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

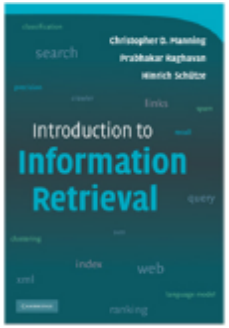
Plan

- 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

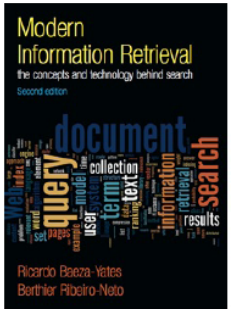
Qu'est ce que la RI ?

- 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

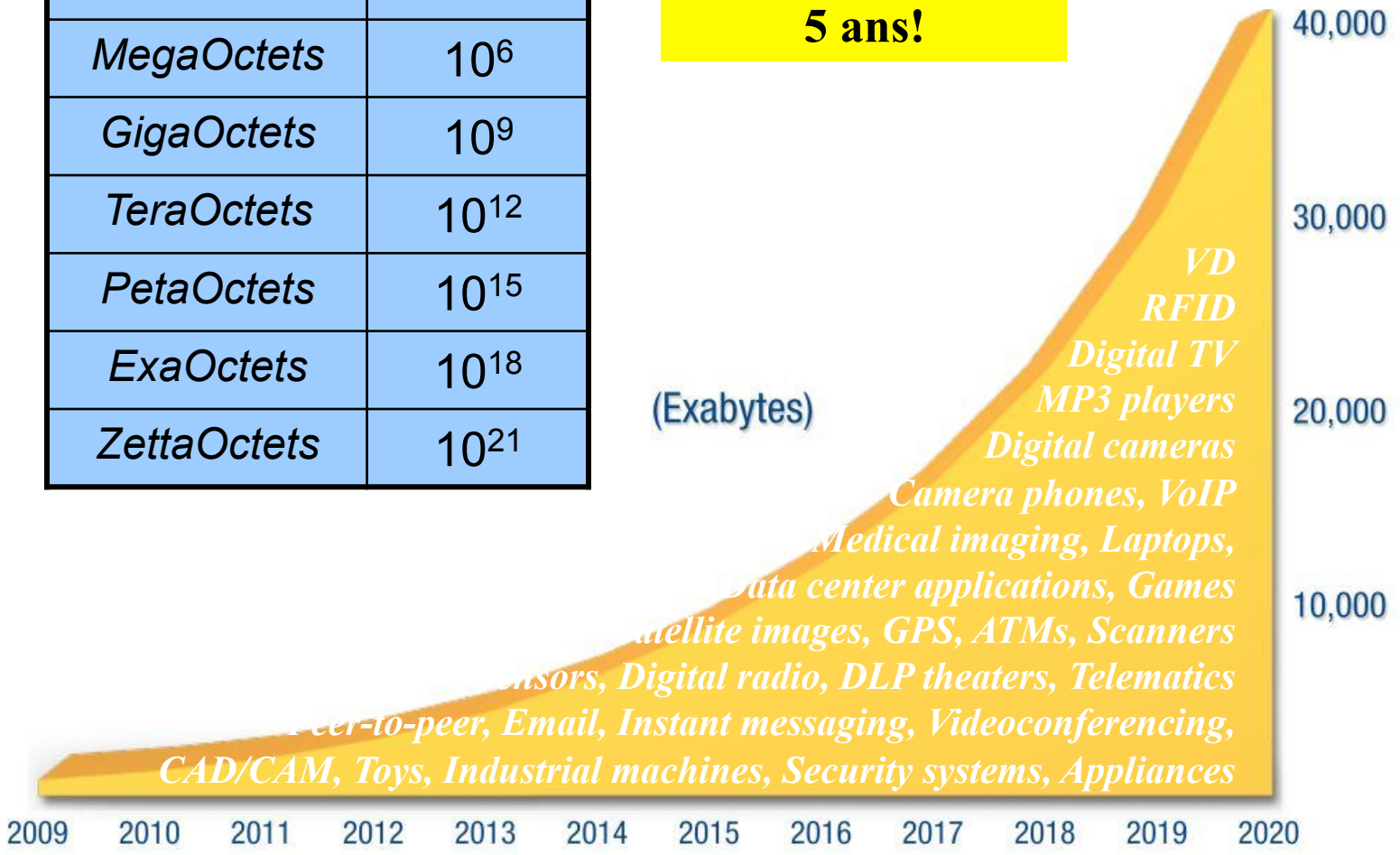
.. Mais pas seulement

- 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!

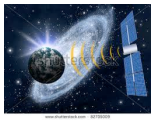


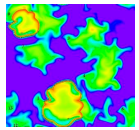
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Source: IDC, 2012

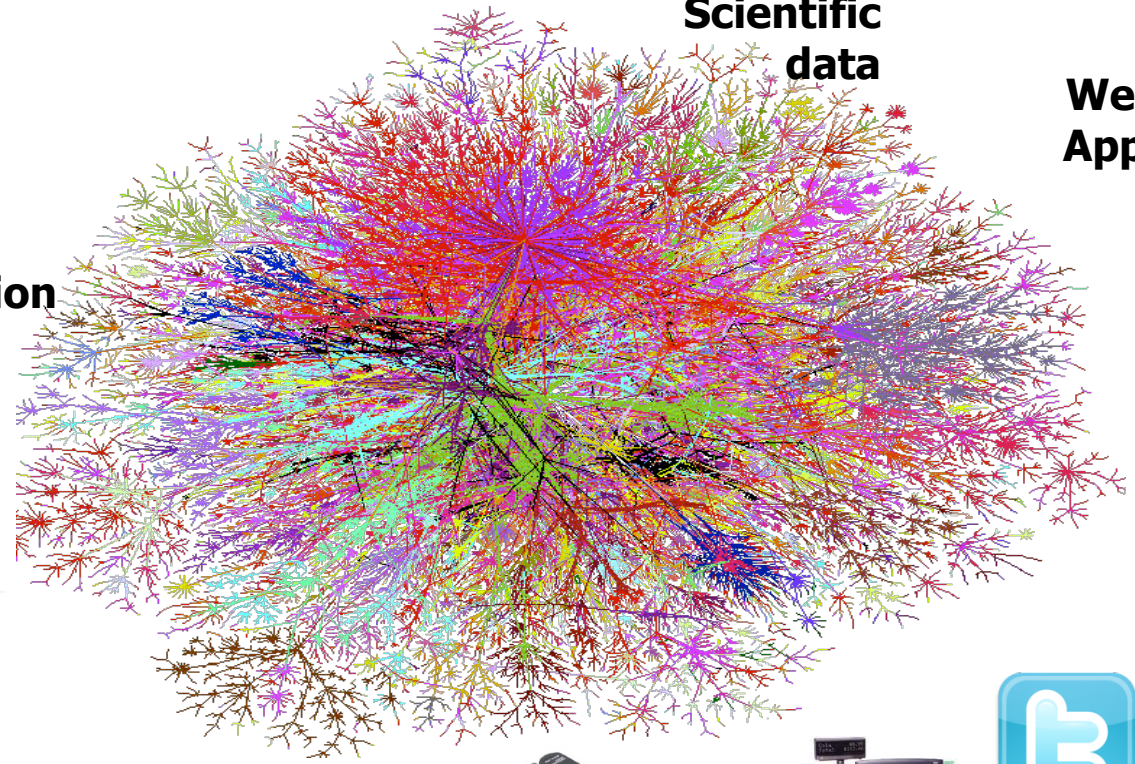
Origine

Enterprise Apps

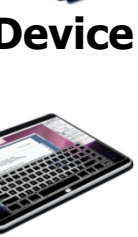




Scientific data



Web Apps



Device explosion

Social Media Data

Machine Data

.. 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...

