
Introduction à la Recherche d'information



M. Boughanem

bougha@irit.fr

<http://www.irit.fr/~Mohand.Boughanem>

Université Paul Sabatier de Toulouse

Laboratoire IRIT

Plan

- Introduction à la Recherche d'information (RI)
- Tour d'horizon sur les techniques de RI
- Recherche d'information sur le Web
- Où en sommes nous

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**



Définition de wikipédia

- **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.
- Automated information retrieval systems
 - 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.

Exemples de Systèmes de RI

Google™

YAHOO!®

msn™

AOL.

Ask Jeeves®
Ask.com

alltheweb
• • • find it all • • •

altavista™

.. . les moteurs de recherche

traitent une petite partie de l'information disponible

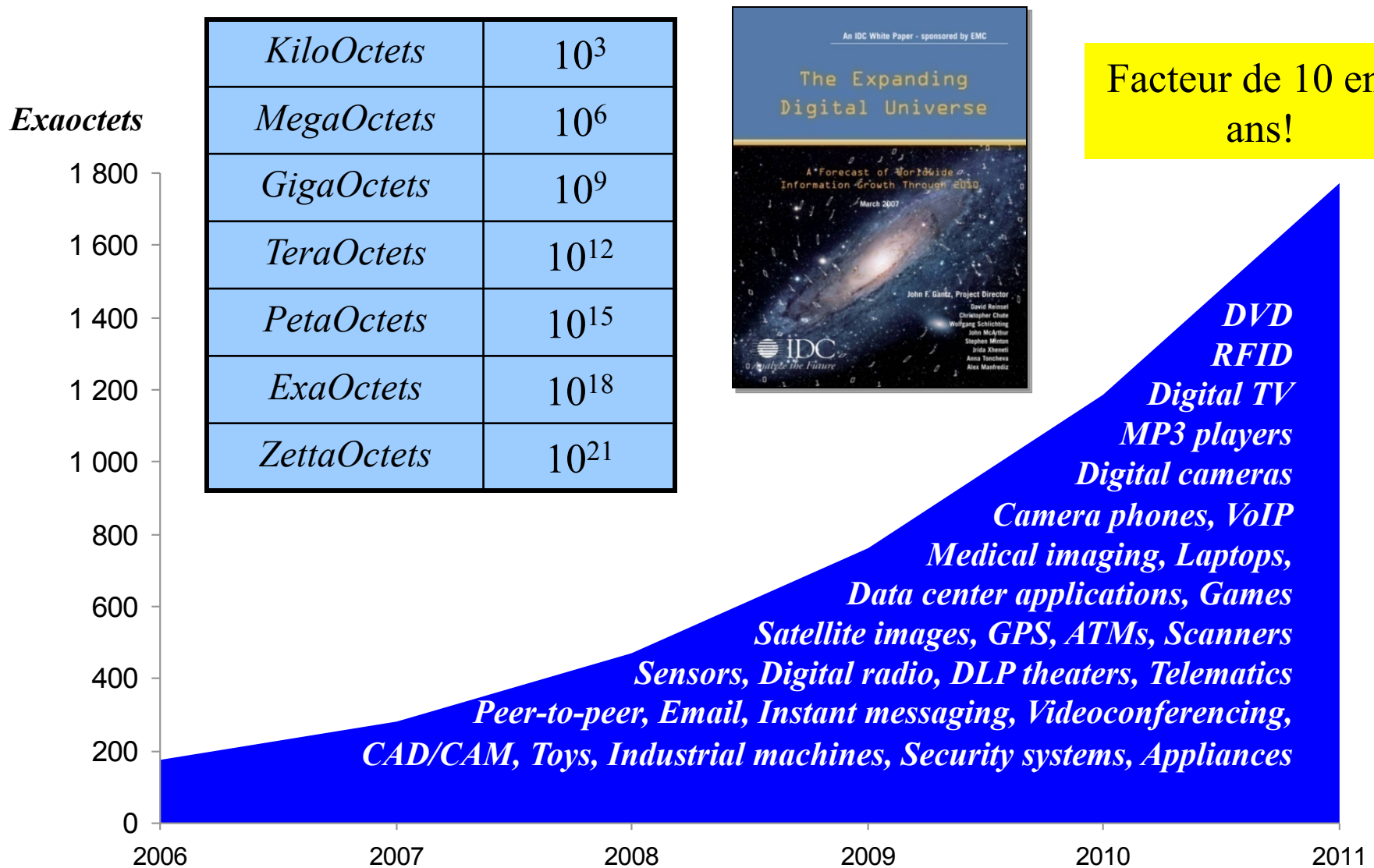
- Internet (web localisé par les moteurs) juste la partie visible de l'iceberg (40 milliards de pages google)
- Intranet (entreprises, laboratoires, ...)



