Introduction à la recherche d'information

Mohand Boughanem
bougha@irit.fr
http://www.irit.fr/~Mohand.Boughanem
Université Paul Sabatier de Toulouse
Laboratoire IRIT , UMR5055

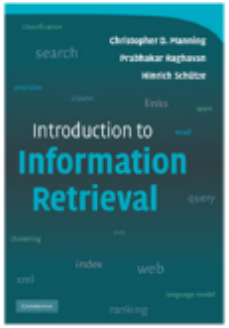
Plan

- Fondements de la Recherche d'information (RI)
  - Introduction : définition, contours de la RI
  - Problématique de la RI
  - Tour d'horizon sur les techniques de RI
- Panorama RI – scénarios et applications
  - Thématiques de recherche en RI
- Conclusion

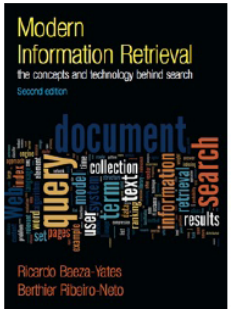
Qu'est ce que la RI ?

- Recherche d'information (RI) :
  - Ensemble des méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information pertinente pour un utilisateur

Qu'est ce que la RI ?



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



Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.

Merriam-Webster

IR: The techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system.

Qu'est ce que la RI ?

- Information retrieval (IR)
  - is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.
- IR is interdisciplinary,
  - based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics, and physics.
  - are used to reduce what has been called "information overload".
  - Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Moteurs de recherche

