université PARIS-SACLAY

Information Retrieval
-
Introduction

Recherche et Extraction d'Information

Anne-Laure Ligozat/Xavier Tannier

What is Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an often unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)

YAHOO!

bing

Google
France

Recherche Google | J'ai de la chance

Reche
Outils

Index

A

Abiteboul S. 799, 948
Abrial J. R. 421
accès aux données distantes – voir RDA
accès direct (index) 846
accès séquentiel (index) 846
accès séquentiel (physique) – voir séquence physique
accepter une prémisse 911
Adachi S. 620
Adiba M. E. 575, 734
Adleman L. 500, 510
administrateur de la base de données – voir DBA
administrateur des données 15
adressage dispersé – voir dispersion
affectation relationnelle 188
    relations cibles 569
agent 707
Agrawal R. 691, 844, 945, 948
Aho A. V. 382, 392, 624, 946
algèbre relationnelle 159
    implantation 603
    objectif 180
    opérations primitives 190, 203
    règles de transformation 181, 592
algorithme de chasse 393
algorithme de réduction de Codd 226
ALL (SQL) voir duplicata
Allen F. W. 424
ALPHA – voir DSL ALPHA
ALTER DOMAIN (SQL) 258
ALTER TABLE (SQL) 100, 261
Alsiana E. B. 873
American National Standards Institute – voir ANSI
analyse incohérente 458, 465
Anderson E. 665
anomalies de mise à jour 345, 349, 356, 378
ANSI 73
ANSI/SPARC 33, 57
ANSI/X3 57
ANSI/X3/SPARC Study Group on Data Base Management Systems – voir ANSI/SPARC
Anton J. 664
appel des procédures distante voir RPC
APPEND (QUEL) 496
applications en ligne 9
arbre de requête 589
arbre de syntaxe abstraite voir arbre de requête
arbres de recherche numérique – voir trie
architecture ANSI/SPARC 33
    vs. SQL 80
ARIES 453
arité voir degré
Armstrong W. W. 326, 328
Arya M. 661
Asbenhurst R. L. 82
assertion (SQL) – voir CREATE ASSERTION
association 12, 405, 406, 410
    OO 790
    récursive 414
association (RM/T) 426
associativité 170
Astrahan M. M. 295, 296, 873
Atkinson M. P. 800, 822
atomicité
    relations 353
    transactions 436, 441
    valeurs scalaire 63, 104, 642
attribut 89, 97
authentification voir mot de passe
autonomie locale 700
autorisation voir sécurité
auxiliaire 427
AVG – voir fonction d'agrégation
axiome 906
axiome déductif 325
axiome de base 900, 921
axiomes de Armstrong 320, 328

B

B-trees 850
Badal D. Z. 534
Banerjee J. 800
Barnes G. M. 872
Barnsley M. F. 884
base de connaissance 839
base de données 3, 10
    avantages 16
    DB2 890
base de données déductive 910
base de données distribuée 55, 695
    principe fondamental 699
base de données expecte 940
base de données catexionnelle 923
base de données intentionnelle 925
base de données logique 940
base de données relationnelle 120
base de données statistique 508
Batory D. S. 813, 874
Bayer R. 478, 851, 875
BCNF 337, 354, 361
BIBA 16
Bockley D. A. 884
Bocch D. 823
Bocci C. 382, 392, 393, 946
BEGIN DECLARE SECTION (SQL) 283
BEGIN TRANSACTION 439
Bell D. 729
Bentley J. L. 884
Bernstein P. A. 365, 419, 479, 534, 576, 729
Bind (DB2) 894, 896
Bitton D. 617
Bjornerstedt A. 880