Volume sans cesse croissant

- Gros volumes d'informations (numériques) créées toutes les minutes
- 1% sur l'Internet (localisé par les moteurs)
- 99% dans des Intranet (entreprises, laboratoires, ...)

.. en perpétuelle croissance



.. produite par tout le monde



Vragenlijst website

Heeft u interesse in een website en wenst u dat één van onze medewerkers met u contact opneemt, wil dan deze onderstaande vragenlijst invullen.

Persoonlijke gegevens

Naam & voornaam

Tel./telefoonnummer

E-mail

.. L'information (numérique) est disponible partout

Social Media Landscape



© google



© google

... dans tous les domaines d'activités

- Domaines d'application
 - Internet (Web, *Forum/Blog search*, News)
 - Entreprises
 - Bibliothèques numériques «*digital library*»
 - Domaine spécialisé (médecine, droit, littérature, ...)
 - Nos propres PC (*Yahoo! Desktop search*)

... la RI a un coût

- 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 la RI

- Données, information et connaissance
- Tâches de recherche d'information

Contours: Données–Information–Connaissance

Système de gestion de
Base de données

Système de Recherche
d'information

Données :

Chaîne de caractères/valeurs
associées à des objets, des
personnes et des événements :
(15)

Information :

*Signification (explication/
description) des données,*
*données intelligible (15° C -
relevé à 18 h, sous abri, à XX)*

Connaissance :

*Information apprise, découverte,
comprise et partagée par une
communauté*
*(étant donné qu'on est à XX 15°C en
avril c'est plutôt froid)*

(information/data
mining, fouille de
données)

Contours :

Tâches de recherche d'information (1 / 2)

- Recherche adhoc (classique)
 - Je cherche des infos (pages web) sur «veille économique »
 - Requête «veille économique» → SRI → renvoie une liste de documents
- Classification /catégorisation (*clustering*)
 - Regrouper les informations (documents) selon un ou plusieurs
- Question-réponses (*Query answering*)
 - Chercher des réponses à des questions
 - par exemple «qui est averroes ? »
 - « Quelle est la longueur du Nil ? »



averroes

Input interpretation:

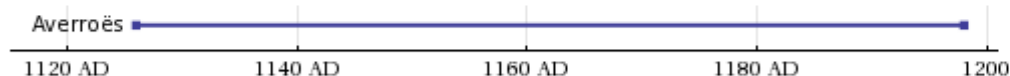
Mathematica form

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

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 »

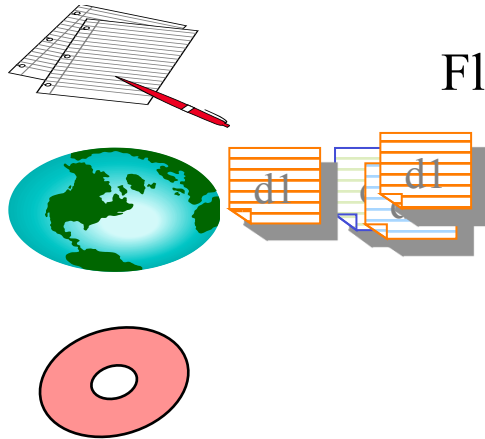
Contours :

Tâches de recherche d'information (2/2)

- Filtrage d'information (*filtering/recommendation*)
 - Dissémination sélective d'information

Filtrage d'information

Sources



Flot de documents



SFI/push

Utilisateurs ayant des besoins
définis au préalable
= Profils



Terminologie :

- Push
- Filtrage d'information
- Recommandation
- Dissémination sélective d'information
- Système d'alerte



Bonjour Boughanem. Découvrez [nos conseils personnalisés](#). ([Vous n'êtes pas Boughanem ?](#))

[Chez Boughanem](#) | [Nos bonnes affaires](#) | [Chèques-cadeaux](#) | [Listes et idées cadeaux](#)

Livraison en 24h

Toutes nos boutiques

Rechercher Toutes nos boutiques

GO

Chez vous

Votre page à votre image

Recommandé pour vous

Évaluer ces articles

Affinez nos conseils personnalisés

Votre profil

Plus

Boughanem, Bienvenue sur Votre Amazon.fr (Si vous n'êtes pas Boughanem, [cliquez ici](#).)