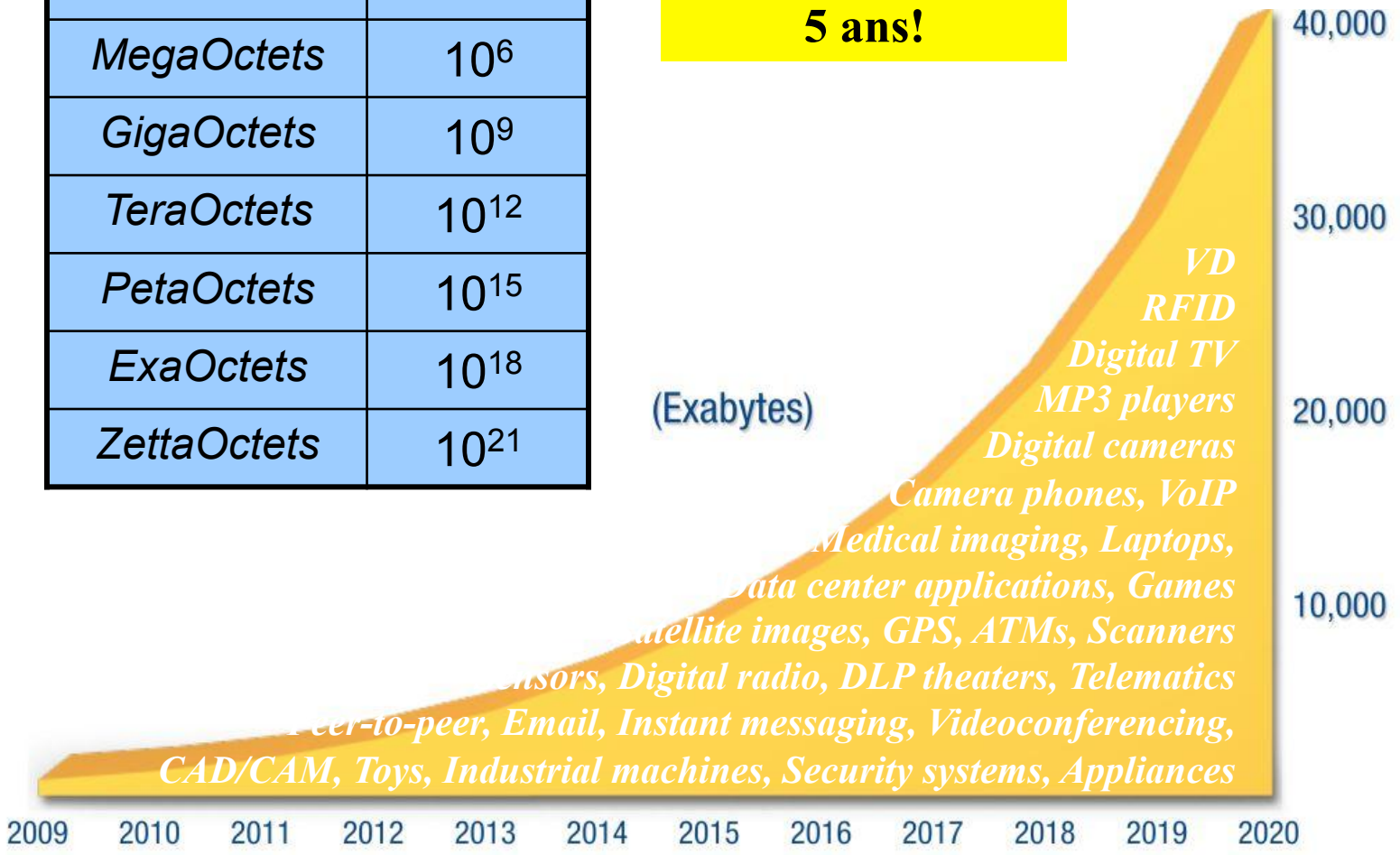
.. Mais pas seulement

- Plusieurs domaines d'application
  - Internet (Web, Forum/Blog search, news)
  - Entreprises (entreprise search)
  - Bibliothèques numériques «digital library»
  - Domaine spécialisé (médecine, droit, littérature, chimie, mathématique, brevets, software, ...)
  - Nos propres PC (Yahoo! Desktop search)

Gros volumes d'information

KiloOctets	10^3
MegaOctets	10^6
GigaOctets	10^9
TeraOctets	10^{12}
PetaOctets	10^{15}
ExaOctets	10^{18}
ZettaOctets	10^{21}

Facteur de 10 en 5 ans!

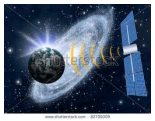


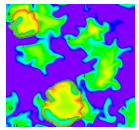
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Source: IDC, 2012

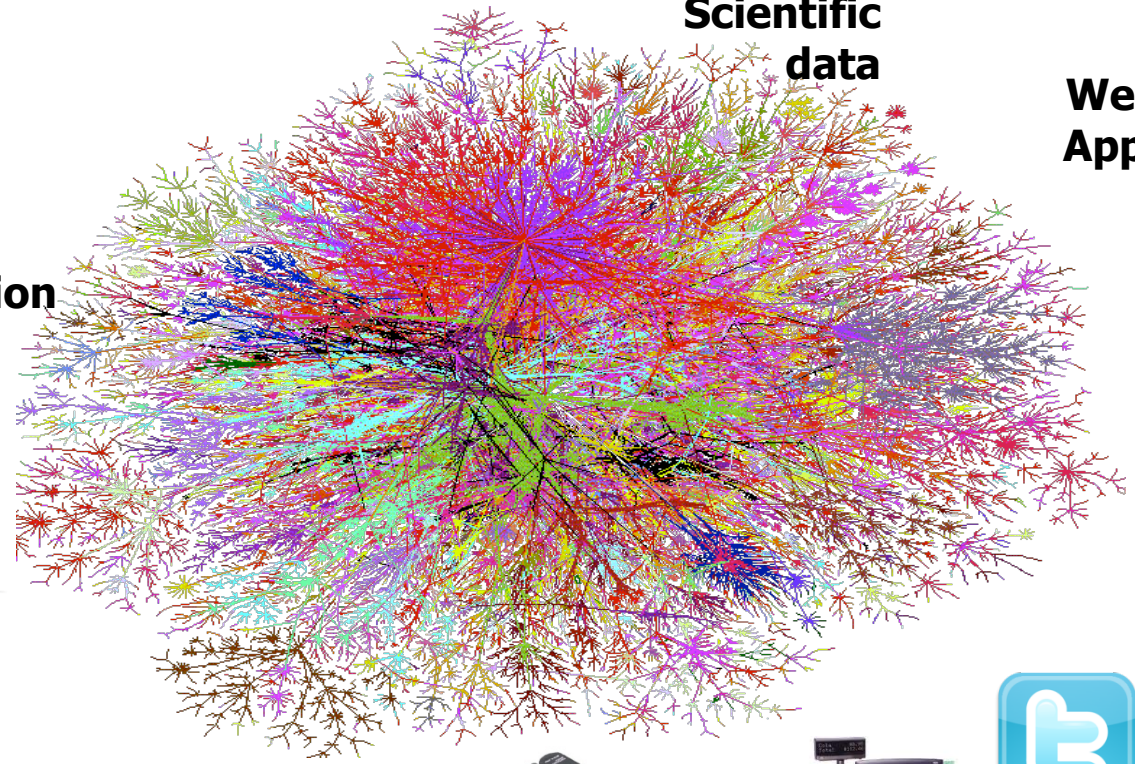
Origine

Enterprise Apps





Scientific data



Web Apps

Device explosion

Machine Data

Social Media Data

Origine

.. L'information (numérique) est disponible partout

Gros volume

- Average Number of Tweets Sent Per Day: 500 million
  - 2 billions queries per day on twitter
- Every minute 510,000 posted comments FaceBook

- 45 milliards (Google), 25 milliards (Bing)
- 672 Exabytes - 672,000,000,000 Gigabytes (GB) of accessible data.