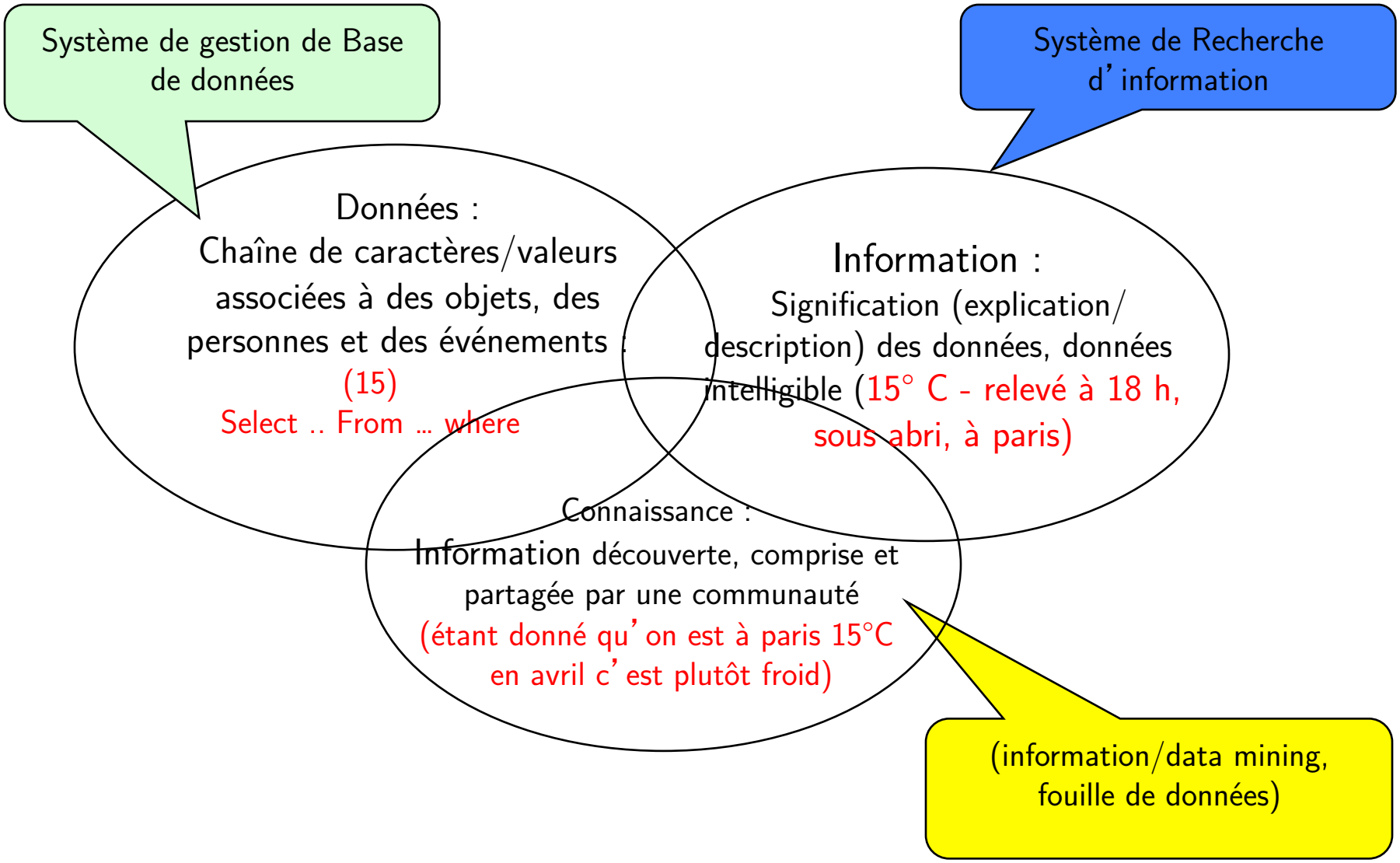
- Rechercher une information a un coût
  - « On» passe (en moyenne) 35% de son temps à rechercher des informations
  - Les managers y consacrent 17% de leur temps
  - Les 1000 grandes entreprises (US) perdent jusqu'à $2.5 milliards par an en raison de leur incapacité à récupérer les bonnes informations

- 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

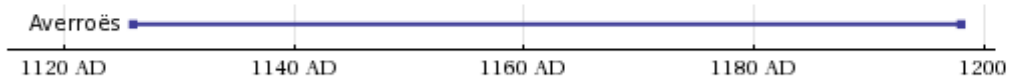
Input interpretation: *Mathematica* form

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

Page D'accueil Le Cop

Musée

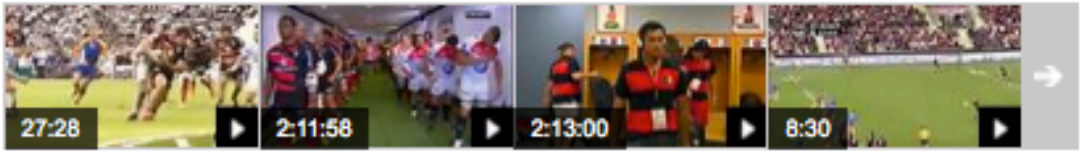
Stade Toulousain - Page d'accueil Translate this page

www.stadetoulousain.fr/index2.php

Saracens / **Stade Toulousain** - Interview de Maxime MÉDARD Election du stadiste de la saison. Le **Stade** dans les Médias . Suivre ...

Videos of stade Toulousain

bing.com/videos



Compilation des essais du Stade ... YouTube

Stade Toulousain - RC Toulon [Final... YouTube

Stade Toulousain - Montpellier [Final... YouTube

stade toulousain compilation YouTube

Stade toulousain - Wikipédia Translate this page

fr.wikipedia.org/wiki/**Stade_toulousain**

Histoire · Palmarès · Les finales du **Stade** ... · Personnalités ...

Stade toulousain Généralités Fondation 1907 Statut professionnel depuis le 1 er février 1998 Couleurs rouge et noir **Stade Stade** Ernest-Wallon (19 500 places ...

Rugbyrama

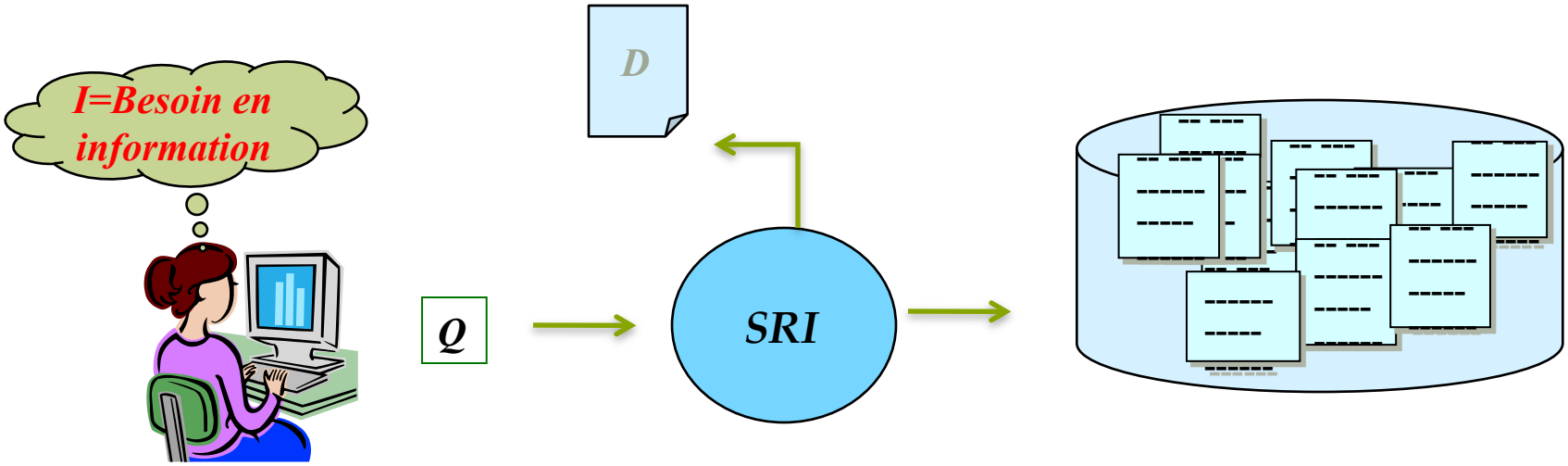
Stade **Toulouse Transferts**

Stade **Français**

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

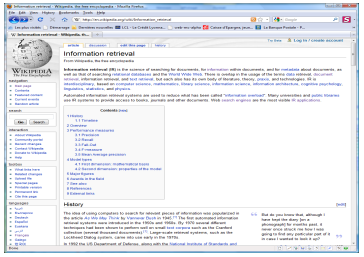
Problématique de la RI



- 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)

