RESEARCH INFORMATION SYSTEMS

Algimantas Juozapavičius
Vilnius University, Naugarduko 24, 2006 Vilnius, Lithuania

Abstract

This survey presents an overview and study of features and capabilities of information systems, which are suitable to support research work in all stages of it: developing of ideas, design, simulation, calculation, searching and exchanging of information, etc. It discusses a wide-spread attempts, purposes (and objectives) to integrate such systems, relying on capabilities of Internet. The role of methods like multimedia, algorithms of spatial data, data compression, telecommunication, in design and developing research information systems, are discussed too. This servey reflects main points in integration of information search and retrieval procedures with other research activities.

An information system is a collection of components that accumulates, processes, stores, analyses, and disseminates information for a specific purpose [1]. Research information system is dealing with specific information (research data, scientific and technical publications, patents, similar information) and in a specific way (sophisticated and prolonged calculations, publication and information activities). Like any other system it processes inputs, produces outputs, operates feedback mechanism and is functioning within an environment. Research activities of professors, academicians, students are bringing a lot of specifics to the design, implementation and development of information systems.

The components of IS are:

- **hardware** – a set of devices that accepts data and information, processes them, and displays them;
- **software** – a set of programs that enables hardware to accept, calculate and process data;
- **database** – a collection of related files, tables, graphics, other data, that stores data, results of calculation, and associations among them;
- **network** – a connecting system that permits the sharing of resources and data by different computers;
- **procedures** – a set of instructions about how to combine the previous components in order to process information and generate the desired output;
- **people** – the most intelligent part of the system;
- **purpose** – a most common purpose is to provide solution and information to a scientific or research problem, and enhance activities for improving productivity, quality, and competitiveness of research;
- **social context** – an understanding of the values and beliefs that determine what is admissible and possible within the culture and ethics of the scientific community.

Research information systems have local and global subsystems [2]:

- **local** – to accumulate and process research and laboratory data in a local computer environment, and to disseminate them among colleagues;
- **global** – to find out and supply publications, similar information related to research projects and achievements, researchers, companies, institutions etc.

Research information systems are more and more relying on information technology (IT) and its trends:

- **cost-performance ratio** (there are evaluations that in 10 years the cost-performance of computers vs. manual work will improve by a factor of 100);
- **information superhighways** (wide installation of fiber optic computer networks is facilitating new ways to manage information);
- **networked computers and client/server architecture** (this architecture is predicted to dominate IT and will allow interconnection of software and hardware in different computing environments);
- **graphical interfaces** (the domination of them provides users with direct control of visible objects and actions to replace complex command syntax);
- **storage and memories** (large memories will enable the use of multimedia and emerging computer technologies);
- **multimedia** (the integration of various types of media will improve calculations, experiment control, decision making and training);
- **object-oriented environment** (this innovative way of programming computers is expected to significantly reduce both the cost of building and of maintaining information systems);
- **emerging technologies** (the increased capabilities of computers enable the implementation of expert systems, natural language processors, neural computing, spatial databases and other powerful intelligent systems);
- **compactness and portability** (the small sizes and portability of computers allow users in the field to enter data and to attach computers to various laboratory devices).

At a present stage of development of IT, the most important issues in research information systems are: **multimedia, hypertext, hypermedia, virtual reality**. Multimedia merges the capabilities of different input/output devices integrating experiment data, graphics, voice, other media into one interactive application. Hypertext and hypermedia address to a set of navigational techniques, incorporated in multimedia software tools. These issues (especially including virtual reality) give sophisticated and powerful tools for modelling of processes in research and education (**representation of complex mathematics, virtual physics laboratory, modelling of earthquakes and galaxy configurations, etc.**) as well as in industry (**automotive and heavy equipment industry, architecture, visual arts, medicine, etc.**).

An information system, implementing many of attributes listed above and suitable for research in a lot of aspects is Internet. The Internet links millions of users

(and researchers too) around the world. Among other activities, the Internet is used to tap into thousands of large databases and search through them. Most university libraries offer database search services using the Internet. Logically, all services of Internet could be considered as a global subsystem of any research information system. It is also an ideal way to interconnect various research information systems for search operations and exchange of data.

