Techniques d'indexation et recherche multimédia

Chap1-Introduction

Amira BELHEDI amira.belhedi@istic.ucar.tn
Fadoua MHIRI fadoua.mhiri@istic.ucar.tn
Wafa ABID wafa.abid@istic.ucar.tn

ISTIC 2021-2022

Objectif du cours

- Comprendre:
  - Les techniques d'indexation des documents multimédia (texte, image, vidéo, son)
  - Les modèles de Recherche Multimédia
  - Les méthodes d'évaluation des performances d'un système de recherche multimédia
  - Fonctionnement des moteurs de recherche

- Le Système de recherche devra retourner le moins possible de résultats non pertinents.

Plan du cours

- **Chapitre 1: Introduction**
  - Problématique et difficulté de la Recherche Multimedia (**RM**)
  - Système de Recherche Multimédia (**SRM**)
    - définition, concepts de base et architecture
  - Un peu d'histoire
- **Chapitre 2: Indexation des documents texte**
  - Modes d'indexation
    - manuelle, automatique et semi-automatique
  - Etapes de l'indexation textuelle
    - Extraction, étalage, normalisation et pondération
  - Construction du fichier indexe (fichier inversé)
  - Indexation textuelle sémantique
- **Chapitre 3: Indexation des images**
  - Indexation textuelle des images
  - Indexation des images par le contenu
    - Types de requêtes visuelles
    - Descripteurs visuels
      - Descripteurs globaux: Histogramme, texture
      - Descripteurs locaux: régions de l'image, points d'intérêt

Plan du cours

- **Chapitre 4**: **Modèles de recherche Multimédia**
  - Modèles classiques de recherche textuelle
    - modèle booléen
    - modèle vectoriel
    - Modèle probabiliste
  - Recherche d'images par le contenu visuel
  - Systèmes de recherche d'images combinant texte et images
- **Chapitre 5: Evaluation des performances des SRM**
  - Notions de rappel et précision
  - Mesures de similarité utilisées dans la recherche d'images par le contenu
  - Méthodes de comparaison
- **Chapitre 6: Recherche d'information sur le web**
  - Principes de fonctionnement d'un moteur de recherche
  - Crawler: robot d'indexation
  - Page RANK: système de classement des pages web

Introduction

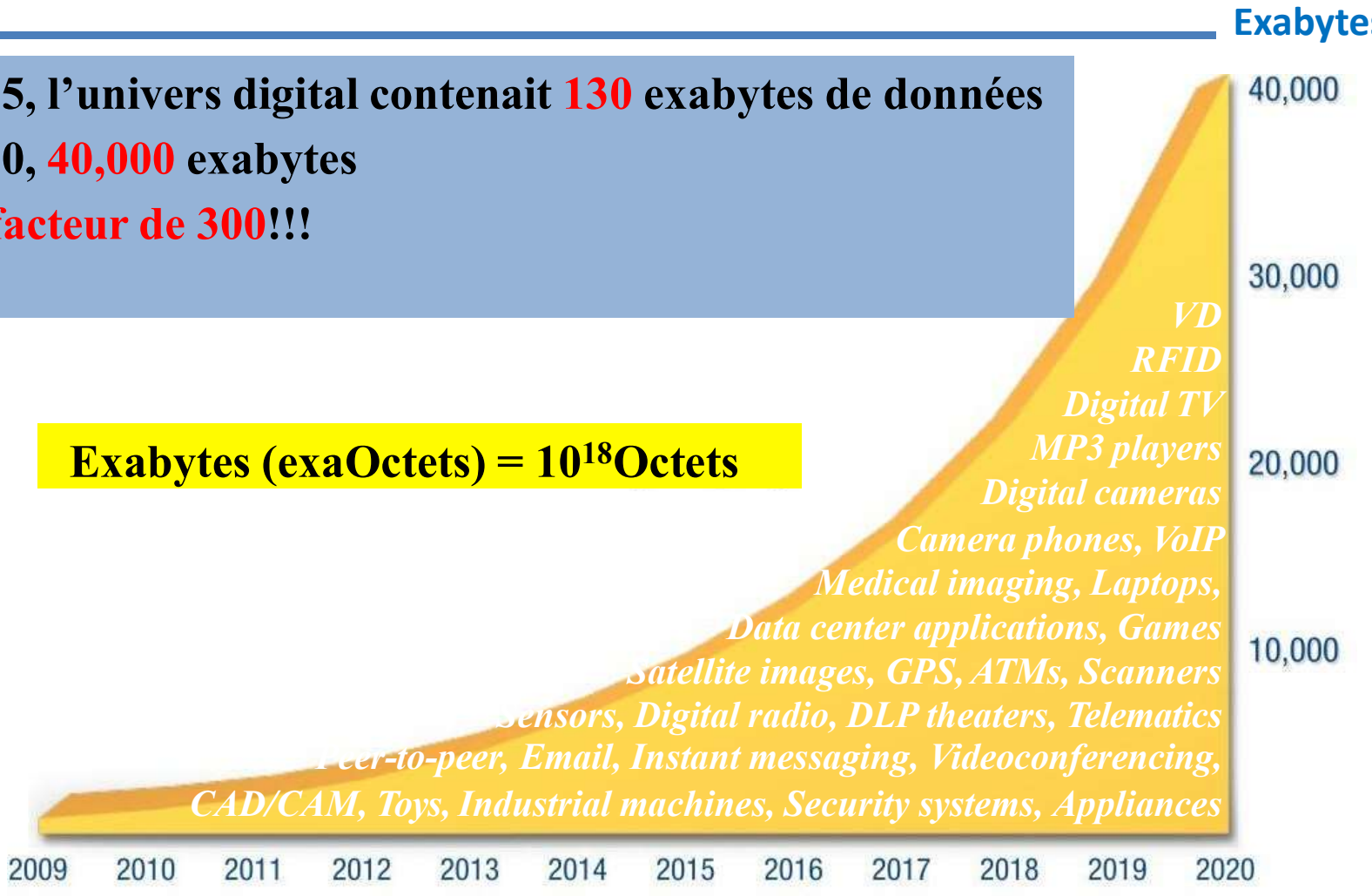
- La recherche d'information (RI) dans les documents multimédias est un problème crucial