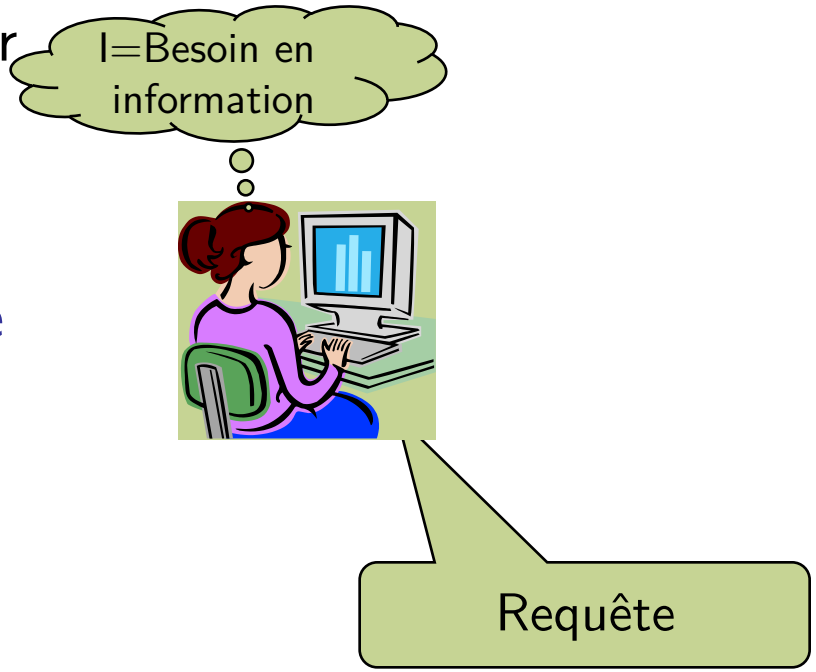


Question

- 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*"

What's in a query?

apple

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

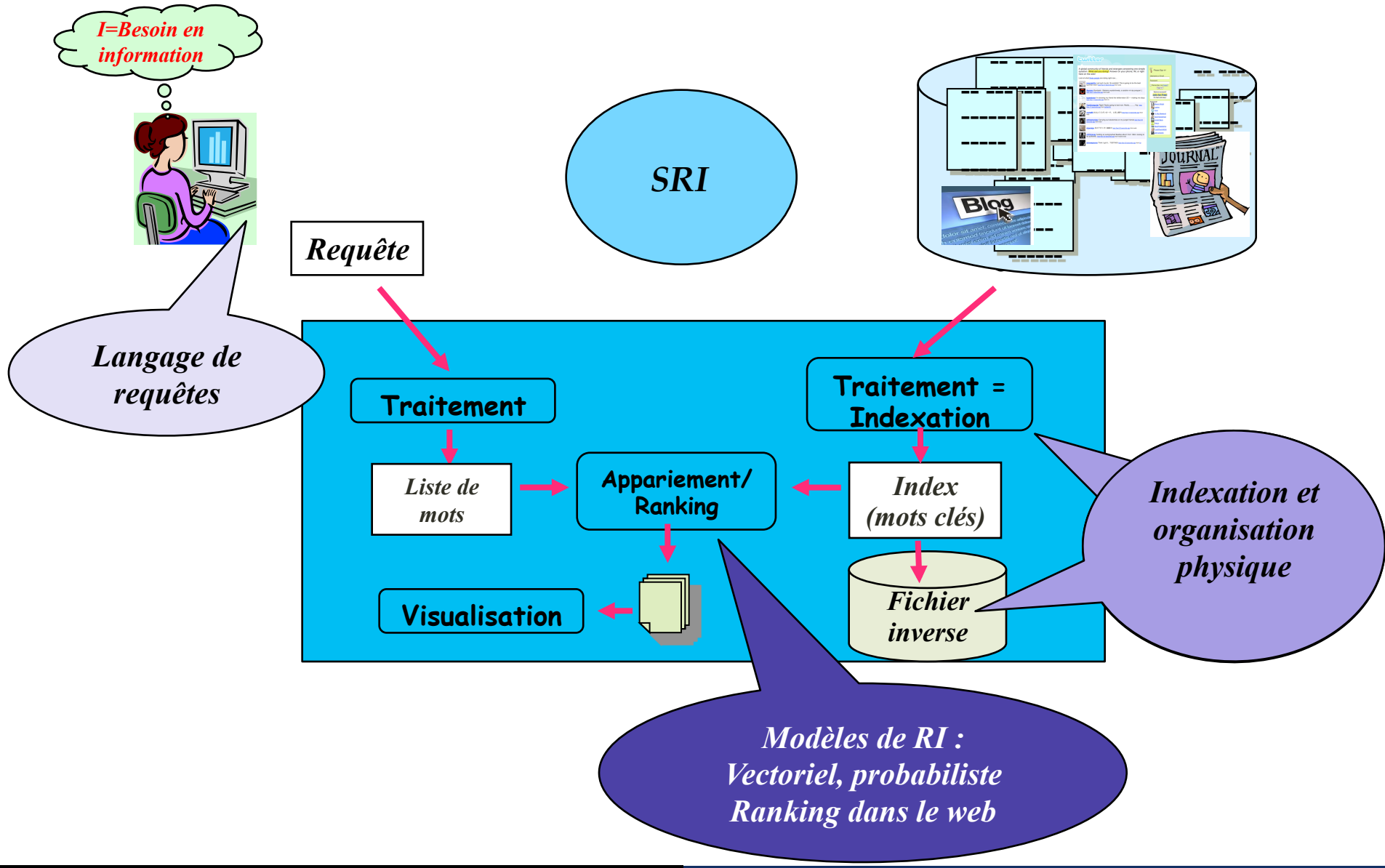
Hypothèses simplificatrices

- 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

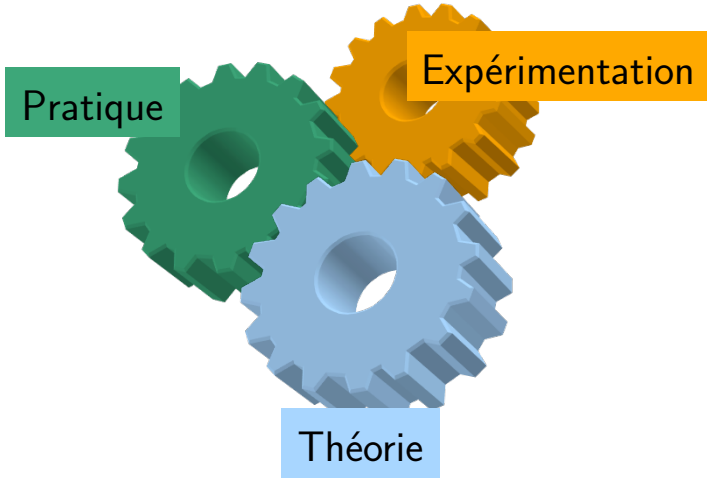


Récap.