Le problème ...

- n'est pas tant la disponibilité de l'information
- MAIS
- sa sélection, son identification → arriver à trouver au bon moment l'information utile

Le problème...

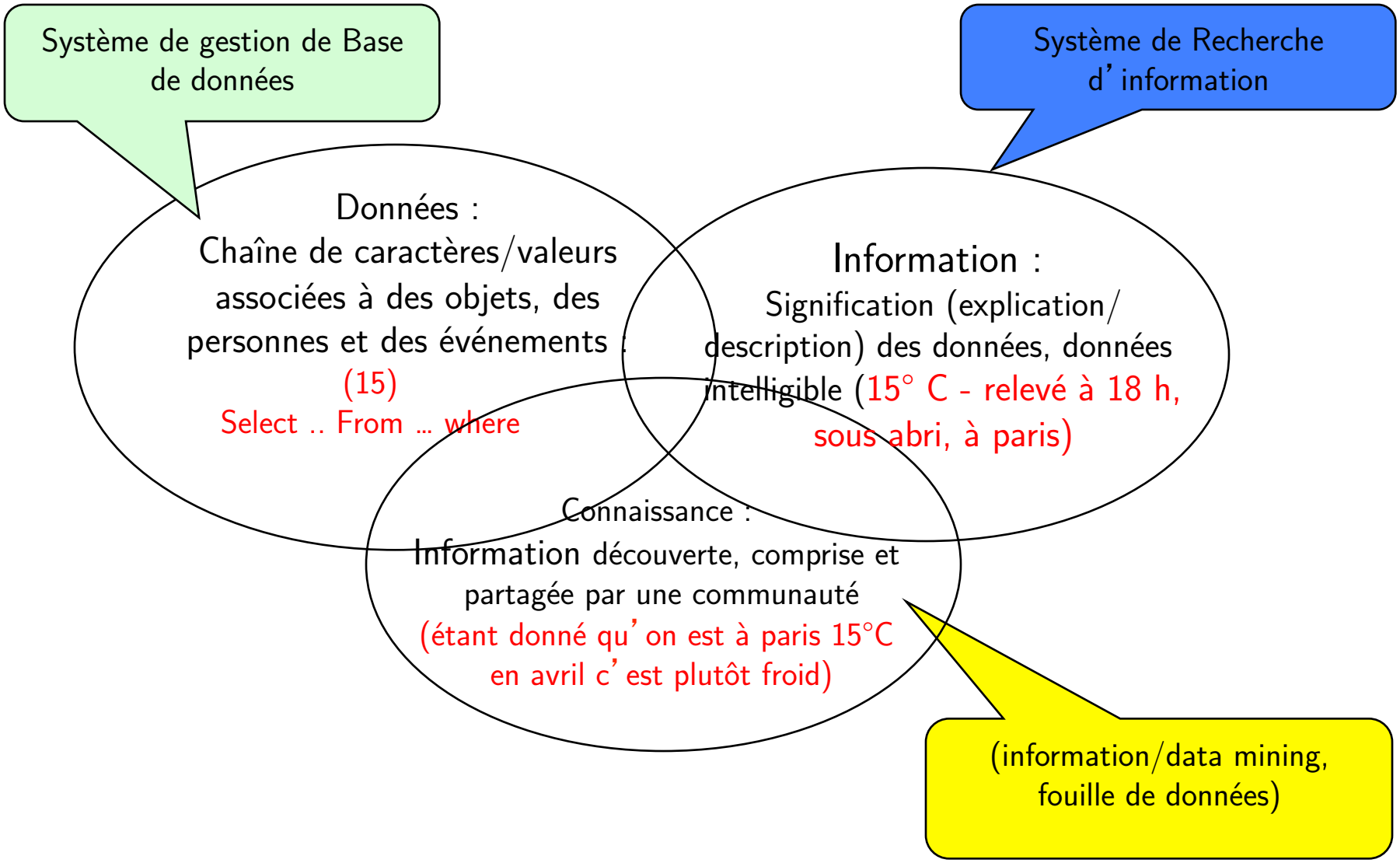
- Rechercher une information a un coût
  - « On» passe (en moyenne) 35% de son temps à rechercher des informations
  - Les managers y consacrent 17% de leur temps
  - Les 1000 grandes entreprises (US) perdent jusqu'à $2.5 milliards par an en raison de leur incapacité à récupérer les bonnes informations