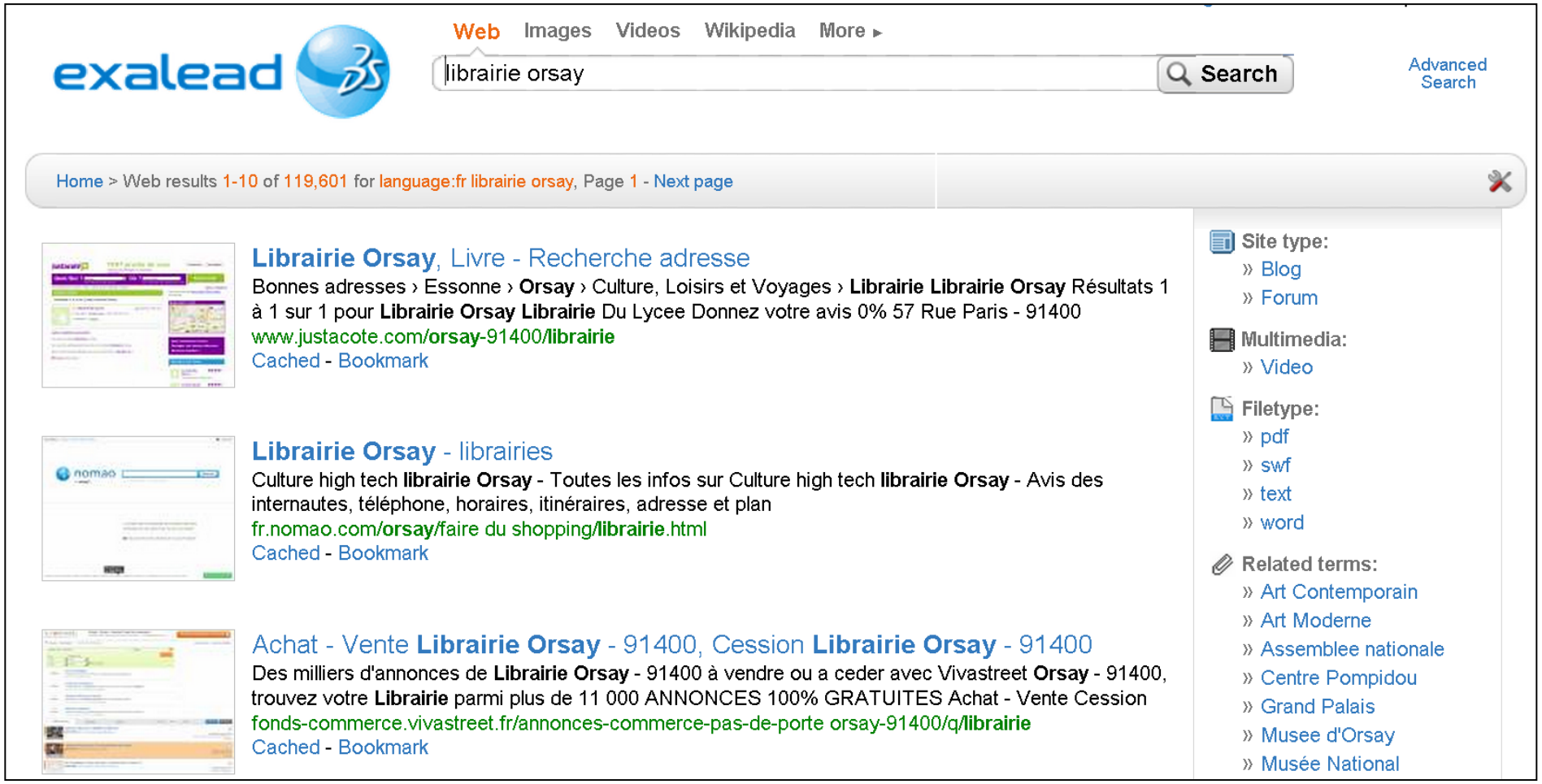
Information retrieval actors

Collection :
a set of
documents

User :
an information
need and/or
a task to perform

IR system: the tool that should find relevant documents for the user information need

Information Retrieval



Information Retrieval

- *Where is the bookstore closest to home?*
- *Who is presently leader of the rugby Top 14?*
- *Quels sont les titres mentionnés à la une du journal Le Monde d'aujourd'hui ?*
- *Que rapporte la une du Monde d'aujourd'hui sur la politique étrangère ?*
- *Quels sont les films qui passent ce soir sur la TNT ?*
- *Dans quels films Jean Rochefort et Philippe Noiret ont-ils joué ensemble ?*
- *Quels sont les logiciels d'installation de logiciels sous Linux/Debian ?*
- *Comment peut-on installer des logiciels sous Linux/Debian ?*
- *What is the English word for "givre" ?*
- *Who was Claude Bernard?*

Questions

- What kind of results are expected?
- How to evaluate the results relevance?
- How to formulate a query?
- …

Information vs. *data*

"**Data** are received, stored, and retrieved by an information endosystem. The data are impersonal; they are equally available to any users of the system.

