Introduction to Information Retrieval

Chapter 3: Retrieval Methodology

Dr. SAID KADRI

Associate Professor

Department of Computer Science, Faculty of Mathematics and Informatics, University Mohamed Boudiaf of M'sila

E-mail: kadri.said28@gmail.com

Website: https://kadrisaid28.wixsite.com/sgadri

Table of content

Information Retrieval Methodology

Retrieval Process:

1. Identification of needs.

- Define precisely the subject to look for.
- Define the pre-acquired knowledge.
- If the subject is not clear, try to explain it.
- Define the general context of the subject.
- Define authors if that is possible.

2. Selection of relevant sources of information

- Information sources can be multiple, heterogeneous, complementary and redundant.
- Give priority to international bibliographic DBs which are well structured, ordered and its scientific and technical information is credible, free or paid.
- Search on the web allows access to huge amounts of information.

3. Analyze the structure and the query syntax of DBs

- Bibliographic DB contains bibliographic references organized in fields describing the document (titles, authors, keywords, year of publication, summary, language, type of publication, etc.)
- Offer search modes on these fields ➔ then, use index and thesaurus in the search, combine criteria with logical operators (and, or, no, except), refine the query, view and export the results.

3. Choice of keywords

- Convert the subject into simple concepts and then into keywords.
- Check the existence of these keywords in the thesaurus or the index of the DB.

4. Building search formulas in DBs

- Establish a logical formula by linking the selected keywords using logical or proximity operators (close to, far from, in front of, next to, ...).
- Launch a first test query to validate the search criteria.
- Refine the final query according to the results obtained by the test query.

5. Adopt a search strategy on the web

To optimize the search:

- Use the advanced search mode (search engine).
- Use complementary services and tools of the selected search engine (google, altavista, yahoo)
- Evaluate the obtained results: nature of the site containing the searched information, date of creation, date of update, identification of the author name and contact details, etc.

6. Extend search to other sources

Diversify the information by consulting:

- Catalogs of libraries rich in gray literature.
- Open archives that provide access to primary information.
- Journal and conference publisher sites with summaries.
- Websites of organizations, research centers and government agencies putting documents online.
- Official statistics sites.

7. Follow new publications

- Set up a watch on your research theme to receive regularly: new publications, updates, date of events.
- Set up an alert system on important sites to receive news about your theme (use google alert https://www.google.fr/...).

8. Manage the selected information

Generally, information is issued from multiple sources and redundant ➔ using bibliographic management software that can:

- Extract structured references from bibliographic DBs and journal sites.
- Create your database
- Sort duplicates
- Edit your bibliography

As an example of free management software we note: mendeley (https://www.mendeley.com), zotero (http://www.zotero.org/)

Characteristics of information on the web

- Heterogeneous (support, content, language, accessibility: paid, free, subscription)
- Dynamic

Information search tools on the web

1. Search engines

- Based on servers specialized in locating resources on the web.
- Store information in BDDs on specific machines.
- A search engine receives a user query consisting of keywords.
- Uses a robot that travels the web day and night, explores web servers, navigates through hypertext links to collect new pages, or news updates.
- Integrate the collected information in its database and then structured it to facilitate access (indexing)
- The metadata provided during the web pages drafting facilitates the indexing of these pages.

Examples of search engines

- Altavista.com (1995, Digital Equipment Compagny, index 28% of the web)
- Google.com
- Yahoo.com
- Lycos.com
- Bing.com
- Hotbot.com
- Ask.com
- Baidu.com
- Yandex.com

Arabic search engines

- Ayne.com
- Araboogle.com
- Yamli.com
- Arabi.com
- 4arab.net

2. The indexes

- Also called directory sites, are lists of servers classified by themes, allow to search Internet services containing relevant information. The user will only have to navigate in the tree of the themes and sub-themes to find the category that corresponds to the searched theme.

Example:

The Yahoo Index (Yet Another Hierarchically Organized Oracle)

- Developed by two students at the university of Stanford: David Filo and Jerry Yank.
- Search by category of themes or directly by theme.
- Has 14 thematic categories of maximum depth 4 levels for each.
- The information is less exhaustive in the indexes because they index only the description of the site.

3. Specific search engines

Allow to look for information on resources other than the web, such as: news, FTP files, directories, … (ex : dejanews.com ➔ news, Filez.com ➔FTP, …)

4. Meta-tools

- **Meta-index**
- **Single shops**

A meta-index : offers a unified search interface in the form of a single form giving access to several search tools simultaneously (eg meta-search engines, ...)

Single shop: are simple web pages giving access to different information search tools (SE, indexes, directories, ...) (ex: allonesearch.com)

5. News :

Means of communication used to exchange information on a specific theme via messaging, this tool uses forums, news groups that are classified hierarchically by theme. An article posted in a group will be read by anyone accessing this group.

6. Mailing lists:

Are discussions via messaging around various topics. The user receives the information via his e-mail. A user can subscribe to one or more mailing lists.

7. Intelligent agents :

They are software used to automatically or on demand search relevant information for a user on the web, they are able to monitor, detect sites according to pre-established criteria.

Which information with which tool?

Type of information	Tools	Results
General Informations	Standard search engine Virtual libraries	Getways of hypertext links companies Personal websites related to the subject
Experts in domain	Standard search engine Specific search engine (news) Mailing Lists	Name of people Functions mails Areas of interest
Press reviews on the subject	Specific and general press sites	Free Articles Paid Articles Directories Abstracts Scientific Publications
Reports, analyzes	Standard search engine Index Public and government service sites Accessible and free Databases	Reports Theses Memories Scientific Publications presentations
Information Exchange	Mailing lists on the subject News	Answers to questions Names of experts Informal information with added value
Industrial propriety	accessible and free databases Sites of public and government providers	Patents (brevets) Trademarks (marques déposées) General information on the filing of patents
Constraint information	Government sites Specific sites	Regulations Legal Environment Security

Tableau 1. ***Information retrieval according to its type***