- Nécessité de développer des systèmes automatisés efficaces permettant
  - Collecter, Organiser, Rechercher, Sélectionner

Contours de RI

- Contenu : donnée-information-connaissance
- Tâches : adhoc; filtrage, classification, question réponse

Donnée-info-connaissance



Tâches de RI

- Recherche adhoc
  - Je cherche des infos (pages web) sur un sujet donné
    - Je soumets une requête → retour liste de résultats
    - Requête «recherche d'info»→ SRI → renvoie une liste de documents traitant de la » recherche d'information »
  - Plusieurs types de RI adhoc
    - Recherche adhoc (tâches spécifiques)
      - Domaine spécifique (médical, légal, chimie, …)
      - Recherche d'opinions(Opinion retrieval) (sentiment analysis)
      - Recherche d'événements
      - Recherche de personnes (expert)

Tâches de RI

- Classification / Catégorisation
  - Regrouper les informations (documents) selon un ou plusieurs critères
- Question-réponses (*Query answering*)
  - Chercher des réponses à des questions
  - par exemple
    - « Qui est averroes ? »
    - « Quelle la hauteur du Mont Blanc ? »

WolframAlpha computational... knowledge engine

averroes

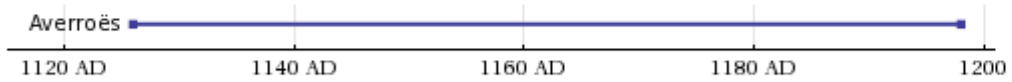
Input interpretation: *Mathematica* form

Averroës (philosopher)

Basic information:

full name	Abu al–Walid Muhammad
date of birth	1126 AD (884 years ago)
place of birth	Cordoba, Spain
date of death	1198 AD (age: 72 years) (812 years ago)
place of death	Marrakech, Marrakech–Tensift–Al Haouz, Morocco

Timeline:



Computed by: Wolfram *Mathematica* Source information » Download as: PDF | Live *Mathematica*

Tâches de la RI

- Filtrage d'information/ recommandation (filtering/ recommendation)
  - Recommandation
  - Dissémination sélective d'information
  - Système d'alerte
  - Dissémination sélective d'information
  - Push
  - Profilage (profiling)

Tâches de la RI

- Résumé automatique (document summarization)

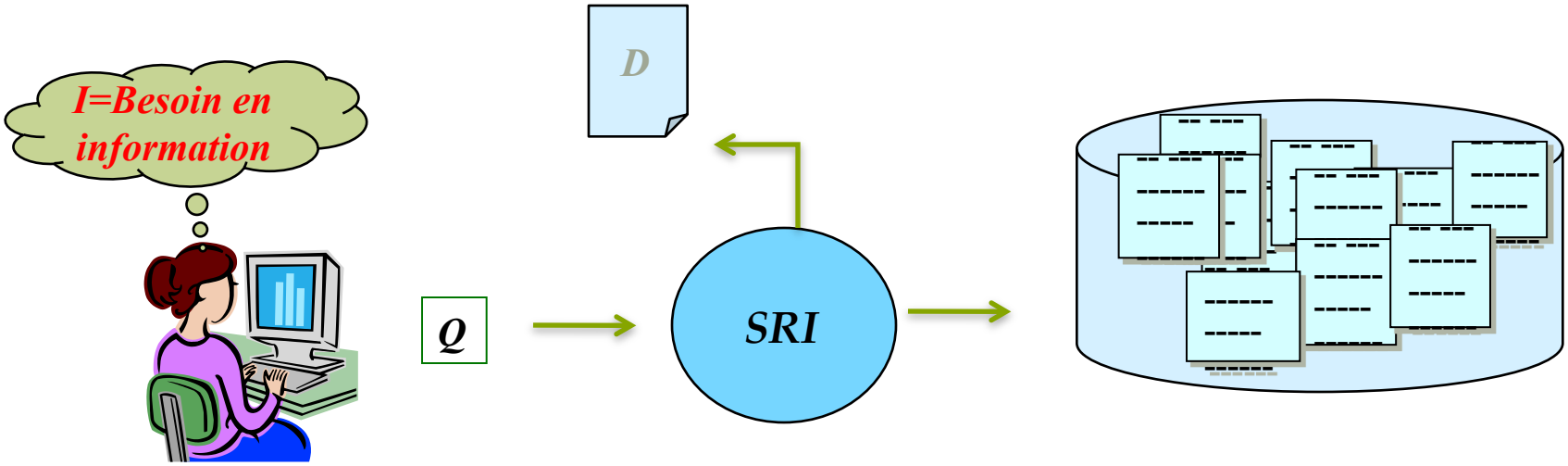
- Recherche agrégée (Aggregated search)
  - Agréger des moteurs : interroger les résultats de plusieurs moteurs (méta-moteurs)
  - Agréger des résultats : interroger plusieurs sources (vertical search)
  - Agréger des contenus : former un résultat à partir de plusieurs contenus