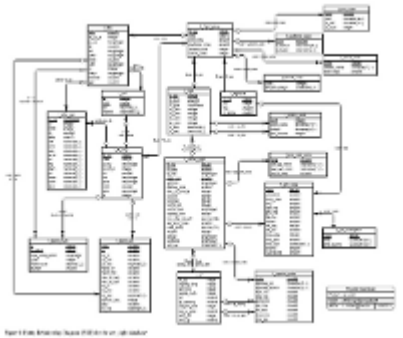
Information, in contrast, is a set of data that have been matched to a particular information need

The concept of information has both personal and time-dependant components that are not present in the concept of data"

(R. R. Korfhage, 1997)

Diversity of information needs (1/2)

- Search for a known element
- The user knows exactly which elements they look for
  - and can recognize them
  - *Example*: search for a bibliography item
  - ➢ Databases (*SQL*, *XQuery*, *etc.*)
- Search for a general information
  - The user searches for information about a subject
  - and may not recognize relevant information
  - or information can be partially relevant
  - *Example*: reforms of research in France
  - ➢ **Traditional information retrieval**



Diversity of information needs (2/2)

- Search for precise information

  The user searched for specific information
  but ignores under which form it is
  - Partial answer not relevant
  - *Example* : What day was president *Kennedy killed?*
  - ➢ **Information extraction and question answering systems**

- Exploration

  The goal is not to answer a particular question,
  but to go through data to discover information
  about a subject or a domain
  - ➢ **Navigation**

Diversity of information sources

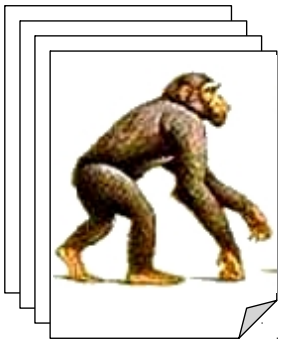
- Location of information
  - Local or distant resources
  - Problems : disponibility, identification, distribution on several sources, format variability (character encoding and content description)
- Nature of the resource files
  - **Databases**: well described formats, non ambiguous query languages (ex : SQL for relational databases)
  - **Annotated files**: more or less described formats, presentation and/or semantic description annotations, query languages (ex : XSLT/XPath for XML files)
  - **Text files**: few or not described formats, known language(s) or not, more or less regularity between documents of a same class, no generic interpretation possible (NLP probleme)

Diversity of problems

- Difficulty to **access, cover and treat**
- Document bases **very large**, distributed on
  - **numerous supports** in **different places**
- Difficulty to define **relevance**
  - How does a document respond to the **user information need**?
  - What is the **relevance**? How to measure it?
- Difficulty to **exploit**
  - Relevant documents may not be in the query language
  - The desired information may not be easy to identity in the document.

IR main evolutions

- In the beginning
  - Small and structured document bases
  - Acces by metadata, and rarely by whole text
  - Use of specific languages by specialists

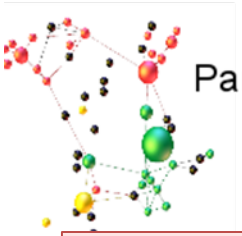


- Nowadays
  - Digital multimedia documents
  - Various representation formats (raw text, HTML, XML, PDF, RTF,...)
  - More and more unstructured data
  - Huge quantity (Web...)
  - Social component

Web IR

- On the Web: massive use by non expert users
  - domain with major economic importance
  - typical query: a few keywords
  - users adapt to existing tools
- Part of the Web is not directly accessible (invisible web, including restricted access pages and dynamic pages)
- Strongly multilingual information: the documents containing query information may be in different languages
- Information not always reliable
- Information visualization particularly important: result ranking, extract presentation, relevant segment extraction etc.