Abondance des archives multimédia

En 2005, l'univers digital contenait 130 exabytes de données

En 2020, 40,000 exabytes

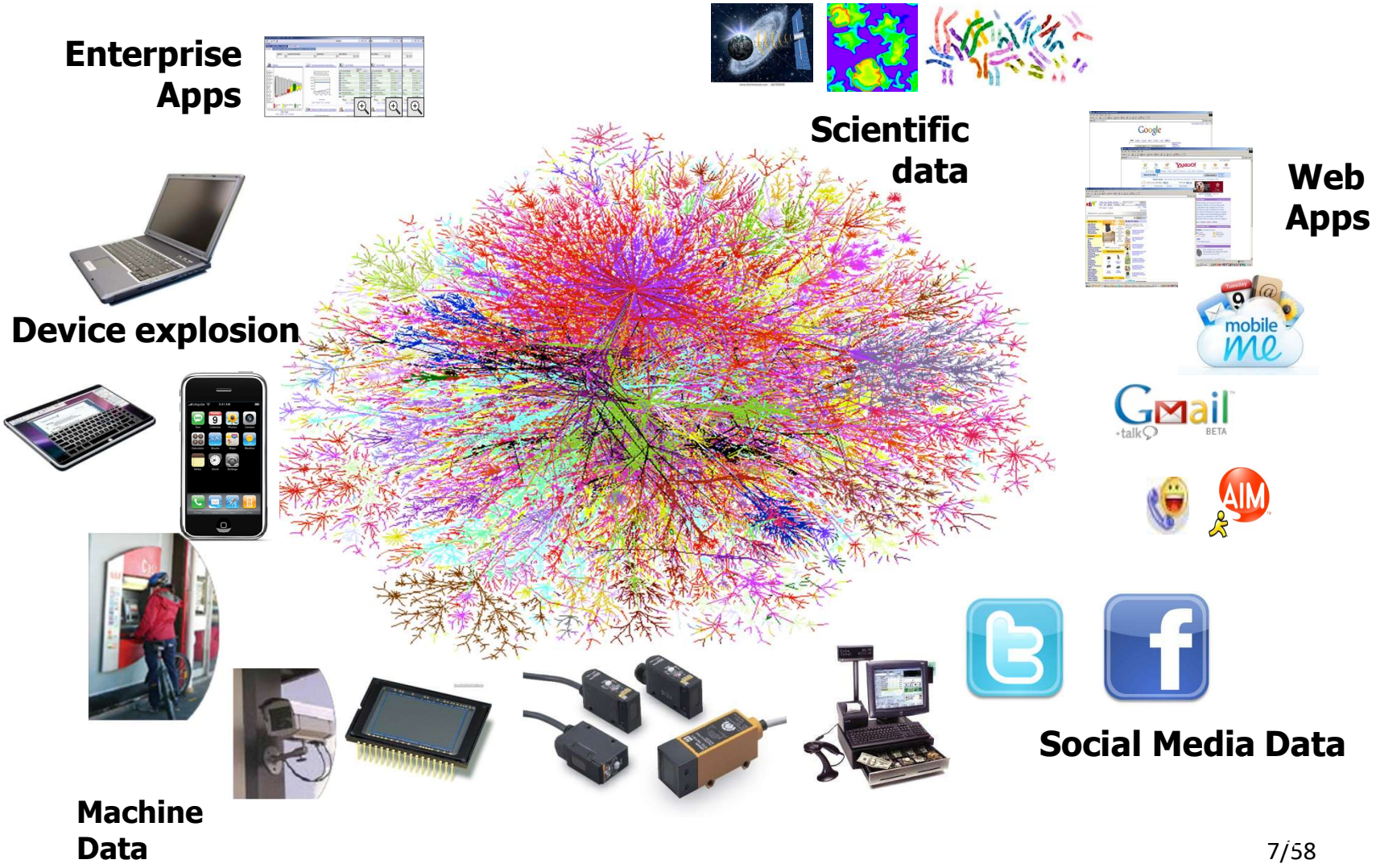
➔ un facteur de 300!!!



Exabytes (exaOctets) = 10^{18}Octets

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

L'information est partout: origine



L'information est partout: origine

Les documents multimédia sont disponibles partout

L'information est partout: gros volume

- Nombre moyen de Tweets envoyés par jour:500 millions
  - 2 billions (2000 millions) de requêtes twitter par jour
- Chaque minute, 510,000 commentaires FaceBook postés
- 45 milliards (Google), 25 milliards (Bing)
- 672 Exabytes - 672,000,000,000 Gigabytes (GB) de données accessibles

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

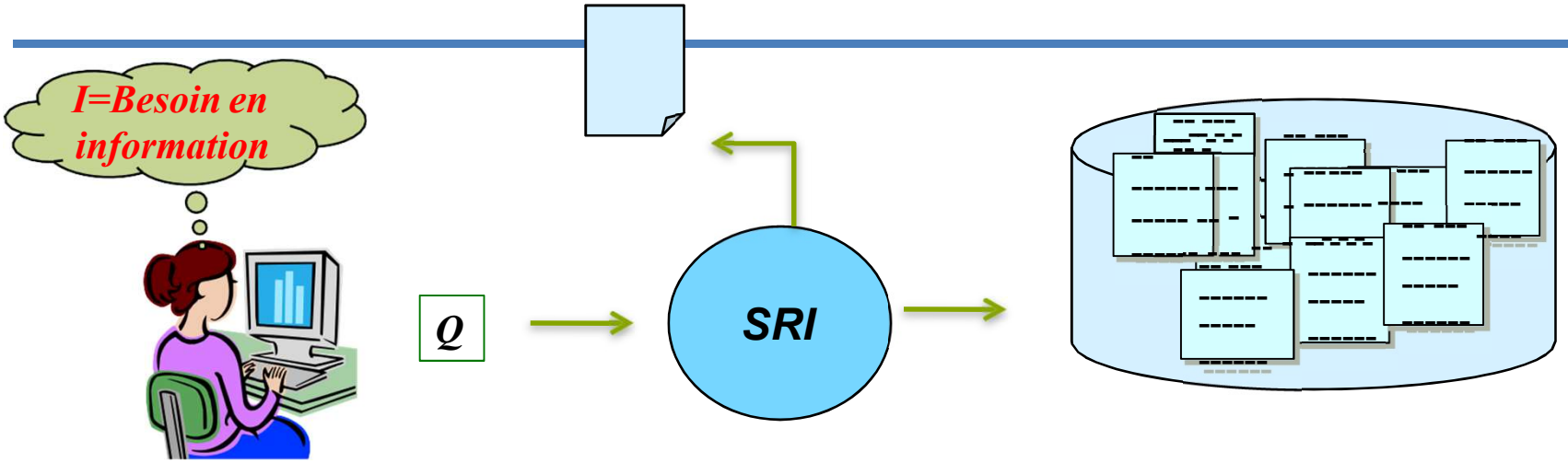
- Nécessité de développer des systèmes automatisés efficaces permettant de
  - Collecter, Organiser, Rechercher l'information pertinente