Nos recommandations du jour pour vous

Voici un échantillon quotidien d'articles qui vous sont recommandés Cliquez ici pour [voir toutes les recommandations](#).

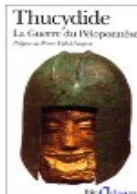
Page 3 sur 4 (Rev



[Le Moyen Âge en Orient : Byzance...](#) (Broché) de Alain Ducellier

EUR 26,51

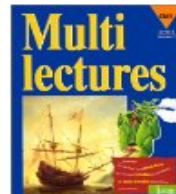
[Corriger cette recommandation](#)



[La Guerre du Péloponnèse](#) (Poche) de Thucydide

★★★★★ (6) EUR 10,83

[Corriger cette recommandation](#)



[Multilectures, CM1, cycle 3, niveau 2](#) (Reliure inconnue) de Gehin

EUR 12,34

[Corriger cette recommandation](#)



[L'Horloger aveugle](#) (Broché) de Richard Dawkins

EUR 21,85

[Corriger cette recommandation](#)



[Bled, cahiers d'activités CP-CE1. Cycle...](#) (Broché) de Bled

★★★★★ (2) EUR 4,56

[Corriger cette recommandation](#)

De votre liste d'envies cadeaux

Vous n'avez pas d'articles dans votre [Liste d'envies cadeaux](#).

Votre compte

Suivi des colis

Modifier votre nom, votre e-mail et votre mot de passe

Vos notifications par e-mail

Toutes les options du compte

Vos achats récents

Articles consultés

récemment (0)