- Vertical search

Plan

- Introduction
- Chapitre 1 : Concepts de base de la RI
- Chapitre 2: Langage de requêtes
- Chapitre 3 : Indexation et pondération
- Chapitres 4 : Modèles de RI
- Chapitre : Evaluation de RI
- Chapitre 5: RI sur le WEB
- Chapitre : XML et RI
- Chapitre : Techniques de reformulation

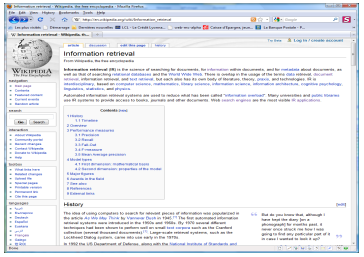
Problématique de la RI



- Sélectionner dans une collection
  - les informations (items, documents, ..)
  - ... pertinentes répondant aux
  - ... besoins en information des utilisateurs

Information

- Formes
  - Texte, images, sons, vidéo, graphiques, etc.
  - Exemples texte : web pages, email, livres, journaux, publications, blog, Word™, Powerpoint™, PDF, forum postings, brevets, etc.
- Hétérogénéité
  - langage (multilingues)
  - media (multimédia)

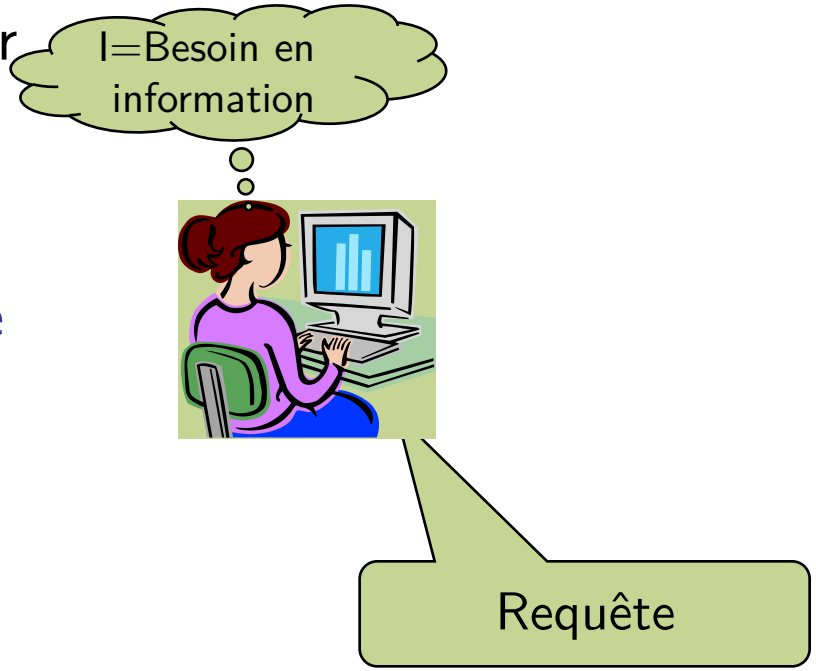


Question

- Comprendre le contenu vs. l'interptérer →Ambiguïté du langage naturel (polysémie, synonymie, ...)
- Information, document, unité/granule/passage

Besoin en information (Information needs)

- Besoin en information est une expression mentale d'un utilisateur
- Requête
  - Ensemble de mots-clés
  - → Une représentation possible du besoin en information



Question

- Comment capturer le besoin de l'utilisateur

Intention de la requête
"*Query intent*"

Pertinence

- Au cœur de tout système de RI
  - Relation entre le document et ... la requête ou le besoin de l'utilisateur ?