Exemple de SRI: les moteurs de recherche

Autres exemples de SRI

- 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)

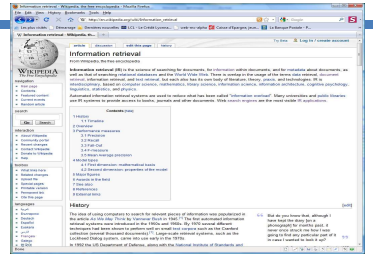
Problématique



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

Problématique

- Hétérogénéité
  - Formes: Texte, images, sons, vidéo, graphiques, etc.
    - Exemples texte : web pages, email, livres, journaux, publications, blog, Word™, powerpoint, PDF, forums, brevets, etc.
  - Langues: Français, Angalais, Espagnel,etc.



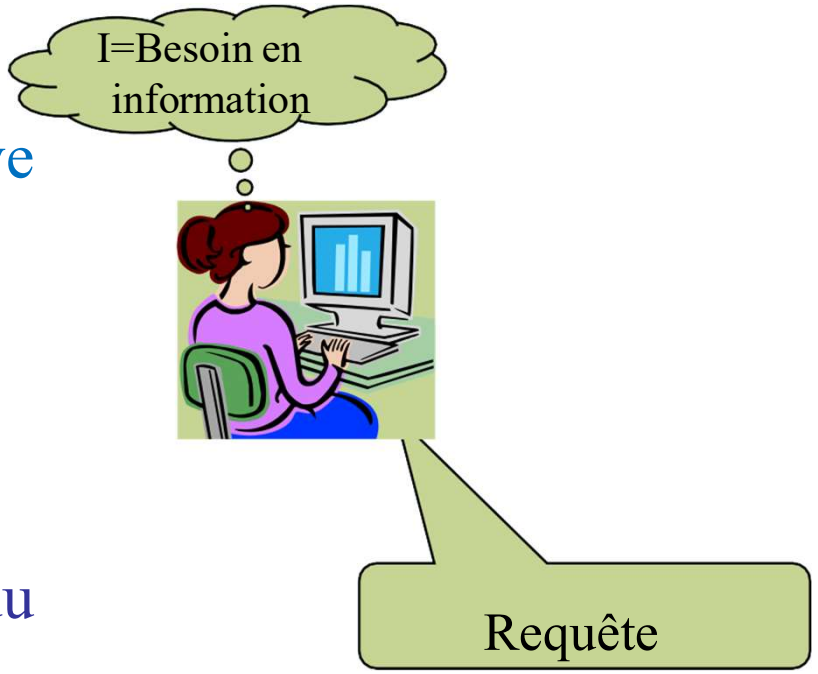
Question

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

Problématique

- L'expression du besoin d'information par l'utilisateur est parfois vague et toujours subjective
  ➔ L'humain est subjectif et il utilise un langage "naturel" !
- Requête
  - Ensemble de mots-clés

  ➔ Une représentation possible du besoin en information



Question

- Comment capturer le besoin de l'utilisateur

Langage de la requête

- Le langage "naturel" est implicite, redondant et ambigu

1. Implicite : tout n'est pas dit dans les textes
   - Donner des exemples ?
2. Redondant: nombreuses façons de formuler le même contenu
   - Donner des exemples ?
3. Ambigu : un même énoncé peut souvent être interprété de différentes façons
   - Donner des exemples ?

Langage de la requête

- Le langage "naturel" est :

1. Implicite : tout n'est pas dit dans les textes
   - Question : Le voisin est-il chez lui ?
   - Réponse : Sa voiture est devant le portail ➔ implique que: Le voisin est chez lui
   - Il a assassiné Henri IV en 1610 ➔ Henri IV est mort en 1610.
2. Redondant: nombreuses façons de formuler le même contenu
   - vélo / bicyclette
   - véhicule / vélo / VTT
   - pédale / pédalier / vélo
   - lave-vaisselle / machine à laver la vaisselle
3. Ambigu : un même énoncé peut souvent être interprété de différentes façons
   - Il vend une tarte aux pommes
   - Il vend une tarte aux clients.

Intention de la requête ?

- Exemple: mot clé recherche « apple »

Problématique

- La pertinence d'un document pour une requête est une notion variable et très complexe à définir.
  - pas de SRI parfait
  - Méthodes d'évaluation d'un SRI différentes de celles utilisées dans l 'évaluation des systèmes informatiques

Problématique

- Comment retrouver une information qui intéresse un utilisateur ?

- Besoin: organiser, accéder et retrouver des informations qui satisfont un besoin utilisateur en terme d'information

Problématique

- A votre avis, pourquoi on n'utilise un SGBD?
- Quels sont les limites des SGBD?

Problématique
Limite des SGBD pour la RI

- Structure des informations
  - SGBD traite des informations structurées
    - exigence d'un schéma de la base
  - RI recherche d'information non structurée

- Appariement des données
  - SGBD appariement exacte: le mapping entre les valeurs des attributs d'une requête et celle de la BD sont exactes (=, like,etc.)
  - RI appariement approximatif: liste des documents triés selon leurs pertinences

Problématique
Limite des SGBD pour la RI

- Langage
  - SGBD langage de requête dédié à des spécialistes
    - Exemple: langage SQL
  - RI langage libre (« langage naturel »): mots clés

- RI prise en compte des variations
  - Morphologiques: étudiant/étudiante/étudiants/étudiantes
    prétraitement/traitement/traiter
  - Syntaxiques: m'entends-tu? / Tu m'entends?
  - Lexicales: auto, voiture, char, automobile

Définition

- La recherche d'information ou RI (ou encore SRI) prend plusieurs terminologies:
  - recherche d'information,
  - informatique documentaire,
  - information retrieval,
  - document retrieval.