Votre panier

Commandes ouvertes & expédiées récemment

Vos listes

Votre liste d'envies cadeaux

Votre communauté

Vos rappels

Votre profil

Votre participation

Vos commentaires

Vos listes thématiques

Listmania

Vos images

Affinez nos conseils perso

L'unicité du savoir

Évaluez cet article

x|★★★★★

☐ Ne pas utiliser pour les recommandations

Articles que vous possédez

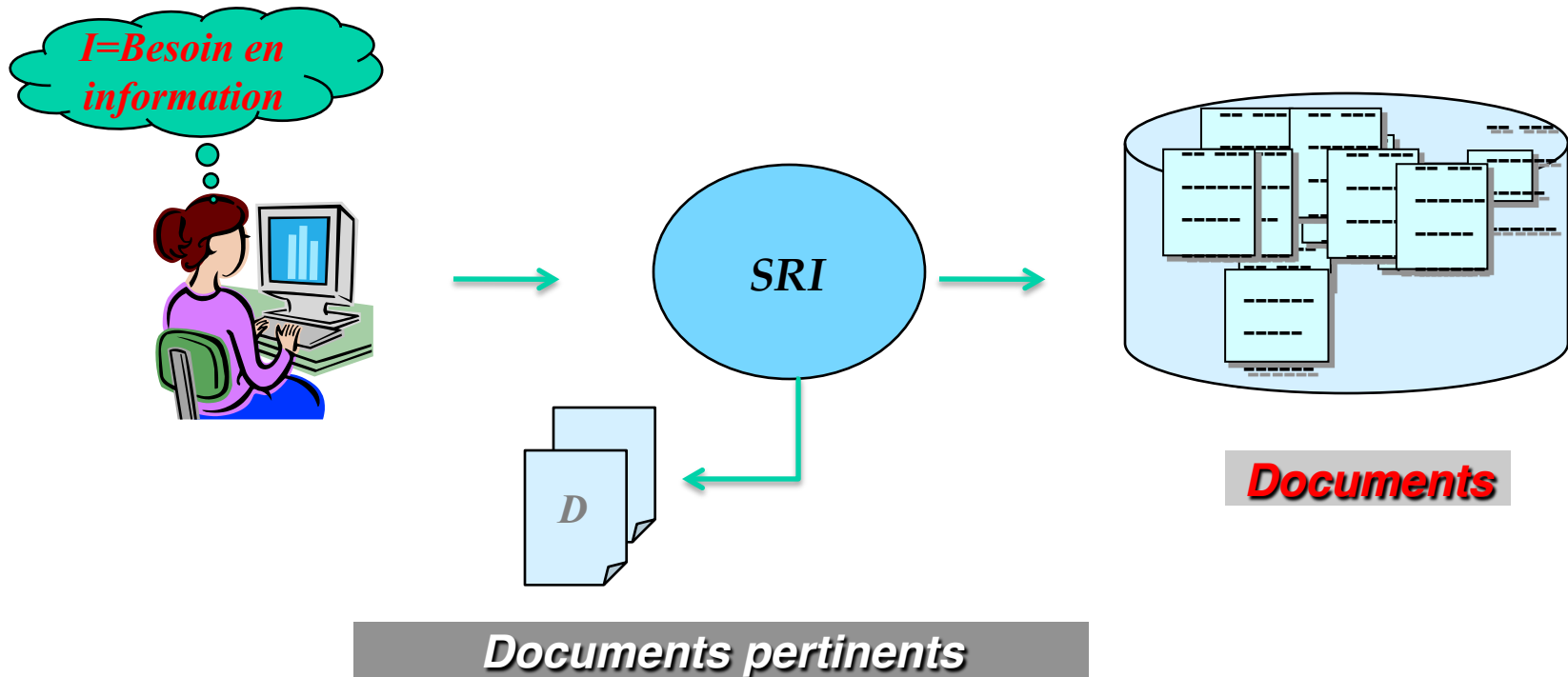
Articles notés

Contours :

Tâches de recherche d'information (2/2)

- Filtrage d'information (*filtering/recommendation*)
 - Dissémination sélective d'information
- Croisement de langues (*cross language*)
 - Rechercher des infos écrites dans une langue autre que celle de la requête
 - Requête en français → retour documents en arabe
- Recherche d'opinions (Opinion retrieval) (sentiment analysis)
 - Pertinence(relevant) +subjective (opinionated)
 - Polarité (Positive, négative et neutre)

La RI adhoc : concepts de base



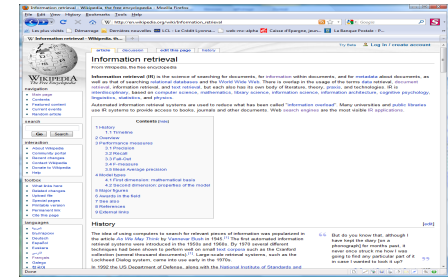
- Sélectionner dans une collection
 - les **informations**
 - ... **pertinentes** répondant aux
 - ... **besoins en information des utilisateurs**

- Formes

- Texte, images, sons, vidéo, graphiques, etc.
- Exemples texte : web pages, email, livres, journaux, publications, messages email, blog, Word™, Powerpoint™, PDF, forum postings, brevets, etc.

- Hétérogénéité

- langage (multilingues)
- media (multimédia)
- structures



Besoin en information vs. requête

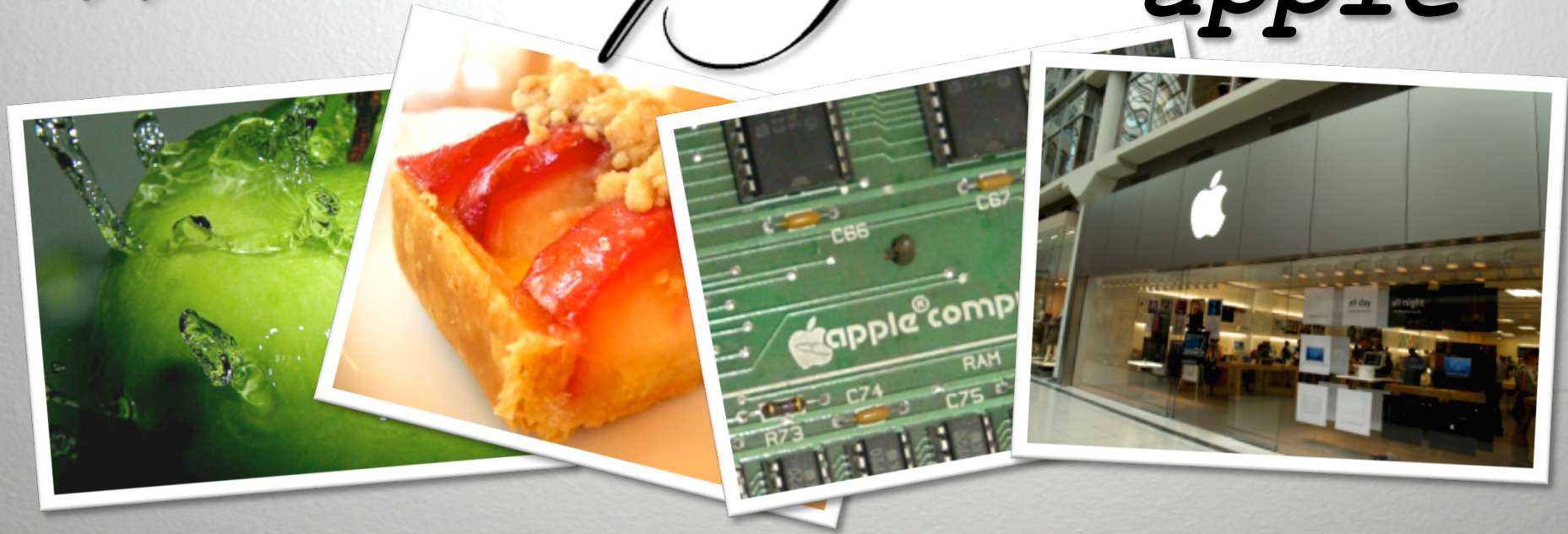
- Besoin en information est une expression mentale d'un utilisateur
- Requête est une représentation possible du besoin en information