- Plusieurs facteurs influencent la décision de l'utilisateur, tâche, le contexte, nouveauté, style, compréhension, temps, ...

- Pertinence par document

> Goffman, 1969: '...the relevance of the information from one document depends upon what is already known about the subject, and in turn affects the relevance of other documents subsequently examined.'

Pertinence

Type of relevance(survey) (Saracevic 2007)

- Plusieurs pertinences
  - Thématique (topical): relation entre le sujet exprimé dans la requête et le sujet couvert dans le document.
  - Contextuelle (Situation) : relation entre la tâche, le problème posé par l'utilisateur, la situation de l'utilisateur et l'information retrouvée.
  - Cognitive : relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée

Question

- Processus subjectif (humain), dépend de plusieurs facteurs
  → difficile à automatiser

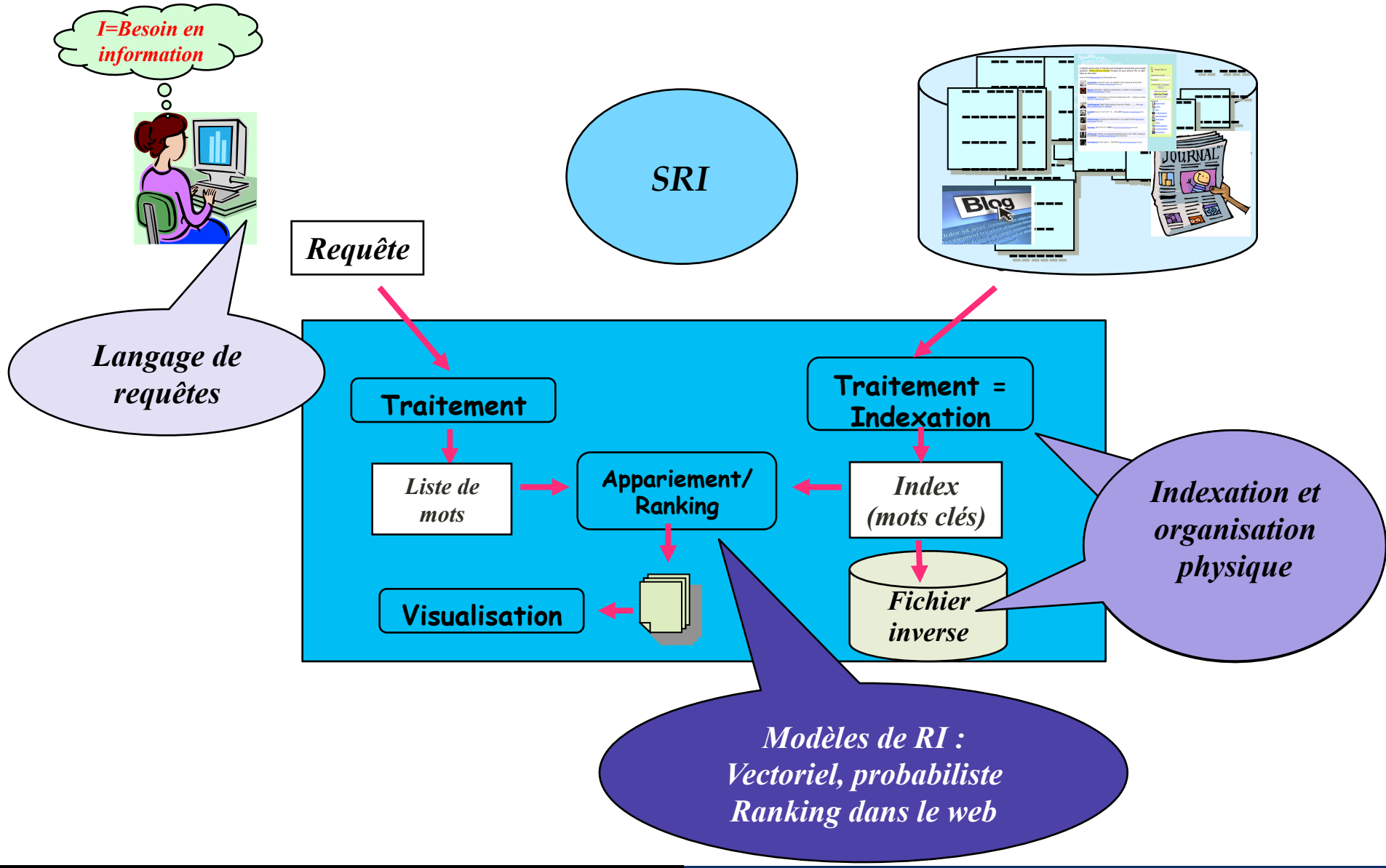
Hypothèses simplificatrices

- Besoin = requête
  - Besoin confondu avec la requête utilisateur (une liste de mots clés)
- Document et requête
  - Représentés par des termes (mots simples, groupes de mots, ...) → Sac de mots
- Pertinence
  - Traduite par la similarité de vocabulaire (mots) entre la requête et le document→ thématique