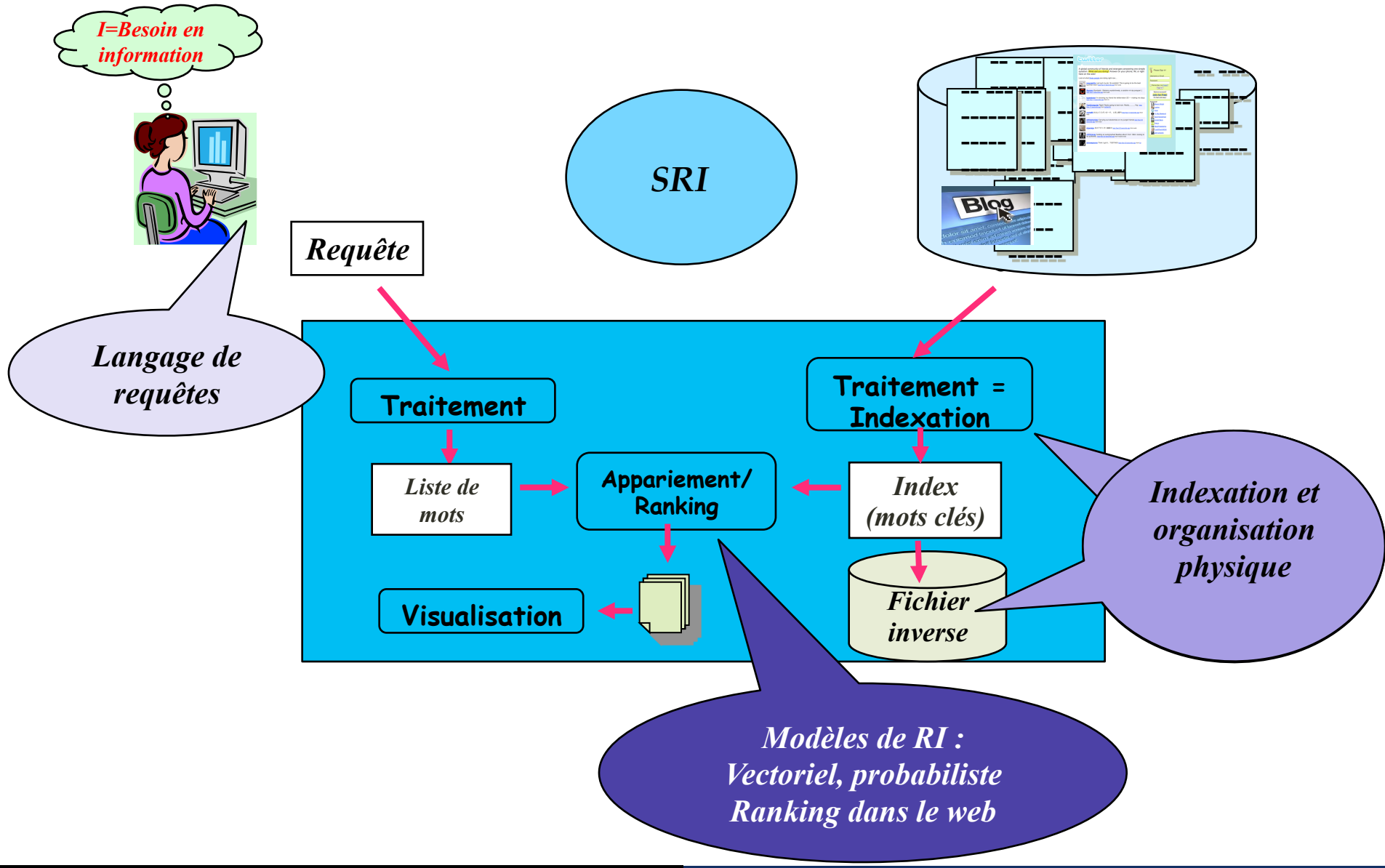
- Représentation (indexation) du document
  - Comment construire une représentation à partir du document ?
  - Quelle organisation physique pour ces index?
- Représentation des besoins
  - Comment capturer le besoin de l'utilisateur ?
  - Comment exprimer le besoin (langage de requêtes) ?
- Comparaison/ranking document-requête (des représentations)
  - Comment mesurer (décider) la pertinence d'un document ?
- Évaluation des performances
  - Comment comparer les SRI ?
  - Quelle démarche (empirique/ analytique) ?
  - Quelles métriques ?

Récap.

- Proposer des solutions
  - modèles, techniques, approches, outils pour répondre à ces problèmes
- ...avec 2 soucis majeurs
  - Quels supports théoriques ?
    - Souvent basés sur des théories mathématiques : Probabilités, statistiques, ensembles, algèbre, logique floue, analyse de données, ...
  - Quel processus pour la validation ?



Processus de RI



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
  - Guidée par un vocabulaire contrôlé vs. Libre
  - 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

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> : information retrieval (IR) is the science of searching of documents

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

Information, retrieval, information, retrieval, IR, science, searching, documents

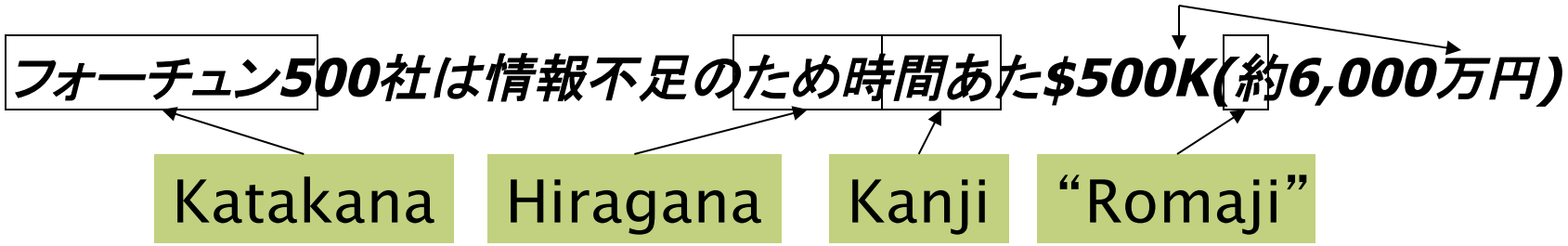
Information, retrieval, information, retrieval, IR, science, search, document

Information 2, retrieval 2, IR 1, science 1, search 1, document 1

Un sac de mots (BOW)

Indexation

- 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

Tour d'horizon
(Ouvrir la boîte noire)

Fichier inverse

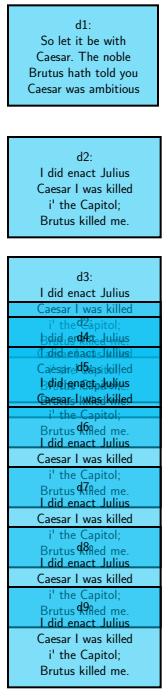
d1:
So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

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



Fichier inverse

Dictionnaire

Mot	Nb Doc	FrqTotal	Ptr
Ambitious	2	6	1
Brutus	2	4	3
capitol	5	15	6

Posting simple

doc	Freq
doc1	3
doc2	2
doc1	1
doc3	7

- Liste triée
- B-Arbre
- Table de hashage (hash-code)
- ...