Définition

- La RI dans les documents multimédias est une discipline qui s'intéresse à la proposition de méthodes et de techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection pertinente des documents multimédia pour un utilisateur

Définition

- Un SRI est un ensemble logiciel qui permet de retrouver une information pertinente par rapport à une requête dans une grande collection de documents

Tâches de la RI

- La RI est un domaine vaste qui se situe dans plusieurs disciplines:
  1. Recherche adhoc
  2. Classification /catégorisation (clustering),
  3. Question-réponse (query-answering)
  4. Filtrage d'information (filtering, recommandation)
  5. Méta-moteur (data-fusion, meta-search)
  6. Résumé automatique (summurization)
  7. Croisement de langues (cross langage)
  8. Fouille de textes (texte mining)

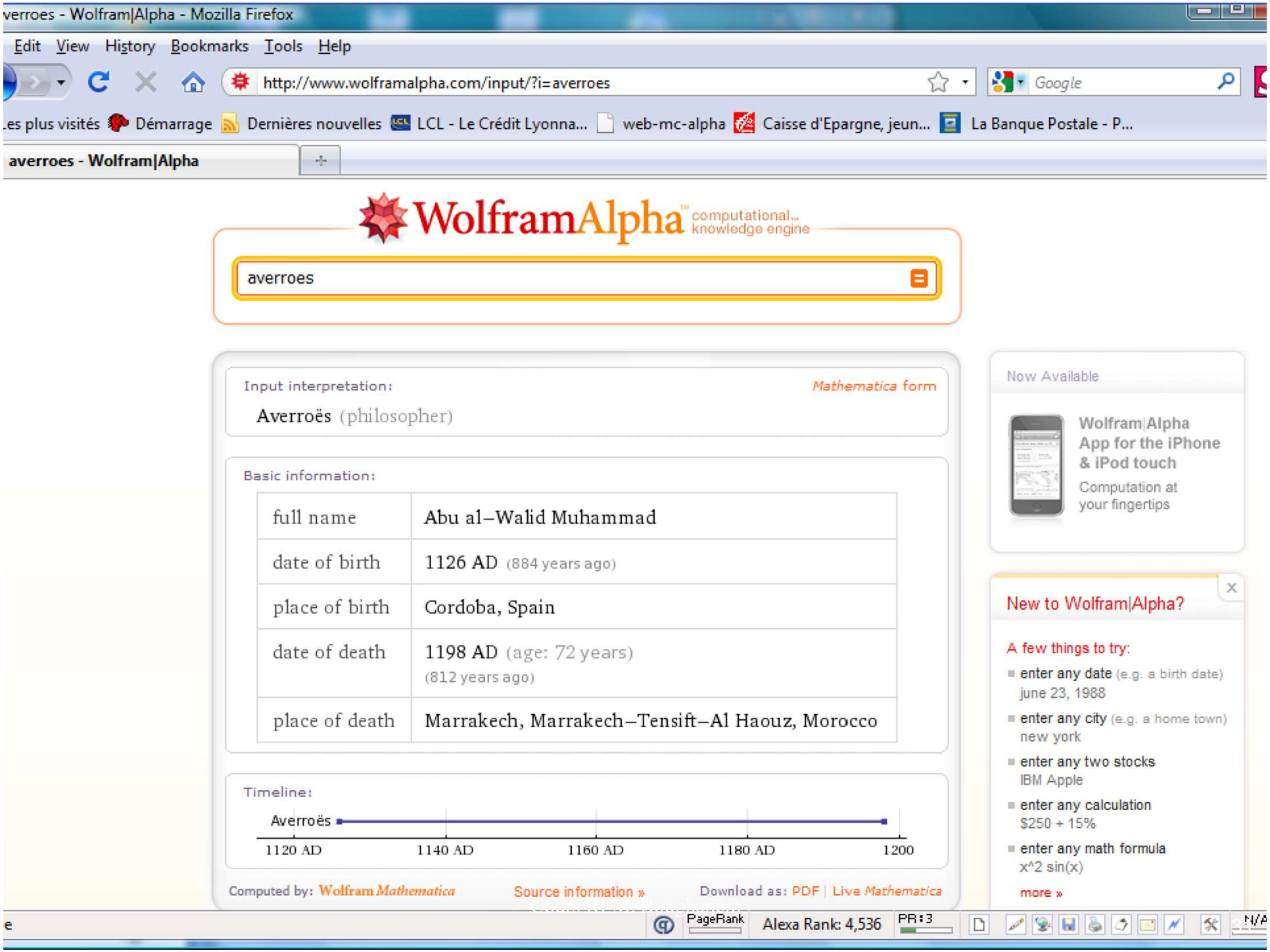
Tâches de la RI

1. Recherche adhoc: recherche dans une collection de documents fixée
   - 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 la RI

2. Classification / Catégorisation
   - Regrouper les informations (documents) selon un ou plusieurs critères

3. Question-réponses (*Query answering*)
   - Chercher des réponses à des questions
   - par exemple: « Quelle est la hauteur du Mont Blanc ? »
   - Exemple d'outil: WolframAlpha = service internet qui répond directement à la saisie de questions par le calcul de la réponse à partir d'une collection de données, au lieu de procurer une liste de documents ou de pages web pouvant contenir la réponse


verroes - Wolfram|Alpha - Mozilla Firefox
Edit View History Bookmarks Tools Help
http://www.wolframalpha.com/input/?i=averroes
Google
Les plus visités Démarrage Dernières nouvelles LCL - Le Crédit Lyonna... web-mc-alpha Caisse d'Epargne, jeun... La Banque Postale - P...
averroes - Wolfram|Alpha
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:
Averroës
1120 AD
1140 AD
1160 AD
1180 AD
1200
Computed by: Wolfram Mathematica
Source information »
Download as: PDF | Live Mathematica
Now Available
Wolfram|Alpha App for the iPhone & iPod touch
Computation at your fingertips
New to Wolfram|Alpha?
A few things to try:
enter any date (e.g. a birth date)
june 23, 1988
enter any city (e.g. a home town)
new york
enter any two stocks
IBM Apple
enter any calculation
$250 + 15%
enter any math formula
x^2 sin(x)
more »
PageRank
Alexa Rank: 4,536
PR: 3

