presentation (output, report, sorting of the results before output, ...) which are supported by the system.

3. Current situation

3.1. System classes

An analysis of the present situation leads to the identification of three different classes among existing systems:

Class A – specialized systems

Very few systems are developed *ex novo* and only for research. This approach on the one hand points towards the resolution of specific requirements (multimedia, sophisticated questioning language, etc.); on the other hand it is possible only through the development in advanced environments, where are available the instruments necessary for developing the functions required (data-base management, compiler generators, lexical analyzers etc.).

Class B – hybrid systems

Many existing PIR applications are partly based upon generalized systems or are, to varying degrees, of general applicability. From this point of view there has been a trade-off (difficult to quantify) between the tendency of developing *ex novo* systems, and the tendency to produce systems which exclusively use existing products: editors, query interfaces, retrieval procedures, data-base management systems for implementing the data structures. The parameters used in the comparison will be both of a strictly technical nature, e.g. amount of memory used compared with search speed, as well as of an economic nature, e.g. time and

development costs compared to the benefits of the application, evaluated not only in quantitative terms.

Class C – generalized systems

Few generalized packages exist on the market, and these are severely limited by their lack of flexibility in adapting to specific applications (paradata structure, document support, maximum number of documents).

3.2. System architectures

There are significant capabilities for distributed applications:

1) PIR systems can connect to the corporate Information System—that is, personal computers linked to a centralized system to exploit the greater IR capabilities in both quantitative (higher speed and greater online memory) and qualitative (more sophisticated user interaction and query) terms, and the possibility for the transfer, revision and storage of the search results, or parts of the centralized document base in local files.

2) PIR systems can be linked to graphical networks for access to (international) online documentation services.

3) PIR systems can be connected in local-area networks, with the possibility of document exchange, or with a file server for storage in a common document base.

4. Trends

This section presents current trends in the design characteristics with reference to the three main classes of systems identified in the previous section.

Class A

This class shows trends towards very specialized systems—that is, systems dedicated for the solution of specific problems, for example:

- storage capacity
- multimedia document support
- user-friendly interface (even natural language)

Such systems usually require powerful and sophisticated workstations and, in a few years, artificial intelligence tools.

Class B

This class corresponds to the use of existing subsystems (that is: editors, database, fourth-generation systems) as multifunctional tools (many different

applications can be supported with the same tools). The result is:
- an integrated environment
- a portable system
- an economic solution

Class C

This class corresponds to the tendency for products dedicated to non-expert end-users (users lacking technical support). Products which are easy to install and to use.

5. An initial analysis of the general design principles of PIR applications and of PIR systems

The design instruments and criteria which have been used so far, referred to in the preceding sections, are suitable for a wide range of specific requirements. The context, however, is a very fragmented one as each problem is resolved on an individual 'one-off' basis. It is thus necessary to develop instruments for analysis and design which can lead to a unified concept. In particular, the following research guidelines have been individuated:

- The composition of a library of modular tools which can be integrated to have different alternatives for each specific requirement. It seems necessary to have different levels of quantitative and qualitative performances:
  - retrieval performances (precision, recall, speed)
  - automatic indexing
  - data dictionary
  - thesaurus

Such a library should include existing tools, and, where possible, generalized tools.

- The development of both methodological and modelling tools (for rapid prototyping, for performance evaluation) supporting the application design, hardware configuration, resources dimensioning and installation of systems.

References

[1] Agosti, M. and F. Spilotro. Sistemi di Information Retrieval: uno schema di riferimento per poterli analizzare e confrontare. In press.

[2] Agosti, M. and F. Spilotro (1982). A reference scheme for personal IR systems and its use for the presentation of their characteristics. Internal report.

[3] Newton, S.J. (1983). *Text Filing and Retrieval Systems*. NCC Publications, Manchester.

[4] Schmidt, J.W. and M.L. Brodie (1983). *Relational Database Systems.* Springer-Verlag, Berlin–Heidelberg.