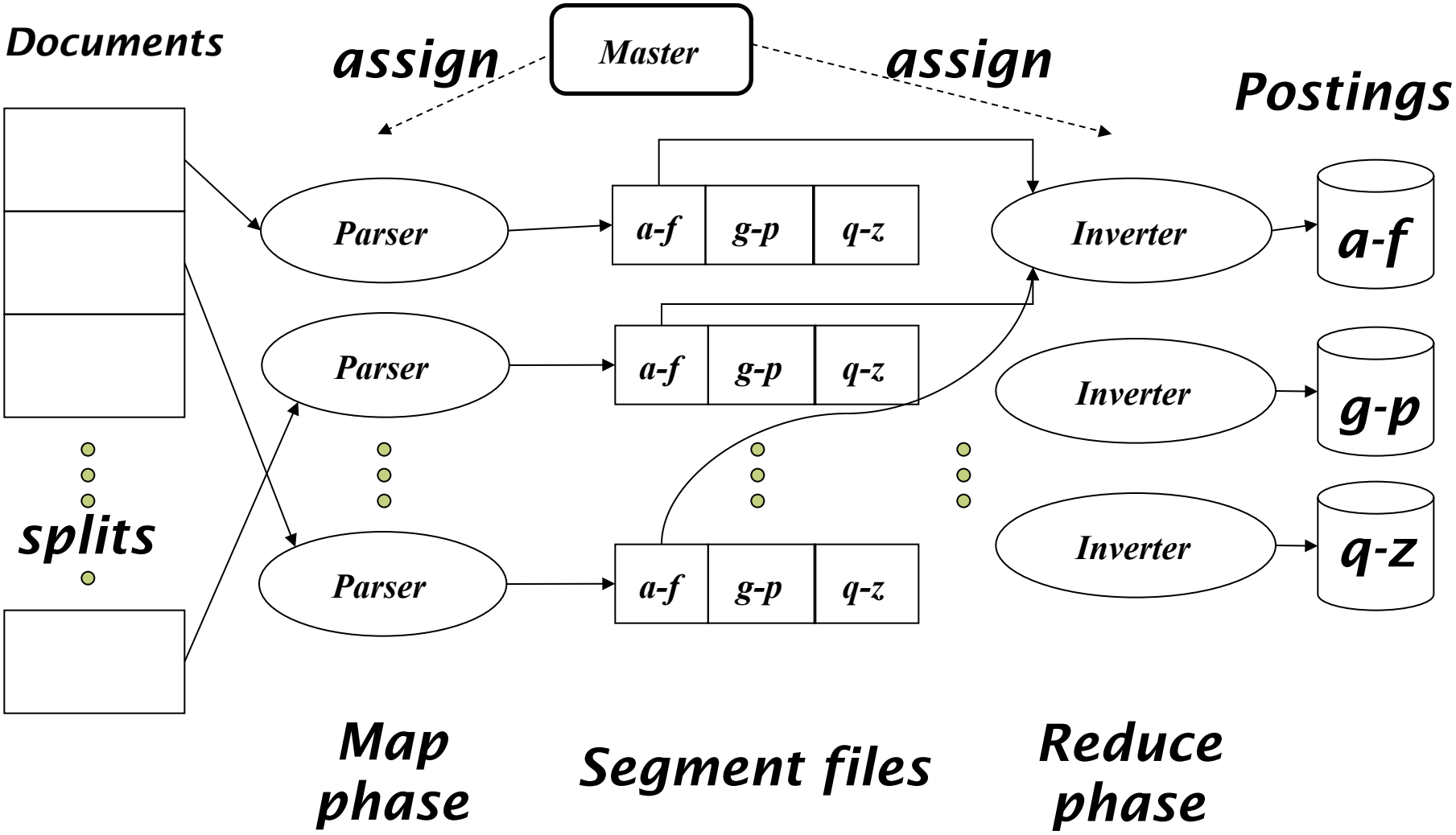
Posting riche

Position du terme dans le document (important pour la recherche d'expressions)

doc	Freq	position	balise
doc1	3	1, 4, 3	1, 5
doc2	2	1	
doc3	2	3	

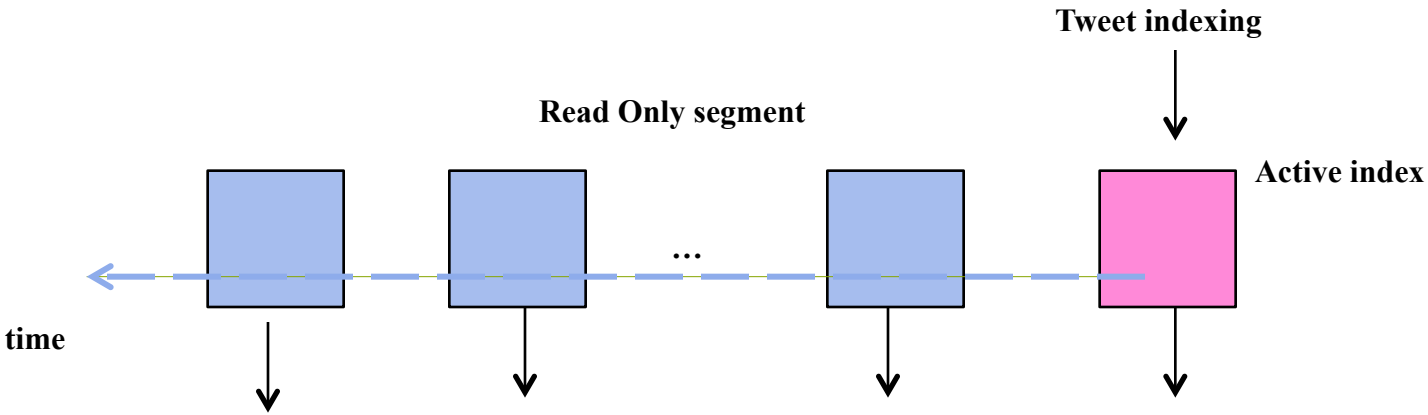
Balises (title, body, anchor, ..)

Construction de l'Index avec MapReduce

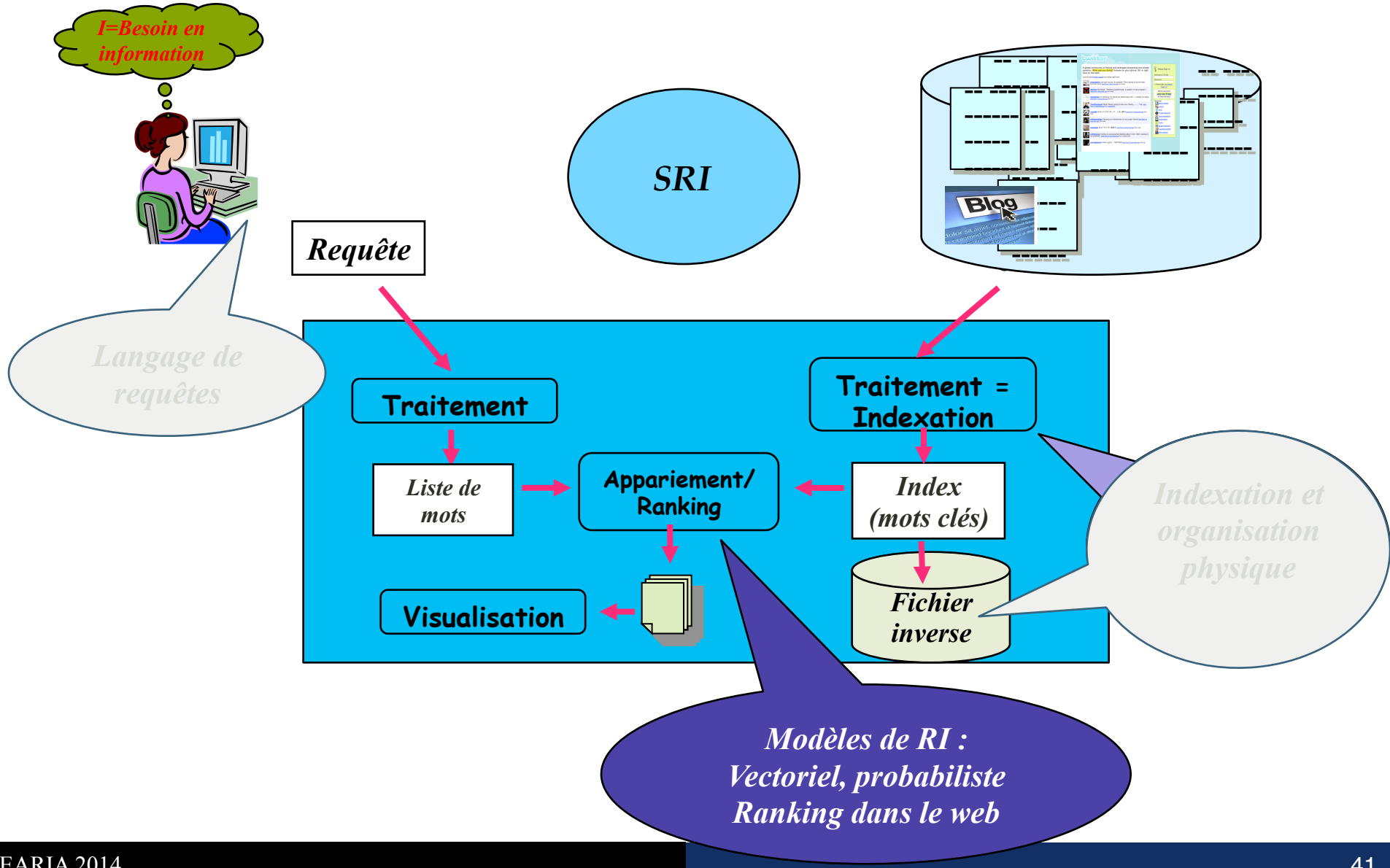


Indexation dynamique

- Real-time indexing
  - Ex. Indexation de tweets (Earlybird (Bush et al ICDE'12))
    - Low query latency (50ms)
    - Tweets are searchable within 10 seconds



Processus de RI



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.

