ITR8 Lecture 1

Data & Information, Storage & Retrieval

Thomas Krichel

2000–01–18

Structure

1. Models of Information Retrieval and Information Storage
2. Korfhage's model of the Information System
3. A very brief history of Information Retrieval

Reading

Korfhage page 1 to page 9 "... are expressed."

BYRN page 1 to page 2 "... as possible."

Korfhage page 16 "Exercises"

BYRN sections 1.1.2 and 1.3

Definition and Aim

Definition: (adapted from BYRN, page 1)

"Information Storage And Retrieval (ISAR) deals with the representation, storage and organization of, and access to information items."

Aim: give users easy access to the information that they need

Basic Model of Information Retrieval

User has an information need.

User formulates a query.

User transmits the query to an information system.

User judges the relevance of the response.

User Tasks

User needs to translate the information need into a query that the system can process. A query is usually a set of words.

The user need to judge the relevance of the information items that are returned. This is usually done by browsing.

The user may make mistakes at both stages.

Data Retrieval Tasks

Data retrieval task with the retrieval of information items that correspond to the query.

It is a part of the information retrieval process.

It is usually done by computer.

It is not error-tolerant.

Data and Information

- Data are information items as stored in a ISAR system. They are impersonal. They are organized in a way that is independent of any particular user.

- Information is data that has been matched to a user need. Any data is information only for a group of users. The organization of information tends to be personalized.

If the user who stores data and the one who retrieves it are not the same, there is problem. One user may store data that he thinks is information, but the user who retrieves data does not find that it is information.

Basic model of information storage

Reality is independent of human observation. It has a physical and abstract side. It is not known in its entirety.

Individual humans make observations about parts of reality. These are often are sufficiently common to be able for them to communicate about them.

Human observers can encode and store descriptions of reality in an information system. These form a collection of information items in the information system.

Storage for Retrieval Principle

A good information system will have the information items stored in such a way as the make the retrieval of these items easy. That means that it will be easy for the user to translate her information need into a query that can be send to the system.

To set up a good information system, the thinking process must address the second (retrieval) stage *first and the first (storage) stage* second.

(This derives from properties of optimal action over time.)

A Model of an Information System

Korfhage distinguishes two "portions" of the information system, the ectosystem and the endosystem.

The ectosystem are all the factors that are not under the control of the designer of the information system.

The endosystem are all the factors that the designer of the information system can control.

The Ectosystem

Three main components

1. Users, who either stores or retrieves information items.
2. Funders, who pay for the information system design and/or operation.
3. Servers, who operate the system as such.

Many projects that have an ISAR system as central component fail because they do not take into account what the members of the ectosystem want.

The Endosystem

Three main components

1. media for storage such as hard-copy, magnetic disks, CD-ROM
2. devices for the storage of media as shelves, computers, scanners
3. encoding algorithms, e.g. classification schemes
4. user interfaces, e.g. command-based, GUI, physical library

These choices are not independent from each other.

Information System Evaluation

Information system performance depends on the choices in the endosystem, but the result is judged in the ectosystem.

- Users care about speed, accuracy, relevance of documents returned.
- Servers care about the efficiency of the system, i.e. how much effort does it take to maintain the system operating.
- Funders care about the overall economy of the system, i.e. if it is a cost-effective solution for the task that it accomplishes.

An efficient system may be popular with users despite meeting few of their information needs.

In the beginning there was the index . . .

BYRN say that there has been ISAR for 4,000 years.

Large set of textual data have been appearing in books for a long while. Information within books has been organized using indexes. These are tables that list terms and location of documents where the term can be found.

In the past, these indexes were created by librarians. Recently such indexes have been made by a computer. Automatic indexes do take little account of user needs.

and then came the Web

BYRN "The Web is becoming a universal repository of human knowledge", because it is "a new publishing medium available to everybody".

But to find something on the web,

- it has to be there
- there must a be a way of finding it

The web search engine is not much different from an old fashioned index. The difference is essentially a matter of scale.

How good are the web and its indexes as tools to find answers to problems?

Korfhage's exercise questions I

1. What is Donald Knuth's first published paper?
2. Who invented the water bed and where was first mentioned?
3. Produce a bibliography on American joinery in the 19th century.
4. Find the earliest documented incidence of the collapse of a structure due to resonance.
5. Who wrote "A Dissertation Upon Roast Pig" and he or she get the idea?
6. What was the ruling of the Indiana legislature on the value of pi π and why did it occur?
7. What did John von Neumann have to say about common set traversals for three or more collections of sets?

Korfhage's exercise questions II

1. Who chose the letters K and W as the first call letters of American radio and TV stations and why?
2. How are the digits of the social security number determined?
3. Who invented cribbage?
4. What information is available on the toxicity of Melongena?
5. Who scored the last home run for the Pittsburgh Penguins?
6. Why was it customary to carve a crescent moon in the door of an outhouse?
7. Who was Joe Pye?
8. Who said "If I was two-faced, I would be wearing this one?"
9. Who invented the ballpoint pen, and when?

Assignment: prepare a presentation

How long did it take you to find the answers. Time out after 2 hours.

Please note fairly precisely what you do. How did you find the answer on the web?

What are your recommendations to finding the answer to a specific question on the web.

How do you judge the overall performance of the web as a retrieval tool.