The services of Internet essential for data and knowledge exchange are:

- **e-mail** – a possibility to send message from-address-to-address like any other surface mail;
- **news groups** – a possibility to "subscribe" interesting information and to post it on "bulletin board";
- **ftp** – a possibility to send/receive files from-address-to-address;
- **telnet** – a possibility to remote log-in with interactive sessions whatever computer is used;
- **gopher** – a system to access any type of textual information, based on a set of hierarchical menus and submenus (many levels);
- **world-wide-web** – a method for providing distributed information, where documents are hypertexted (and hypermediated) so enabling user to go from one item to related items without a need to go through menus or any other addressing method.

The Internet provides very good possibility to cooperate in research work on-line. There are a lot of bibliographies and other information and documentation enabling user to find all necessary ideas and what is done in the field of interest. Using e-mail and ftp services it's possible to exchange the results achieved or do research work simultaneously in a very quick way. Internet is an ideal environment for "invisible" groups of researchers to do their studies. It also gives a possibility to integrate even networks of computers (information systems), not just people.

The history of Internet is also related very much to research and development [3]. The prototype of Internet began with a U.S. Department of Defence Advanced Research Projects Agency (DARPA) contract initiated in 1969. A need for such project was that the existing virtual-circuit-oriented data communications were too centralized. The DoD expected to need general-purpose, peer-to-peer communications at high data rates between widely differing computers in the future, and directed researchers accordingly. The evolution of services developed from the beginning of DARPA contract to the nowadays Internet is as follows:

- **creation of TCP/IP** – a focus to connectionless networks, implementing new protocols on various host systems, and applications allowing end users to access the network, services as remote login, file transfer and electronic mail developed;
- **growth of TCP/IP** – a development of commercial and standartized LANs and distribution of 4BSD UNIX, both including a family of TCP/IP protocols, were widely adopted by workstations and minicomputer vendors and gave a great impetus in the commercial research and engineering sector;

- **port TCP/IP to PC** – a widely accepted distribution of electronic mail, file transfer, gopher services, many different implementations and interoperability issues, covering various types of computers and operating systems;
- **recent TCP/IP protocol family** – a new technologies (like authentication services, electronic data exchange, privacy-enhanced mail, etc.) are adding existing services of Internet, causing its explosive growth, both in geographic range and number of users.

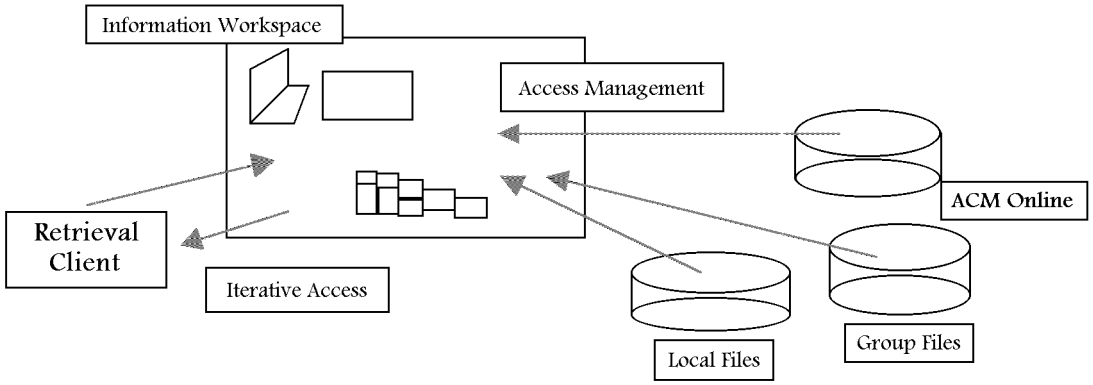
The Internet is also a domain for research activities in the subject of information search and retrieval. A huge collection of information distributed in heterogenous environment of Internet is an perfect universe to test miscellaneous ideas how to search data and conceptions effectively. There are a lot of so-called "search engines", based on different approaches to searching and indexing: **menus and submenus, thesaurus, classification schemes, agents, databases with various indexing strategies, etc.** Such search facilities as Altavista, Yahoo, Lycos, Infoseek, Crawler, SavySearch, etc. (there are more than 130 search engines on Internet) are acquiring various seek procedures, with topical links, multiple databases, good searching parameters, comprehensive data, evaluation of information methods. The search queries are to be based on combinations of words, logical and proximity operators, wild-card symbols, categories of data and addresses, special type of links. As data volumes, user base and data diversity are growing, search facilities, or more exactly, resource discovery tools must scale with the diversity of information systems, number of users, size of information space. The special efforts in this realm are expected to be focused on scalable content-based searching algorithms and on servers specialized to support particular user communities (e. g. academic and research community) [4].