Query intent → what's a query

What's in a query?

apple

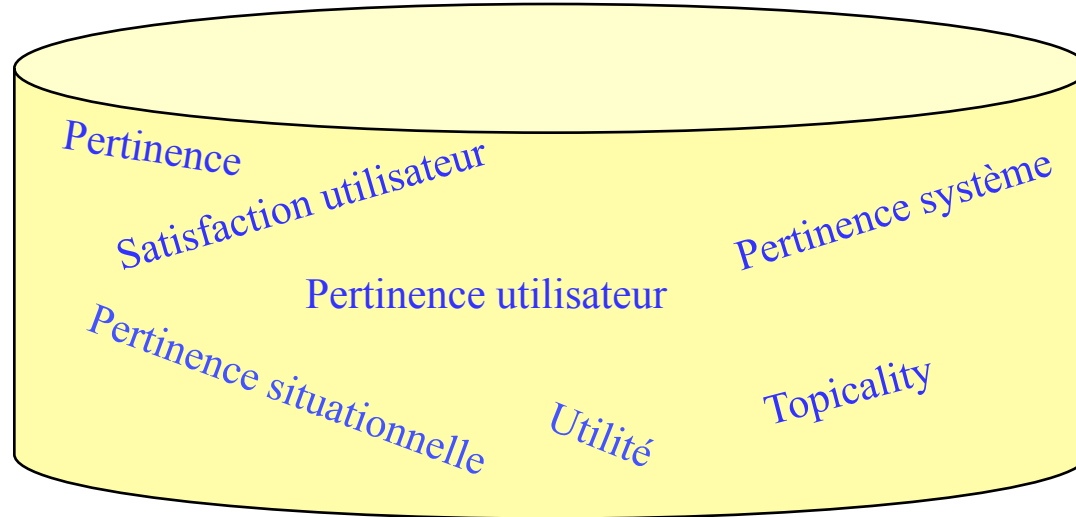


Qu'est ce qu'il y a derrière une requête ?

- Une même interface pour des besoins variés
- .. Deviner ce qui est caché derrière la requête...
- Exemples :
 - BBC → (site web de la BBC)
 - $5+4$ → calcul
 - Air France 454 → Infos sur un vol
 - 1kg en livres, 1 euro en dollars → Conversion
 - dentiste Lyon → Adresse professionnelle

Pertinence

- Quelle pertinence ?



- Relation (correspondance,...) entre un document et

..... une **requête**

ou.....un **besoin en information** ?



Selon **l'utilisateur** ...ou **le système** ?