Information Retrieval



Parcours du Web
(robots)

Analyse des
documents

Indexing
(document model)

dynamic vs. static
collection

Fonction de similarité

Search
model

Evaluation

Query

Analyse du
besoin

IR difficulties

Humans and languages

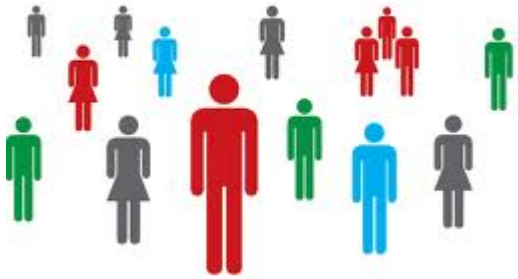
IR difficulties : the human factor

- The user's information need may be vague and is always subjective.
  - The information loss between the real information need and its expression can be large.
  - The relevance of a document to a query is a variable notion, complex to define.

- ⇨ There can be no perfect information retrieval system.

- A system evaluation is different from usual CS performance criteria

- Humans are subjective, volatile and use natural language



IR difficulties : the language factor

- Contrary to artificial languages, natural languages is:
  - **Implicit**: not eveything is written in texts and comprehension requires important knowledge about context and world
  - **Redundant**: language offers numerous ways to formulate the same content
  - **Ambiguous** : a same text can often be interpreted in different ways

- IR is even more complex as:
  - Words can play different roles in texts
  - Sense items can be words or group of words (terms)

➢ Difficult to formulate an information need (information loss between need and query)

Implicit character of language

- Language knowledge and language **conventions**

  Q : *Le voisin est-il chez lui ?*
  R : *Sa voiture est devant le portail*

  (implicature conversationnelle)

- Knowledge of **contexte**

  C'est la deuxième fois qu'il reçoit un carton

  (Sport ? Courrier ? Accident ?)

- **World** knowledge

  La Nouvelle-Zélande va tailler la France en pièces.

  (métonymie + langage figuré + actualité du rugby)

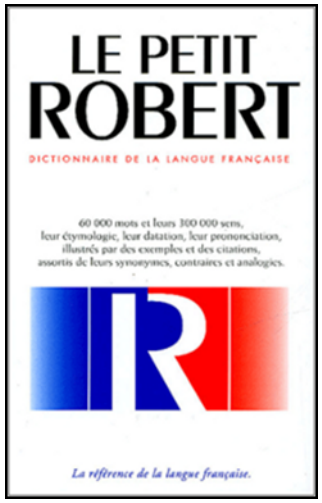
- **Deduction** (presupposition)

  Ravaillac a assassiné Henri IV en 1610.

  ⇨ Henri IV est mort en 1610.

Redundant character of language

- Lexical level
  - **Synonymy:** *vélo* and *bicyclette*
  - **Hypernymy** and **hyponymy**: *véhicule* ◁ *vélo* ◁ *VTT*
  - **Meronymy** and **holonymy**: *pédale* ◊ *pédalier* ◊ *vélo*
- Abbreviations and acronyms
  - *S'il-vous-plaît* and *SVP*, *VTT* and *Vélo Tout Terrain*
- Between mots and expressions
  - **Periphrases**: *lave-vaisselle* and *machine à laver la vaisselle*
  - **Definitions**: *selle* and *petit siège, le plus souvent de cuir, d'un cycle ou d'un véhicule à deux roues à moteur*
- Shifting of sense (contextual synonymy)
  - *Il a écrit un **papier/article** sur la recherche d'information*
  - **Vos **papiers/articles** s'il-vous-plaît !*



Redundant character of language

- **Paraphrase** (synonymy at the syntax level)
  - *Qui sera élu par le peuple en 2022?*
  - *Qui le peuple choisira-t-il dans 5 ans ?*
  - *Qui sortira vainqueur des urnes pour le prochain quinquennat ?*

- Synonymy and paraphrase are not transitive!

*Paul ressort souvent **excité** de la **récréation**.*
*Paul profite bien de la **distraction** de la récréation.*
*Entre deux cours, les **loisirs** sont bénéfiques à Paul.*
*Une période de **détente** entre deux cours ne fait pas de mal à Paul.*
*Paul se **repose** entre deux cours.*

Ambiguous character of language

Homonyms = words that have the same spelling but different senses

WIKIPÉDIA
L'encyclopédie libre

Article | Discussion | Lire | Modifier | Afficher l'hist

Noyau

Cette page d'homonymie répertorie les différents sujets et articles partageant un même nom.