One of potentialities for resource discovery tools based on possibilities of Internet are information networks, representing special interest groups. There are a lot of content-based servers and agglomerations of links, like CAIN (Computer Algebra Information Network), Chaos Group (Network for Nonlinear Dynamics), SymbolicNet, etc. These groups are providing content-related data: topics of research, membership-lists, calendars of special events, publications or referencies to publications, bibliographies, sponsoring funds, etc. The Internet is moreover a testfield for approach-based projects, like "Digital Libraries" [5]. Libraries exist in many forms and are of many types. Adjoining traditional libraries, there are object-libraries, code-libraries (in programming), types of multimedia libraries (image, audio, digital video libraries, etc.), digital collections (residing in databases, knowledge bases, text bases, gopherbases, world-wide-web). Much of the power of the digital library is the flexibility it permits in allowing processing of collections of tangible objects and its electronic representations. However, the knowledge developed over the years is quite flexible too, and it is feasible, perhaps even desirable, to apply it also to the collections of things without direct physical analogs, for example, algorithms, real-time data feeds, computational states, relationships among versions of a physical object showing the historical progression of an idea, multimedia annotations, and tours. Of course, such kind of goals require a lot of afforts to investigate and implement. Techniques no doubts discovered in completion of such project will make a huge influence to research information systems.

Research information systems, while resolving the dilema of information search and retrieval are influenced by commercial information networks too. Many of such networks like STN, Dialog, Orbit, etc. are offering information, which cannot be find in open cyberspace of Internet: patent descriptions, detailed publications, technical releases and drawings, full-text articles. Coverage of commercial databases embraces a wide variety of scientific and technical topics, as well as business and industry. The information of these databases is usually placed in research information systems for applications to follow and is affecting these systems in two aspects: regarding the content of data found and methods of representation of data.

The digital library and commercial information networks are illustrations for the specific concept of research information system. Projects to analyse this concept are to be implemented via Internet and are based on points of view:

- heterogeneous sources;



In order to implement these concepts into real life research methods from mathematics, computer science, artificial intelligence are applied. Retrieval client is operating in user/system environment and is managing "query server" and "display server", allowing to formulate queries for fulltext/attribute/multimedia searches, document match requests, and receive documents, nodes, edges, as well. Iterative access allows for client to compare search results and queries on some semantic background, to reformulate queries and produce retrievals again. The "document server" and "semantic knowledge server" are serving "information workplace" in order to deal with databases of documents and to manage hypertext and hypermedia links, both effectively. Of course, "access management" is designed to direct heterogeneus sources, coming from local, group, Internet, X.25, etc. environment.

The multimedia databases are giving rise to much more complicated models for content-based information search and retrieval. The Dexter-Model, a standartized model for interconnecting multimedia documents, consist of three layers and two interfaces:

- **run-time layer**;

- **presentation specification (interface)**;
- **storage layer**;
- **anchoring (interface)**;
- **within-component layer**.

The class-hierarchy of multimedia data includes:

- **timeless component (text, still image)**;
- **with-time component (audio, video)**.

Multimedia data are large in volume, in comparing to ordinary texts and programs. While transfering them via communication lines, they are to be compressed. Nowadays procedures of compression include four stages, with specific possible modi:

- **refreshing of multimedia multidimensional document (type of file, type of format)**;
- **processing of multimedia multidimensional document (prediction, FDCT)**;
- **partition of multimedia multidimensional document**;
- **entropy-coding**.