Démarche RI

- Interpréter le texte au lieu de le comprendre
- Exploiter les propriétés statistiques (comptage de mots) du texte plutôt que ses propriétés linguistiques

Processus de RI



Fichier inverse

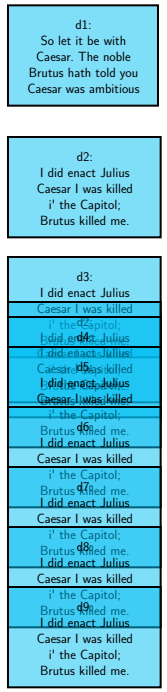
d1:
So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

d2:
I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Traitement
=
Indexation

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1



Matching/appariement : modèles de RI

Caesar, brutus

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

d1:
So let it be with
Caesar. The noble
Brutus hath told you
Caesar was ambitious

d2:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

d3:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

- Facteurs utilisés par la majorité des modèles
  - Fréquence du terme dans le doument (tf), sa fréquence dans la collection (idf), sa position dans le texte(p), taille du document (dl) ...

$$Score(D) = fonction(tf, idf, dl)$$

  - Plusieurs modèles théoriques pour formaliser cette fonction
  - Elle peut être apprise (apprentissage automatique, approche utilisée par la majorité des moteurs de recherche)

- Théorie des ensembles :
  - Boolean model (±1950)
- Algèbre
  - Vector space model (±1970)
  - Spreading activation model (±1989)
  - LSI (Latent semantic Indexing)(± 1994)
- Probabilité
  - Probabilistic model (±1976)
  - Inference network model (±1992)
  - Language model (±1998)
  - DFR (Divergence from Randomness model) (±2002)

Références bibliographiques

- Ouvrages en ligne
  - Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. 2008 http://nlp.stanford.edu/IR-book/information-retrieval.html)
  - Baeza-Yates, R. and Ribeiro-Neto, B. (2011). Modern Information Retrieval - the concepts and technology behind search.
  - Ricardo Baeza-Yates and Berthier Ribeiro-Neto.Modern Information Retrieval. Addison-Wesley, 1999
  - Van Rijsbergen (1977) Information Retrieval, Butterworths
  - Frakes and Baeza-Yates, eds. (1992) Information Retrieval: Data Structures & Algorithms, Prentice Hall
  - Witten, Moffat and Bell (1994) Managing Gigabytes plus software, Van Nostrand-Reinhold
  - Baeza-Yates and Ribeiro-Neto, eds. (1999) Modern Information Retrieval Addison-Wesley (site miroir)
  - Recherche d'information : état des lieux et perspectives (M. Boughanem et J. Savoy)

Conférences et Journaux

- Conférences
  - ACM SIGIR : Special interest group on Information Retrieval
  - CIKM : Conference on Information and Knowledge Management
  - ECIR : European Conference on Information Retrieval Research, University of Sunderland, U.K.
  - WSDM: Internationl conference on Web Search and Data Mining
  - RIAO (OAIR): Coupling approaches, coupling media and coupling languages for information retrieval
  - CORIA : Conférence Francophone en Recherche d'Information et Applications
- Journaux
  - JASIST : Journal of the American Society for Information Science and Technology
  - IP&M : Information Processing & Management
  - IJODL : International Journal on Digital Libraries
  - JDOC : Journal of Documentation
  - JIR : Journal of Information Retrieval
  - ACM-TOIS : Transactions on Information Systems

Ref. bibliographiques

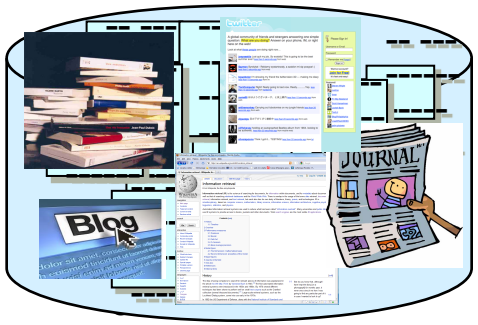
- Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne

- Recommended reading for IR research students A.Moffat J.Zobel D. Hawking (2005)

- http://sigir.org/resources/

Organisation de la suite du cours

SRI



Besoin

Traitement

Liste de mots

Appariement/ Ranking

Visualisation

Traitement = Indexation

Index (mots clés)

Fichier inverse

Chapitre 2: Langage de requêtes

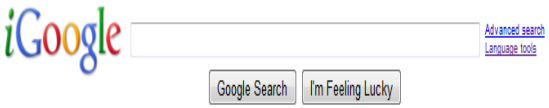
Chapitre 3 : Indexation et organisation physique

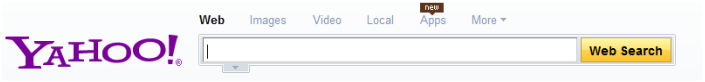
Modèles de RI : Vectoriel, probabiliste Ranking dans le web

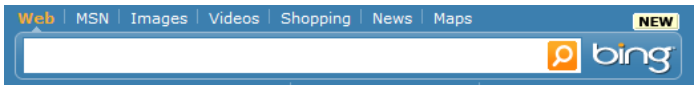
Chapitre 2 : Langage d'interrogation

Du besoin en information à la requête

- Besoin peut être
  - Ponctuel(adhoc)/ Récurrent (filtrage, recommandation)
- Expression des besoins (Langage de requêtes)
  - Texte libre, Liste de mots clés
  - Avec / sans opérateurs booléens (AND, OR, NOT)
  - Images (…)
  - Aucun : navigation dans une liste de concepts (Yahoo,…)
- Requête est le résultat
  - de l'expression des besoins ?
  - ou du processus de représentation des besoins ?
- Ces deux notions sont souvent confondues



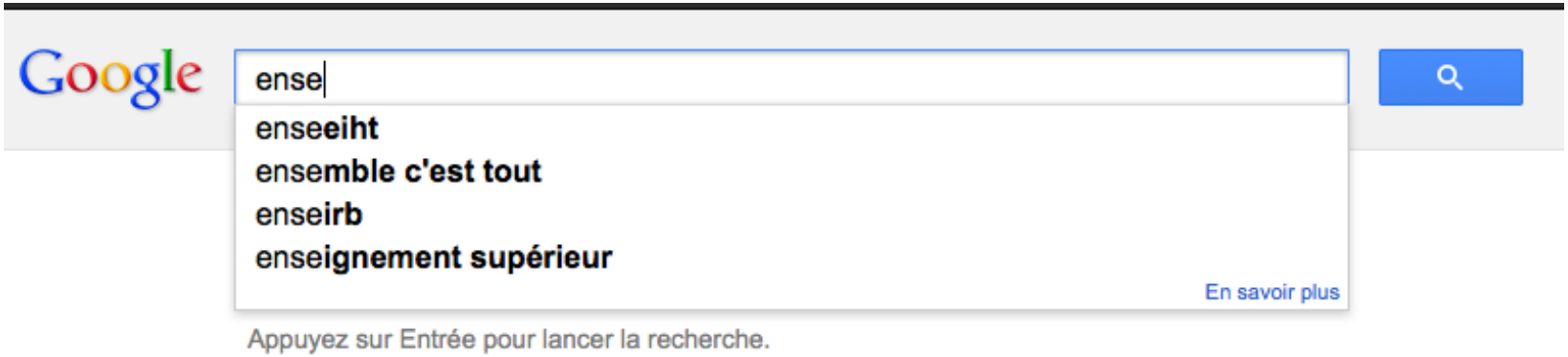




Du besoin en information à la requête

- Paradoxe de la RI
  - Une requête «idéale» doit comporter toutes les informations que l'utilisateur recherche → la similarité serait maximale
  - Or, l'utilisateur recherche une information qu'il ne connaît pas à priori, il ne peut donc pas l'exprimer (décrire) de manière précise (idéale)
  - Ce phénomène est qualifié par Belkin ASK "Anomalous State of Knowledge"

Suggestion de requêtes



Recherches associées à **enseeiht**

enseeiht **classement**	**bde** enseeiht
moodle enseeiht	**rcmail** enseeiht
enseeiht **stage**	**intranet** enseeiht
adresse enseeiht	enseeiht **classement 2011**

Gooooooooooogle ›

1 2 3 4 5 6 7 8 9 10 **Suivant**