Matching/appariement : modèles de RI

- 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)

Modèles de RI

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

Luhn's idea (1958): automatic indexing based on statistical analysis of text

Hans Peter Luhn (IBM)

"It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements." (Luhn 58)

LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development*, **1**, 309-317 (1957).
LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, **2**, 159-165 (1958).

Les statistiques au coeur de la RI

- tf (fréquence des termes) : La fréquence d'un terme est un indicateur de son importance

$$tf = \begin{cases} freq(t,d) \\ 1+\log(freq(t,d)) \\ \dfrac{freq(t,d)}{\max_{\forall t' \in d}(t',d)} \\ \dfrac{freq(t,d)}{\sum_{\forall t' \in d} freq(t',d)} \end{cases}$$

- IDF (Inverse Document Frequency) la fréquence du terme dans la collection

$$\log\left(\frac{N}{n_i}\right)$$

Le modèle vectoriel (VSM)

- Proposé par Salton dans le système SMART (Salton, G. 1970)
- Idée de base :
  - Représenter les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents :

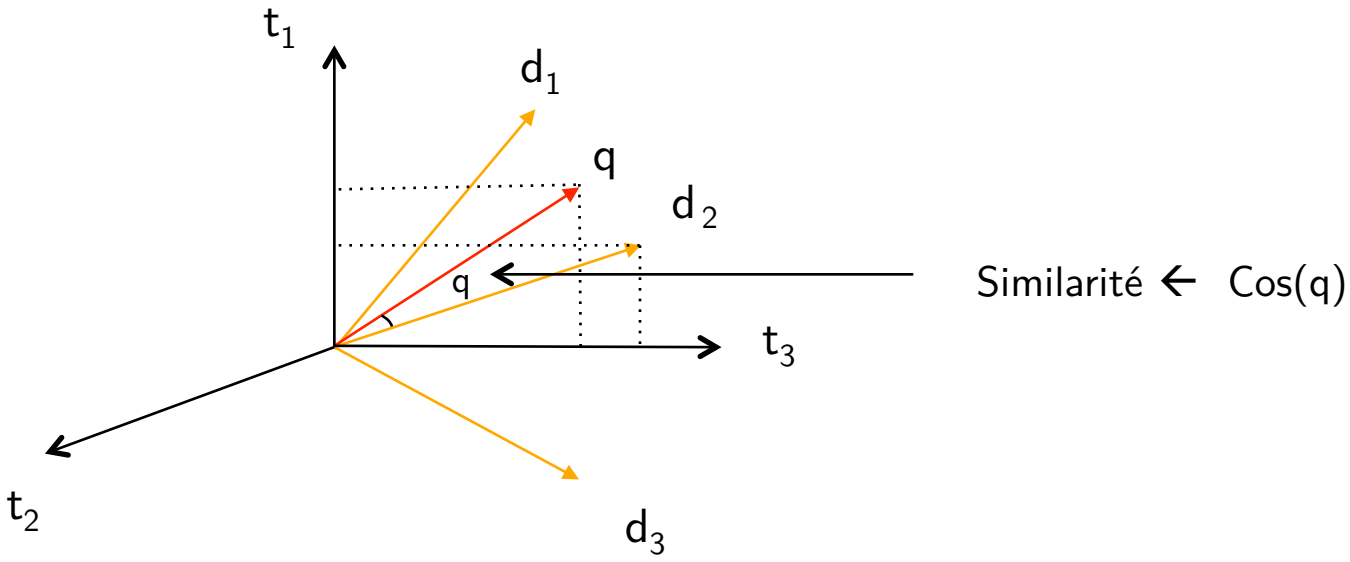
$$T<t_1,t_2, \ldots, t_M> \quad (\text{un terme} = \text{une dimension})$$

  - Document : $dj= (w_{1j}, w_{2j}, \ldots, w_{Mj})$
  - Requête : $q= (w_{1q}, w_{2q}, \ldots, w_{Mq})$

$w_{ij:}$ poids du terme t_i dans le document d_j → tf*idf

Le modèle vectoriel (VSM)

- Pertinence est traduite comme une similarité de vecteurs



La pertinence est traduite en une similarité vectorielle : un document est d1 est d'autant plus pertinent à une requête que le vecteur associé est similaire à celui de la requête.

Le modèle vectoriel (VSM)

Dot product

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i est le poids du terme t_i dans la requête
d_i est le poids du terme t_i dans le document

Le modèle probabiliste

- Le modèle probabiliste tente d'estimer la probabilité d'observer des événements liés au document et à la requête
- Plusieurs modèles probabilistes,
  - se différencient selon les événements qu'ils considèrent
    - P(R/d, q) : probabilité de pertinence (Relevance R) de d vis à vis de q
    - P(q,d)
    - P(q|d)
    - P(d|q)
  - Les distributions (lois) qu'ils utilisent

PRP : BIR et BM25

Modèle PRP (Probabilistic Ranking Principle)

$$RSV(q,d) \overset{rank}{=} \frac{P(R \mid d)}{P(NR \mid d)}$$

$$RSV(q,d) = \frac{P(d \mid R)}{P(d \mid NR)} * \frac{P(R)}{P(NR)} \overset{rank}{=} \frac{P(d \mid R)}{P(d \mid NR)}$$

$$\frac{P(d \mid R)}{P(d \mid NR)} = \frac{P(t_1 = x_1, t_2 = x_2, \ldots, t_n = x_n \mid R)}{P(t_1 = x_1, t_2 = x_2, \ldots, t_n = x_n \mid NR)} = \prod_{i=1}^{n} \frac{P(t_i = x_i \mid R)}{P(t_i = x_i \mid NR)}$$

Bernoulli

$$RSV^{BIR} = \sum \log \frac{\frac{r + 0.5}{R - r + 0.5}}{\frac{(n - r + 0.5)}{(N - n - R + r + 0.5)}}$$