Processing of multimedia documents in information systems is much more complicated and requires sophisticated theories and algorithms. Sampling procedures, extracting different levels of pixels, according to the base of coloring, are subject for various attempts of research. The compression algorithms are well-known and well-developed for unidimensional data, there is only one for twodimensional data (discrete cosine transformation), and there are no algorithms suggested for multidimensional data. Such kind of questions are closely related to problems like content-search in imaging databases, compression and decompression of data presented in graphics formats, etc.

The primary role in solution of problems above belongs to data structures and corresponding algorithms. There are a lot of classical algorithms developed for one-dimensional data, allowing to search and retrieve data in an efficient and comfortable way. These algorithms without any exception are exploiting the peculiarity of such data – the data always could be totally ordered. Multidimensional data, appearing in geographic information systems, pictorial databases, very large data bases, etc. require for search and retrieval much more complicated procedures. The role of data structures are caused by necessity to have support for search operations at the physical level. Typical operations for spatial data bases include the point query (find all objects that contain given search point) and the region query (find all objects that overlap a given search region). Many years of research in multidimensional data bases have resulted in a great variety of multidimensional access methods to support such operations. The main goal of data structures and corresponding algorithms is to manipulate analyzed multidimensional data and that unanalyzed images or multimedia objects are only handled as the source from which spatial data can be derived. The challenge for the developers of spatial information system lies not so much in providing yet another collection of special-purpose data structures.

Important issues in the context of research information systems include the handling of spatial representations and data models, multidimensional access methods, as well as pictorial or spatial query languages and their optimization [6]. The data structures for multidimensional data are also important in various areas of research, not just in information systems:

- **systems of symbolic and algebraic computations**;
- **computer graphics**;
- **theory and praxis of telecommunications**;
- **systems for design and engineering**.

The theory and development of data structures are based on methods, coming from algebra, geometry, mathematical logic, discrete mathematics, programming, computer architecture, etc.

Algorithms based on data structures depend very much on the sort of memories, data are located in: primary memory (RAMs); secondary memory (hard disks); ternary memory (CD-ROM, streamers, DVD, etc). Data of multimedia information systems are located in all three sorts of memories mentioned. Spatial data could be dimensioned as follows:

- **spatial data** (1-D: audio, temperature, time series, etc.; 2-D: raster images; 3-D: video/animation, holograms/voxel; 4-D: tomograms, modells of holograms video, modells of weather forecasting; 5-D: defined by ISO/IEC standarts of visual PIKS data; ¿5-D: multidimensional analysis of operating data in finance institutions);
- **data of computer geometry**;
- **data of virtual reality**.

The analysis of complexity of data structures and algorithms seeks to evaluate ratio time/memory, as a function of data volume, as well as estimate this ratio (or to find suitable asymptotics). Explicit situation in this field of research is far from being evaluated as good. There are no mathematical or formal models, describing dependency as "almost functional" between major parameters of the complexity of algorithms and basic parameters, such as data volumes or time utilization. Many new methods, originating from mathematics, programming, artificial intelligence, related areas, have to be discovered or applied to this field of human activities in order to force computers to serve people in a much efficient way.

References

1. Efraim Turban, Ephraim McLean, James Wetherbe, *Information Technology for Management*, New York, Chichester, John Wiley & Sons, Inc., 1996.

2. A. Juozapavičius, *New Information Systems – Impact to the Society* (Article), *In: Proceedings of the International Conference "Science in a Small State"*, Vilnius, 1996, 6 p.

3. *PC/TCP Interoperability*, North Andover, MA, FTP Software, Inc., 1993.

4. C. Mic Bowman, Peter B. Danzig, Udi Manber, Michael F. Schwartz, *Scalable Internet Resource Discovery: Research problems and Approaches*, *Communications of the ACM*, **37**(8), p. 98–108 1994.

5. Edward A. Fox, Robert M. Akscyn, Richard K. Furuta, John J. Leggett, *Digital Libraries*, *Communications of the ACM*, **38**(4), 1995.

6. Volker Gaede, Oliver Guenther, *Multidimensional Access Methods*, Institut für Wirtschaftsinformatik, Humboldt-Universität zu Berlin, 1996.