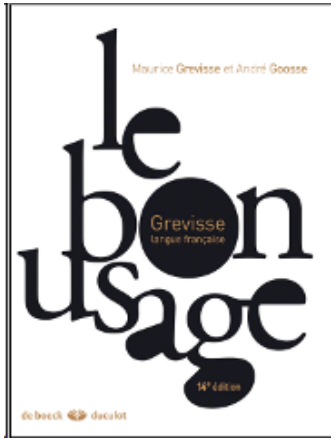
De manière générale, un **noyau** est la partie centrale située au milieu d'un autre objet. Plus particulièrement, le terme peut faire référence à :

- en biologie, un **noyau** est un organite qui contient la plupart du matériel génétique ;
- en linguistique, un **noyau** est partie fondamentale du syntagme, entourée de ses satellites ;
- en botanique, un **noyau** est la partie centrale, dure, d'une drupe ou fruit à noyau ;
- en électrotechnique, un **noyau** est la pièce magnétique sur laquelle un fil conducteur est enroulé afin de réaliser une l
- en fonderie, un **noyau** est la partie d'un moule permettant la réalisation des parties creuses d'une pièce ;
- en géologie, un **noyau** est la partie centrale approximativement sphérique de la Terre ou d'une planète ;
- en informatique, un **noyau** (aussi appelé ***kernel***) est la partie fondamentale de certains systèmes d'exploitation ;
- en mathématiques,
  - en algèbre, le **noyau** d'un morphisme de groupes est un sous-groupe particulier du groupe de départ,
  - en analyse fonctionnelle, un noyau est une fonction permettant de définir un opérateur intégral ;
- en physique, un **noyau** est la région centrale constituée des nucléons d'un atome ;
- en bande dessinée, **Noyau** est le nom de l'illustrateur Yves Nussbaum ;

Accueil
Portails thématiques
Index alphabétique
Un article au hasard
Contacter Wikipédia
Contribuer
Aide
Communauté
Modifications récentes
Accueil des nouveaux arrivants
Faire un don
Imprimer / exporter
Boîte à outils

Ambiguous character of language

- **Syntactic ambiguities**:
  - *Jean vend une tarte **aux pommes**.*
  - *Jean vend une tarte **aux clients**.*
  - *Jean a rapporté un vase **de Chine**.*



- **Anaphores**:

  *Ségolène trahit Martine. **Son** ancien mari **lui** en voulut longtemps.*

- **Ellipses**:
  - *Quelle est la taille **de la tour Eiffel** ? Le poids ?*
  - *Les Stéphanois portent des **écharpes** vertes et les*
  - *Toulousains des rouges et noires.*

Compound words

- **Compound words** are less polysemous
- Searching for them together in texts is useful (but difficult)
- Sense ≠ composition of the items
  - *Homme-grenouille*
  - *Pomme de terre*
  - *Traitement de texte*

© M. Heinrich, J. Negra

Morphology

Morphology is the study of word construction (their structures, variations and similarities)

- Morphological analysis allows to decompose a word and extract:
  - **stem**
  - **lemma**
  - **POS tag** (grammatical category)
  - **morphological features**

Construction de mots

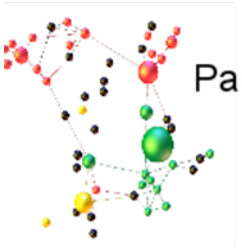
Flexion

- **Composition**

- **Derivation** (affixation)
  - **Prefixation**
  - **Suffixation**

  (to combine)

Recherche d'Information



Parcours du Web
(robots)

Facteur humain, facteur langage (rédaction)

Analyse des documents

Fonction de similarité

Facteur humain, facteur langage (évaluation)

Facteur humain, facteur langage (expression du besoin)

Analyse du besoin

Information Retrieval

Information retrieval is a statistical treatment of strings

- Enables to search large amount of information
- Applies to natural language text, does not require interventions from website creators, or particular representations of knowledge

- Machines do not understand information sense
- A search engine cannot perfom inference of information cross-checking