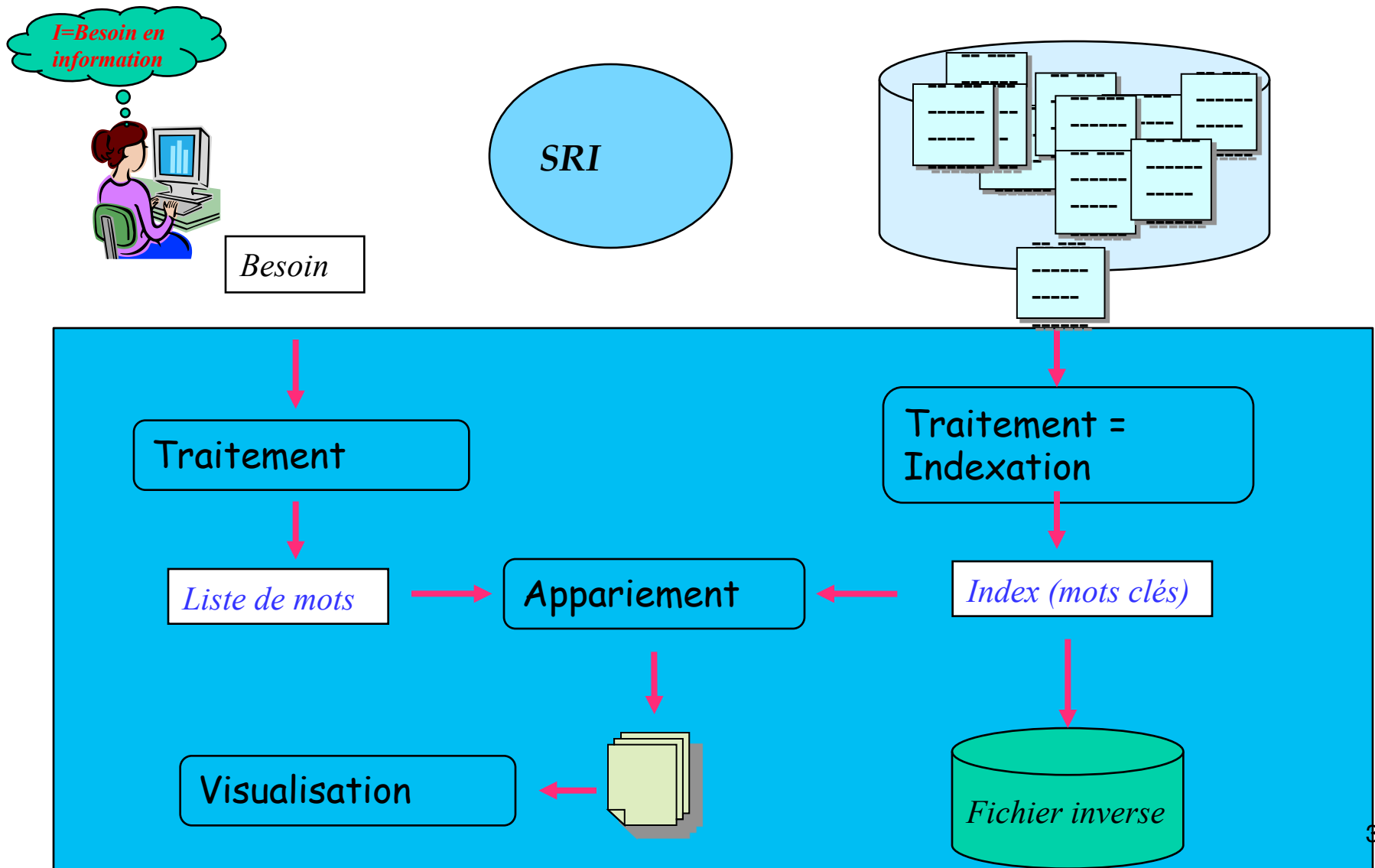
Plusieurs pertinences

- Pertinence Algorithmique : mesure algorithmique dépendant du modèle de RI
- Pertinence thématique : relation entre le sujet exprimé dans la requête et le sujet couvert dans le document.
- Pertinence cognitive : relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée.
- Pertinence situationnelle/contextuelle : relation entre la tâche, le problème posé par l'utilisateur, la situation de l'utilisateur et l'information retrouvée.

Hypothèses « simplificatrices » en RI

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

Tour d'horizon sur les techniques de RI : Ouvrir la boîte noire



Tour d'horizon sur les techniques de RI : Indexation

- Indexation
 - Processus permettant de construire un ensemble d'éléments « clés » permettant de caractériser le contenu d'un document / retrouver ce document en réponse à une requête
- Approches :
 - Statistique (distribution des mots) et/ou TALN (compréhension du texte)
 - Approche courante est plutôt statistique avec des hypothèses simples
 - Redondance d'un mot marque son importance
 - Cooccurrence des mots marque le sujet d'un document

Information retrieval - Wikipedia, the free encyclopedia - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://en.wikipedia.org/wiki/Information_retrieval

Les plus visités Démarrage Dernières nouvelles LCL - Le Crédit Lyonna... web-mc-alpha Caisse d'Épargne, jeun... La Banque Postale - P...

Information retrieval - Wikipedia, th...

Try Beta Log in / create account

Information retrieval

From Wikipedia, the free encyclopedia

Information retrieval (IR) is the science of searching for documents, for **information** within documents, and for **metadata** about documents, well as that of searching **relational databases** and the **World Wide Web**. There is overlap in the usage of the terms **data** retrieval, **document** retrieval, information retrieval, and **text** retrieval, but each also has its own body of literature, theory, **praxis**, and technologies. IR is **interdisciplinary**, based on **computer science**, **mathematics**, **library science**, **information science**, **information architecture**, **cognitive psychology**, **linguistics**, **statistics**, and **physics**.