Tâches de la RI

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

Tâches de la RI

5. Résumé automatique (document summarization)

6. 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

Tâches de la RI

- Vertical search

Système de recherche d'information SRI

Les différents acteurs de la RI

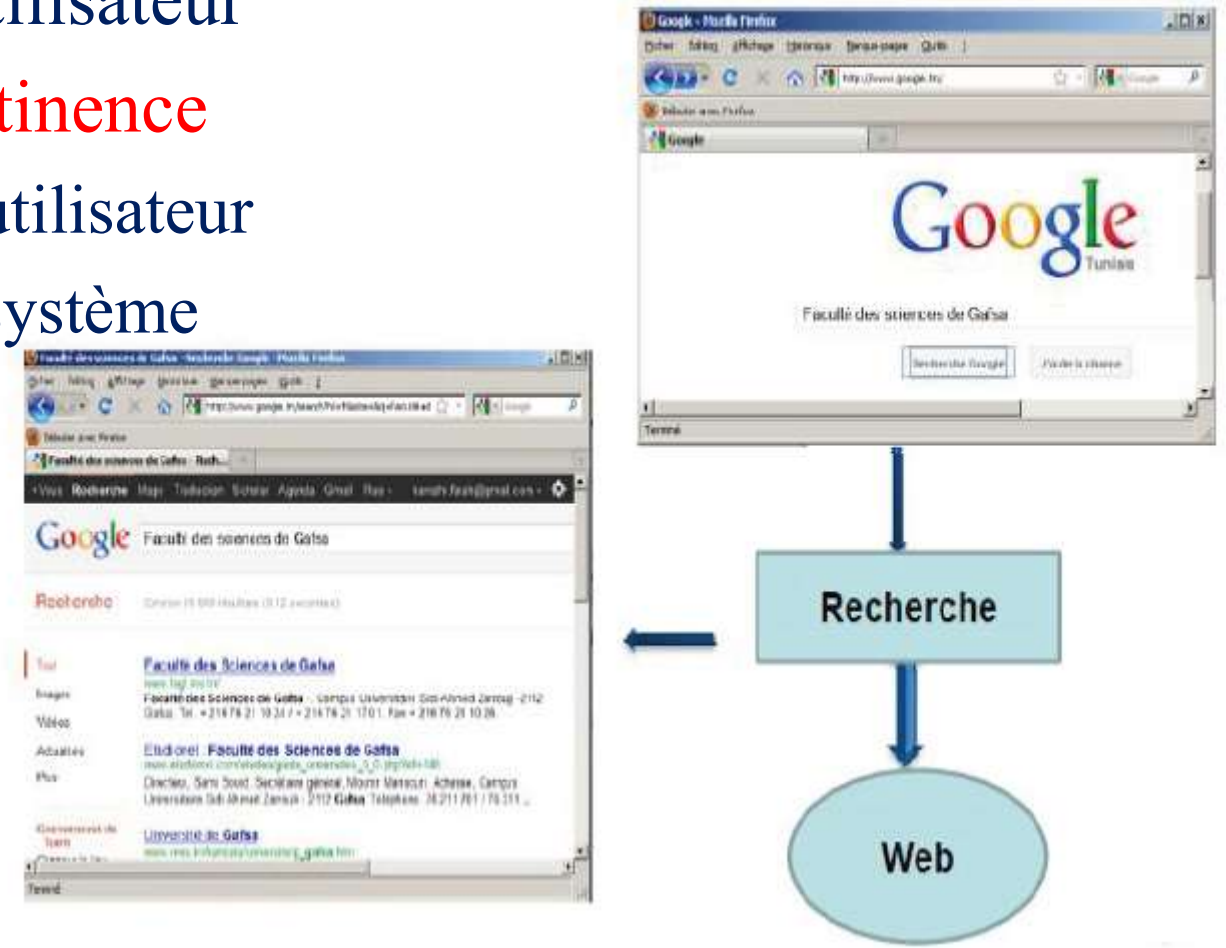
Collection :
un ensemble de documents
+
Un ensemble d'images, vidéos et fichiers sons

Utilisateur :
un besoin d'information et/ou une tâche à accomplir

Système de RI : l'outil qui doit retrouver les documents pertinents pour le besoin de l'utilisateur

SRI

- Résultats de la recherche d'information relatives à une requête utilisateur
- Notion de pertinence
  - Pertinence utilisateur
  - Pertinence système



Pertinence

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

- Pertinence: « degré de corrélation entre la requête et le document apporté », la pertinence est un concept-clé de la RI
  - Pertinence utilisateur
  - Pertinence système

Pertinence

- Pertinence utilisateur (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

Type of relevance(survey) (Saracevic 2007)