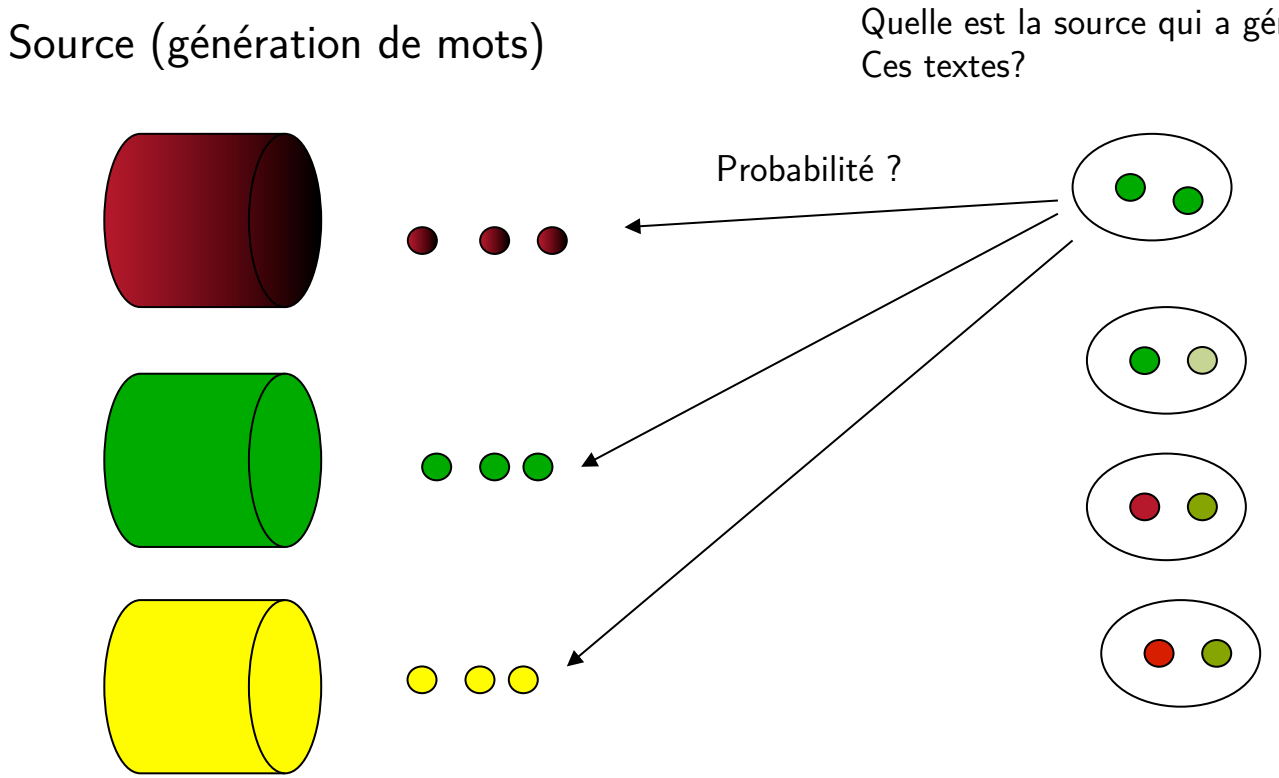
$$P(R \mid d) \overset{rank}{=} \prod_{i=1}^{n} \frac{P(t_i = x_i \mid R)}{P(t_i = x_i \mid NR)}$$

2-Poisson

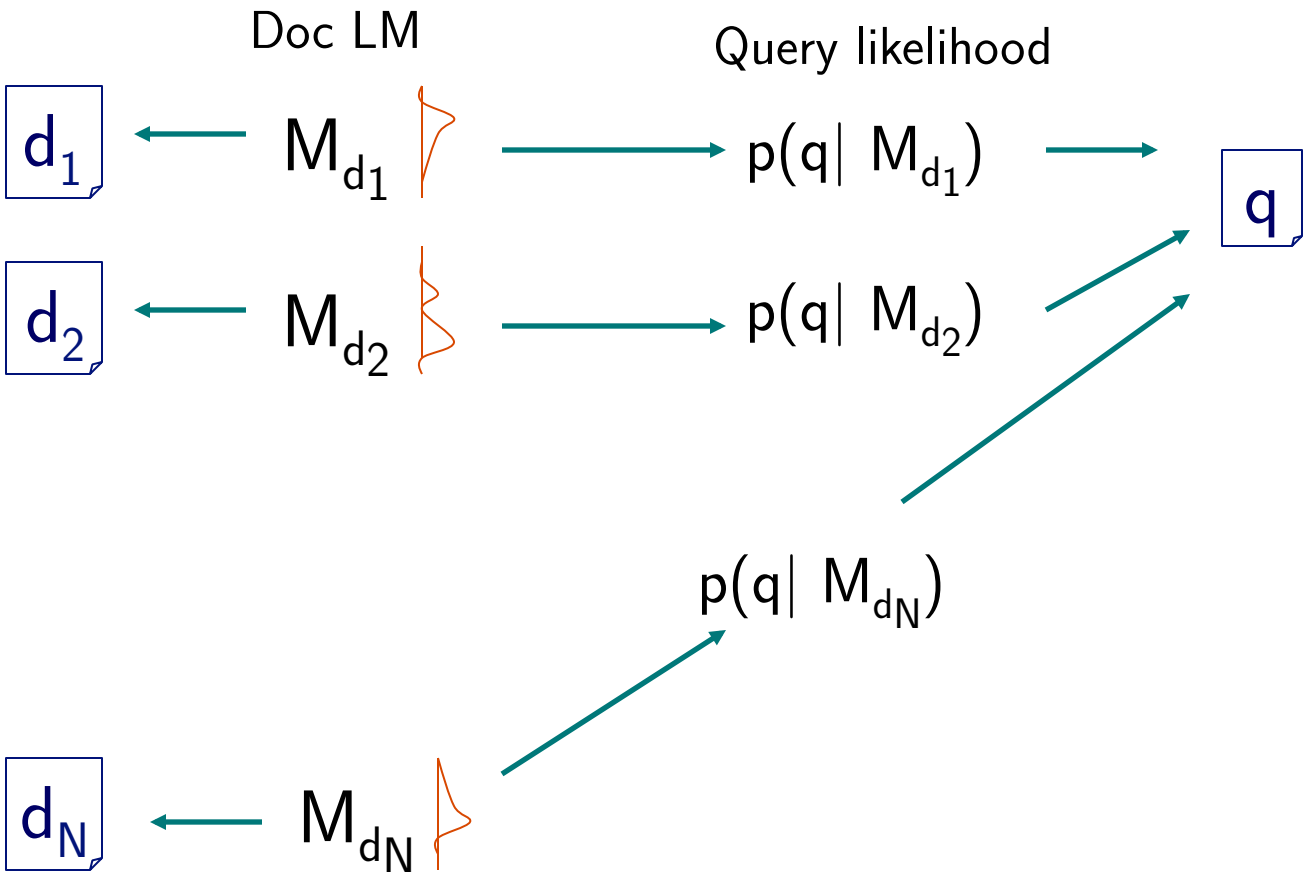
$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1) tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

Les modèles de langue

- Vu comme une source ou un générateur de textes
  - Mécanisme probabiliste de génération de texte (mots, séquence de mots)
    →On parle de modèle génératif



Les modèles de langue



Les modèles de langue

- Estimer de modèle de document (requête) la distribution des termes dans le document → Md

$$RSV(Q,D) = P(q / M_D) = P(t_1,t_2,...,t_n / D) = \prod_{t_i \in Q} P(t_i / D)$$