Automated information retrieval systems 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.

Contents

- 1 History
 - 1.1 Timeline
- 2 Overview
- 3 Performance measures
 - 3.1 Precision
 - 3.2 Recall
 - 3.3 Fall-Out
 - 3.4 F-measure
 - 3.5 Mean Average precision
- 4 Model types
 - 4.1 First dimension: mathematical basis
 - 4.2 Second dimension: properties of the model
- 5 Major figures
- 6 Awards in the field
- 7 See also
- 8 References
- 9 External links

History

The idea of using computers to search for relevant pieces of information was popularized in the article *As We May Think* by **Vannevar Bush** in 1945.^[1] The first automated information retrieval systems were introduced in the 1950s and 1960s. By 1970 several different techniques had been shown to perform well on small **text corpora** such as the Cranford collection (several thousand documents).^[1] Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992 the US Department of Defense, along with the **National Institute of Standards and**

“ But do you know that, although I have kept the diary [on a phonograph] for months past, it never once struck me how I was going to find any particular part of it in case I wanted to look it up?

Tour d'horizon sur les techniques de RI indexation

- Décomposer le texte
- Décomposer les mots
- Supprimer les mots communs
 - Basé sur une “short list” “the”, “and”, “or”
- Radicaliser les mots
- Regrouper les mots

Title: Information retrieval
Corps du texte : IR is the science of searching of documents

Information, retrieval, IR, is, the science, of ,searching, of, documents

Information, retrieval, IR, science, searching, documents

*Information, retrieval, IR, science, **search**, document*

*Information 1, retrieval 1, IR 1, science 1, **search 1**, document 1*

Organiser les termes et les documents dans un 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
1	1
2	1
2	1
1	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.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Répondre à une demande (requête)



caesar

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	1
1	1
2	2
1	1
1	1
2	1
1	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

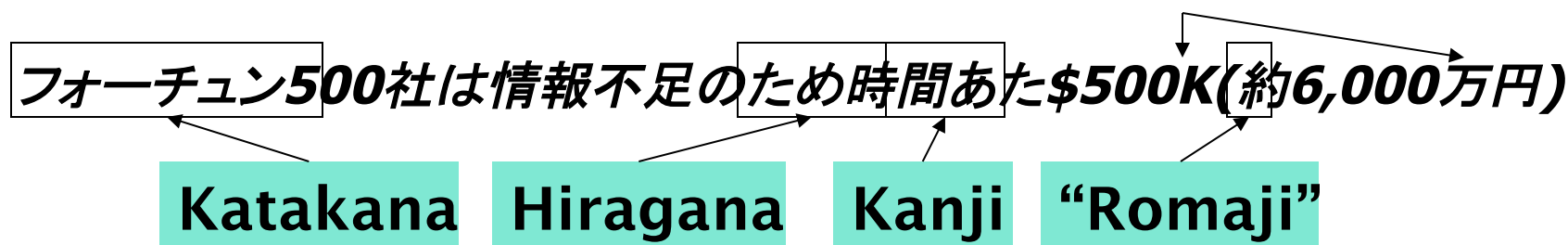
d1:
So let it be with
Caesar. The noble
Brutus hath told you
Caesar was
ambitious

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

Indexation dans d'autres langues

- Pas d'espaces en chinois et en japonais
 - Ne garantit pas l'extraction d'un terme de manière unique
- Japonais encore plus compliqué avec différents alphabets



L'utilisateur peut exprimer sa requête entièrement en Hiragana

Matching/appariement : modèles de RI

- Facteurs utilisés par la majorité des modèles
 - Fréquence du terme dans le document (*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*

IR models

- 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)
- Learning to rank

Facteurs utilisés par les moteurs pour trier des pages

- Plusieurs facteurs plus de 200
- Classés selon deux groupes : dépendant /indépendant contenu de la page
 - Facteurs dépendant du contenu de la page web (mots-clés dans la page web)
 - Classement des pages dépend fortement de la localisation (position) des mots de la requête dans la page
 - → Titre page, mots dans une ancre (anchor), domaine, URL, Meta...