Question

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

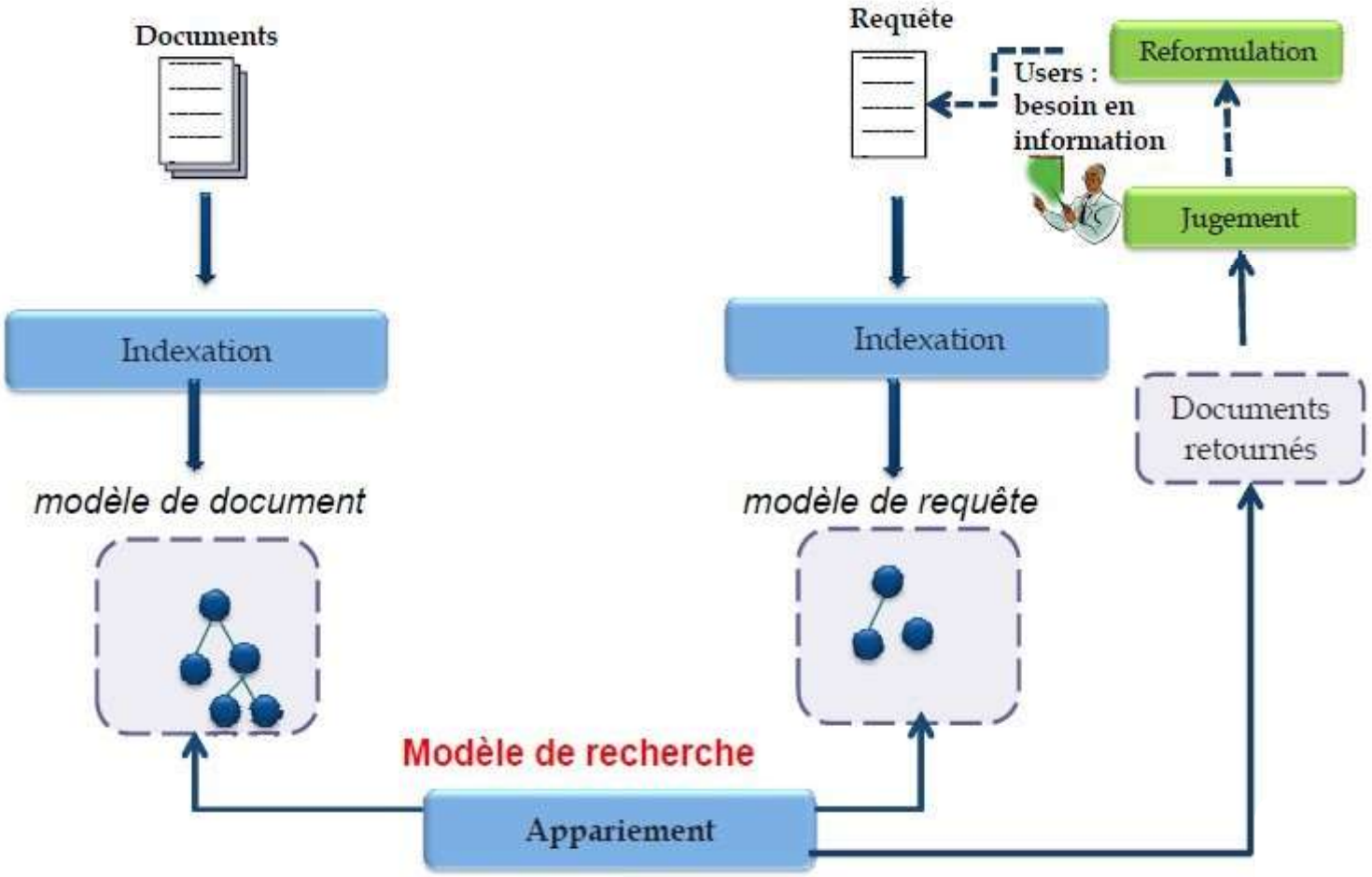
Pertinence

- Pertinence système: c'est la pertinence calculée par le système en comparant la représentation des documents et celle des requêtes

- L'enjeu de la RI est de rapprocher la pertinence système de la pertinence utilisateur

- C'est sur cette notion que les SRI sont jugés

Architecture d'un SRI



Indexation

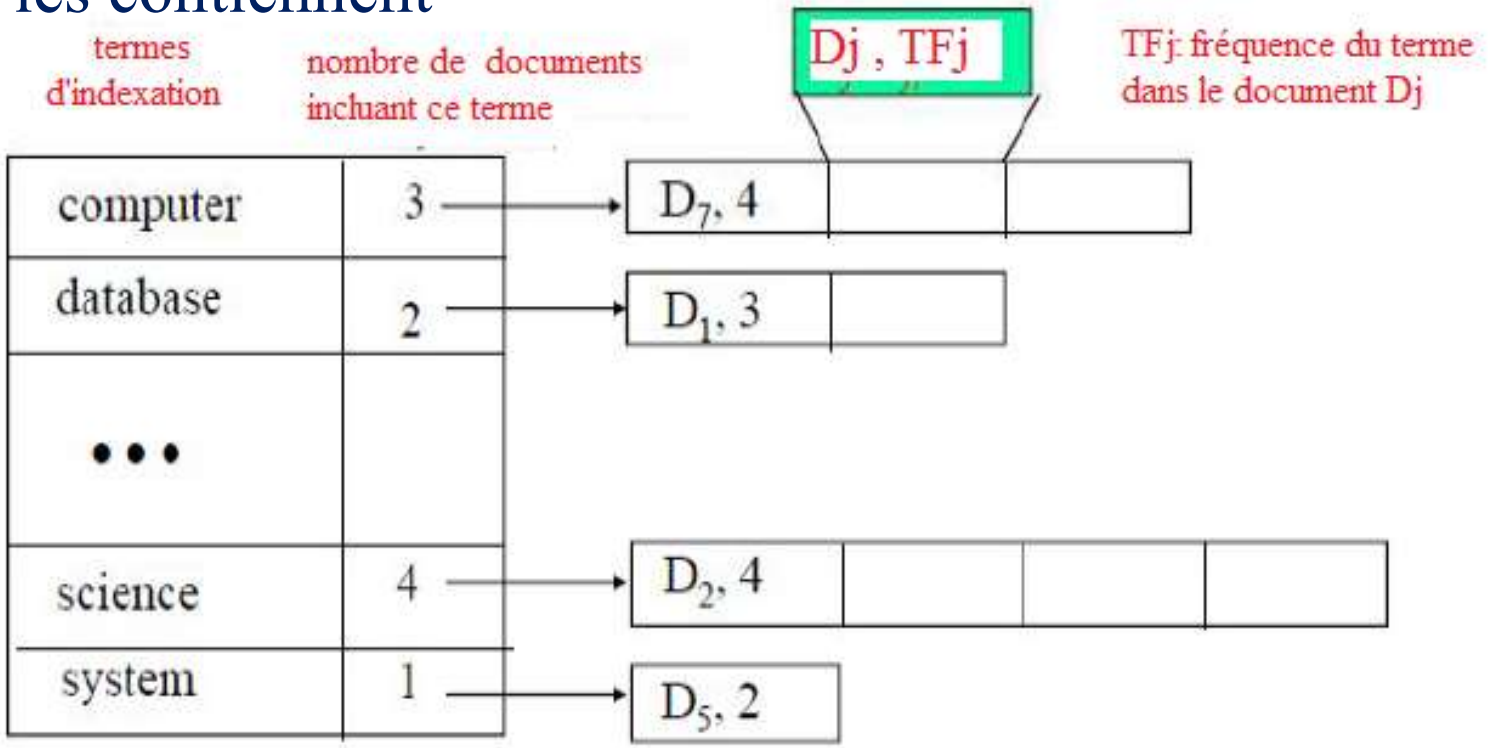
- Le terme «indexation» est parfois ambigu, car il est utilisé pour deux problèmes distincts :
  - le processus d'extraction des descripteurs à partir des documents,
    - descripteur représente le contenu d'un document ou d'une requête
    - doit refléter au mieux le contenu
  - la représentation de cette information.