Pour plus de détails sur le « ranking » sur le web



SEARCH ENGINE
RANKING FACTORS V2

- <http://www.seomoz.org/>
- <http://www.seomoz.org/article/search-ranking-factors>

Exemple top acteurs utilisés par Google pour trier les pages

- Mots dans le titre de la page (Keyword Use in Title Tag)
- Mots dans les liens ('anchor') provenant de l'extérieur (Keyword Focused Anchor Text from External Links)
- Popularité du site (PageRank)
- Diversité des liens externes (Diversity of Link Sources)
- Fiabilité du domaine (fonction de la distance de votre site vis-à-vis d'un site de confiance, cette liste est inconnue)
- Mots dans le nom du domaine (www.ummtto.dz) (Keyword Use in the Root Domain Name)
- Mots dans le nom du sous-domaine (mots.domaine.fr) (Keyword Use in the Root Domain Name)

Exemples de Facteurs utilisés par les moteurs pour trier les pages

- Mots clés dans la balise H1 (Keyword Use in H1)
- Mots dans les balises H2, H3 Page URL
- Mots clés dans les Meta Description / Keywords Tags
- Mots dans le Body (Keyword Use in Body Text)
- HTML Validation

Facteurs utilisés par les moteurs pour trier les pages

- **Top 5 facteurs négatifs**
 - Serveur souvent inaccessible
 - Contenu similaire ou dupliqué (copie de site)
 - Liens vers des pages de qualité douteuse (Spam)
 - Liens provenant de sites (courtier/"Vendeurs" de liens)
 - Sur-utilisation de mots clés

*D'après Amit Singhal (Architecte de Google)
lors de sa conférence à ECIR '2008
Google utilise 263 facteurs*

Plan

- Recherche d'information (RI)
 - Intérêt et contours
 - Tour d'horizon sur les techniques de RI
- Cas d'étude: Recherche d'information sur le Web
- **Ou en sommes nous**

Points clés dans l'histoire de la RI

- 1952 Calvin N. Mooers invente le mot « *IR* »
- 1959 Luhn (RI-statistique)
- 1960 Cranfield I (démarche de validation)
- 1960 Maron and Kuhns (modèle probabiliste)
- 1961 (-1965) Smart (le modèle vectoriel)
- 1968 Premier livre de Salton
- 1975 Livre C van Rijsbergen (accessible sur le web, ver. 1979)
- 1977 Modèle probabiliste (PRP) S. Robertson
- 1978 Première conférence SIGIR
- 1983 Début d' Okapi (modèle probabiliste)

Points clés dans l'histoire de la RI

- 1985 RIAO-1 Grenoble
- 1986 Modèle logique («Keith» van Rijsbergen)
- 1990 (tout début du) Learning to rank (développement dans les années 2000)
- 1990 Modèle LSI (Dumais, Deerwester ...),
- 1992 TREC-1
- 1998 Modèle de langue
- 1998 Google
- 2000 CLEF
- 2002 INEX
- 2004 CORIA (Conférence francophone en recherche d'information)

États des lieux

- Plusieurs avancées :
 - Techniques de pondération (luhn, ... BM25, modèle de langue)
 - Plusieurs modèles
 - Reformulation de requêtes (*relevance feedback*, expansion ...)
 - RI multilingue
 - Évaluation des performances (métriques, benchmark)
 - Architecture physique (inverted file)

Constat : les techniques statistiques sont effectives en RI

Etat des lieux

- Ces techniques “marchent” aussi pour
 - D’autres langues,
 - eg Chinois
 - Cross-language:
 - English $\leftrightarrow$ chinese
- Les méthodes statistiques “marchent”
 - D’autres types de documents -
 - eg homepages, links & URLs

Etat des lieux

- Les méthodes statistiques se basent sur la redondance
- Toutes les utilisations de la langue ont une redondance
- → Les techniques statistiques semblent être les outils de base pour le traitement de l'information

Etat des lieux

- Plusieurs programmes d'évaluation :
 - TREC (Text REtrieval Conference)- DARPA,NIST,ARDA
 - NTCIR (Evaluation campaign on Asian documents)-
 - CLEF (Cross Language Evaluation Forum European Language)
 - INEX (INitiative for the Evaluation of XML Retrieval)
 - Collections volumineuses (topics, documents, relevance judgments)
 - plusieurs participants
- comparaisons riches

Quelques remarques sur le domaine

- Relativement peu de théorie
- Très forte tradition d'expérimentations
- Plusieurs problèmes et verrous à investir
- Domaine difficile : les méthodes intuitives ne sont pas forcément celles qui sont efficaces dans la pratique

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
- CORIA : Conférence Francophone en Recherche d'Information et Applications
- SPIRE : String Processing and Information Retrieval
- ECDL : European Conference on Digital Libraries
- ACM-SAC-IAR : Information Access and Retrieval (IAR)
- RIAO : Recherche d'Information Assisté par Ordinateur

- **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
- Foundations and Trends® in Information Retrieval

Merci