Indexation textuelle

- Utiliser des mots-clés (texte)
  - termes significatifs se trouvant dans le document / requête

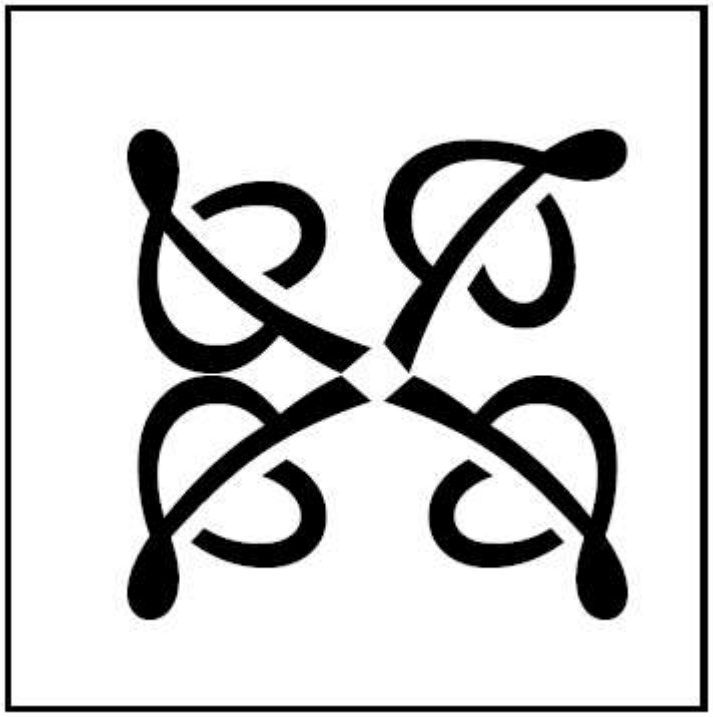
Indexation textuelle
structure d'un fichier inverse

- Un fichier inversé associe des index aux documents qui les contiennent



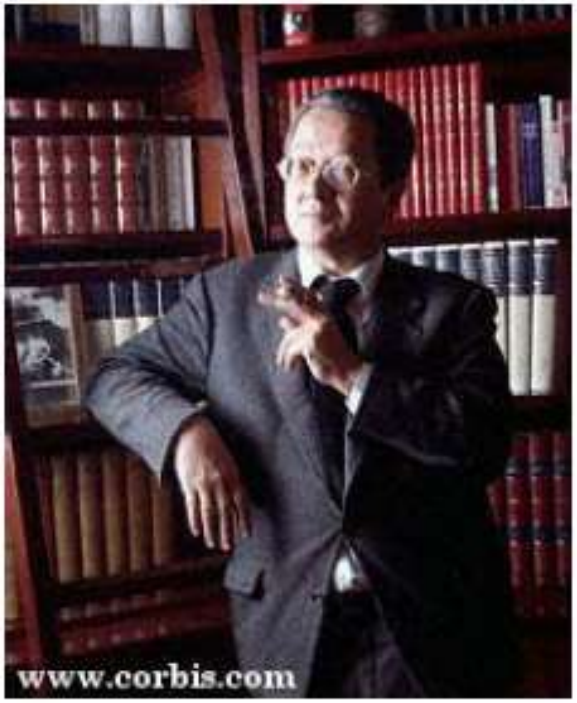
➔ Sera présenté en détail dans le chapitre 2

Indexation par mot-clé?



Recherche d'images

"*Avocat*" ?



Requête textuelle « Port du voile »

Indexation multimédia

- Utiliser des descripteurs multimédias
- L'image: exemples de descripteurs
  - Descripteurs visuels pour les base d'images génériques
    - couleur, texture, forme, position et relations spatiales
  - Segmentation et points d'intérêt pour les base d'images spécifiques
    - Détection et signatures de *visages*,
    - Signatures d'*empreintes digitales*

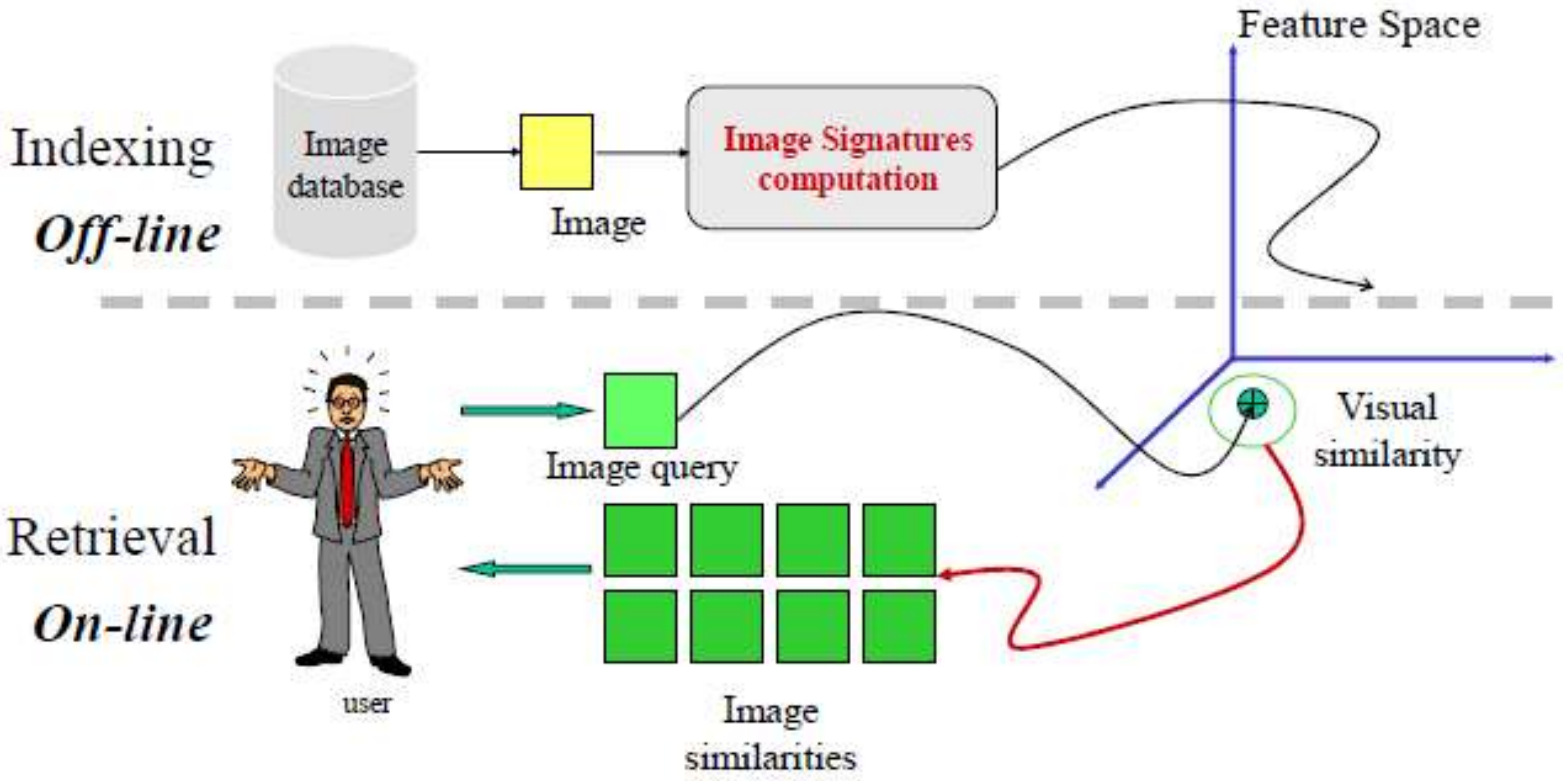
➔ Sera présenté en détail dans le chapitre 3

- Le son ne sera pas traité dans ce cours

Intérêt de l'indexation multimédia

- Indépendante de la langue de recherche
- Description moins subjective que le texte
- Parfois plus riche que le texte
- Souvent plus efficace (ambiguïté)

Architecture d'un système de recherche d'images par le contenu



Application de recherche d'images par le contenu

- Internet (images et vidéos),
- Audiovisuel (personnage dans les JT, documentaires, sport, …)
- Médecine (recherche à but diagnostic ou pédagogique)
- Art et Design (archives archéologiques, peintures, tissus, …)
- Authentification (visages, empreintes digitales, logos)
- Sécurité (surveillance vidéo, objets d'arts volés …)
- Education (recherche encyclopédique)

Appariement

- L'appariement (ou la correspondance) consiste à comparer la représentation de chaque document à celle de la requête
  - basé sur une fonction de similarité (ou de correspondance)

Fonction de similarité

- 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)

➔ Seront présentés en détail dans le chapitre 5

Fonctions de similarité

- Fonctions de similarité utilisées dans la recherche d'images par le contenu visuel
  - Distance euclidienne
  - Distances entre histogrammes
  - Distances quadratiques
  - Distances entre distributions
  - …

➔ Seront présentées en détail dans le chapitre 5

Modèle de RI

- Le modèle de la représentation + la fonction de similarité = Modèle de RI

- Modèle classiques de recherche textuelle
  - Modèle booléen
  - Modèle vectoriel
  - Modèle probabiliste

➔ Seront présentés en détail dans le chapitre 4

Reformulation

- Consiste à réécrire la requête initiale jusqu'à la satisfaction de l'utilisateur
  - Il est rare que la réponse à une question retournée par le SRI satisfait l'utilisateur dès le 1er essai
- Techniques de reformulation: la réinjection de pertinence (« relevance feedback » en anglais):
  - Réinjection positive: ajouter à sa requête des termes auxquels il n'aurait pas pensé et qui apparaissent dans des documents retournées qui sont pertinents
  - Réinjection négative: enlever de sa requête des termes qui apparaissent dans des documents retournées qui ne sont pas pertinents

Un peu d'histoire

- 1940 : arrivé des ordinateurs: la RI se concentrait sur les applications dans des bibliothèques.
- 1950 : Début de petites expérimentations en utilisant des petites collections de documents (références bibliographiques) + utilisation du modèle booléen.
- 1960-1970 : développement de méthodologie d'évaluation du SRI + conception de corpus pour évaluer les différents SRI.
- 1970 : Développement du système SMART (G. Salton): implantation et test pour la première fois du modèle vectoriel et la technique de relevance feedback + beaucoup de développements sur le modèle probabiliste.

Un peu d'histoire

- 1980 : intégration des techniques de l'IA en RI, par exemple, système expert pour la RI, etc.
- 1990 : la problématique est élargie, par exemple, on traite maintenant plus souvent des documents multimédia qu'avant. Cependant, les techniques de base utilisées dans les moteurs de recherche sur le web restent identiques.
- De nos jours, beaucoup de travaux sur la recherche des documents multimédias: développement de nouveaux systèmes de recherche d'images par le contenu

Références

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze.Introduction to Information Retrieval. 2008
- Baeza-Yates R. and Ribeiro-Neto B. Modern Information Retrieval - the concepts and technology behind search, 2011
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. AddisonWesley, 1999.
- Butterworths – Frakes and Baeza-Yates. Information Retrieval: Data Structures & Algorithms,1992
- Prentice Hall Witten, Moffat and Bell Managing Gigabytes plus software, 1994
- Salton Gerard and Mcgill, Michael J. Introduction to modern information retrieval. New York: Mcgraw-Hill Book Company.