- Estimation du modèle de D
  - Maximum de Vraisemblance

$$P_{ml}(t_i \mid D) = \frac{tf_{(t_i,d)}}{|d|}$$

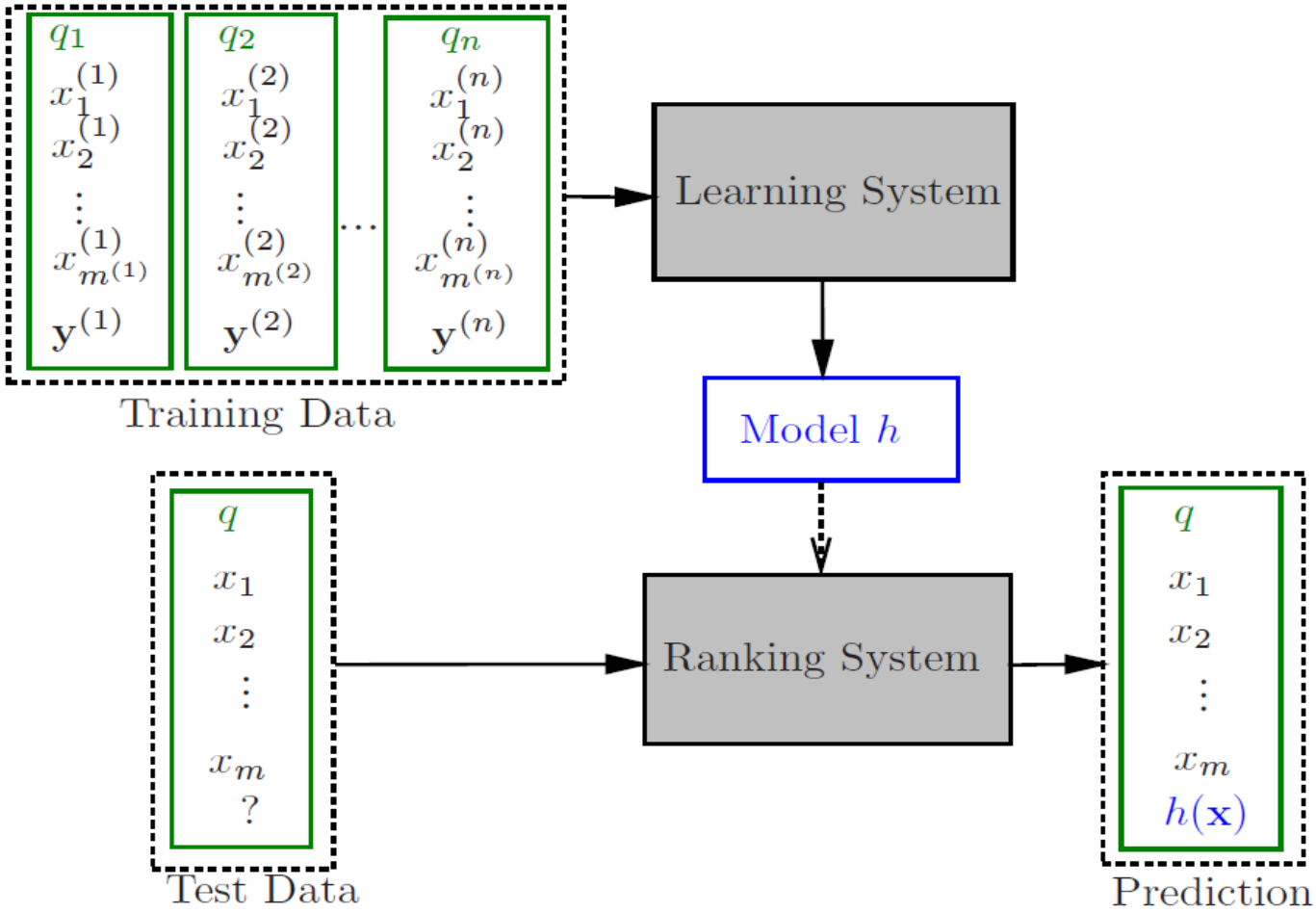
  - Modèle lissé (Dirichlet)

$$P_{Dir}(t \mid d) = \frac{|d|}{|d|+\mu} \times \frac{tf(t,d)+}{|d|} + \frac{|\mu|}{|d|+\mu} P_{ML}(t \mid C)$$

Apprendre à ordonnancer (LTR)

- Apprendre la fonction d'ordonnancement (ranking)
  - Collection d'entrainement
    - Input/: requête, doc (ou docs classés) → vecteur caractéristiques de q et d
    - Output : résultat voulu (pertinence, score..)
- Besoin d'une collection d'apprentissage
  - Collecte manuelle
  - Collecte → Automatique clickthrough data

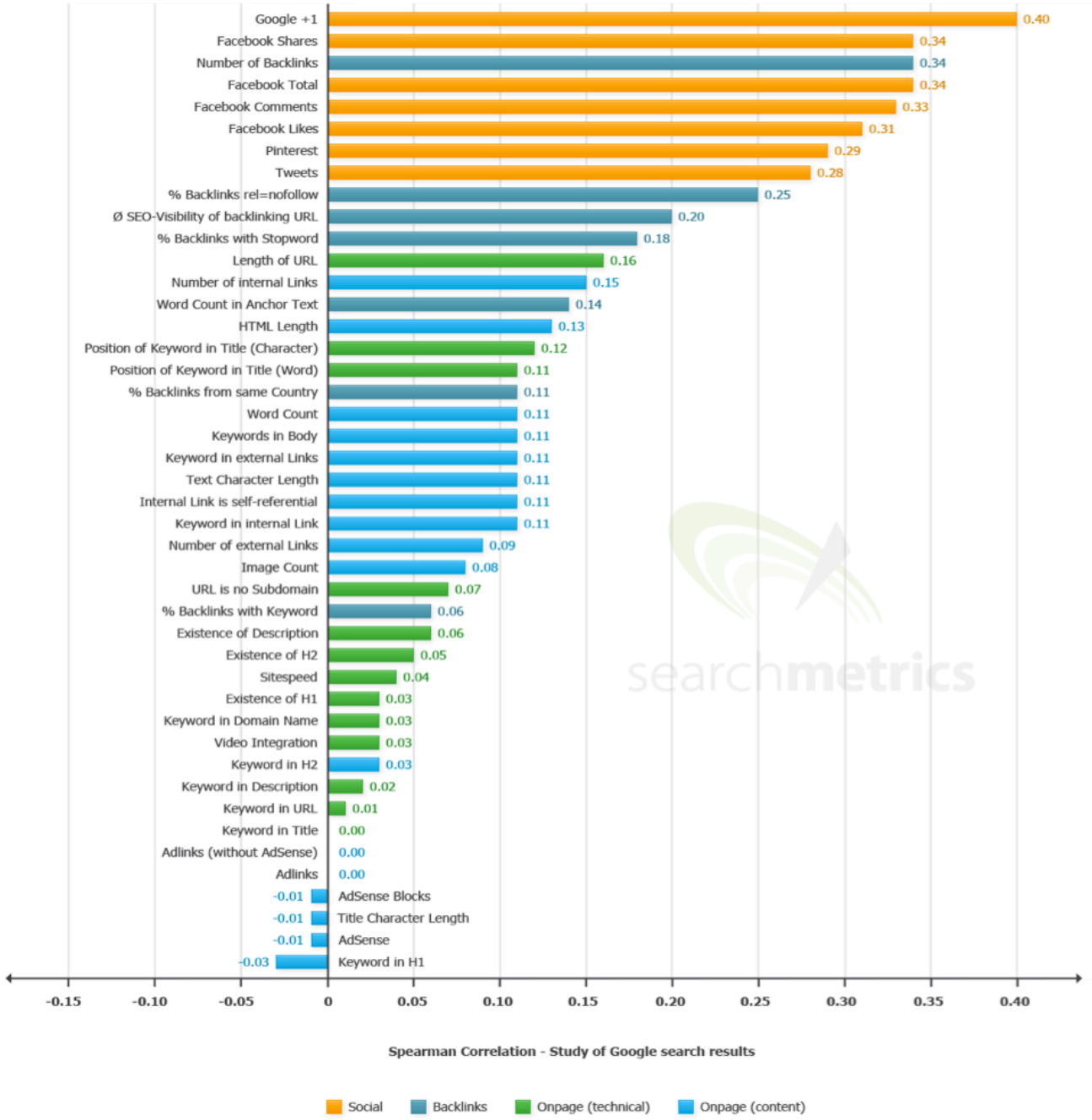
Learning to Rank Framework



Learning to Rank Algorithms

Least Square Retrieval Function (TOIS 1989)
Query refinement (WWW 2008)
SVM-MAP (SIGIR 2007)
Nested Ranker (SIGIR 2006)
ListNet (ICML 2007)
Pranking (NIPS 2002)
MPRank (ICML 2007)
LambdaRank (NIPS 2006)
Frank (SIGIR 2007)
MHR (SIGIR 2007)
RankBoost (JMLR 2003)
Learning to retrieval info (SCC 1995)
LDM (SIGIR 2005)
Large margin ranker (NIPS 2002)
Ranking SVM (ICANN 1999)
IRSVM (SIGIR 2006)
RankNet (ICML 2005)
Discriminative model for IR (SIGIR 2004)
SVM Structure (JMLR 2005)
OAP-BPM (ICML 2003)
Subset Ranking (COLT 2006)
GPRank (LR4IR 2007)
QBRank (NIPS 2007)
GBRank (SIGIR 2007)
Constraint Ordinal Regression (ICML 2005)
McRank (NIPS 2007)
SoftRank (LR4IR 2007)
AdaRank (SIGIR 2007)
CCA (SIGIR 2007)
ListMLE (ICML 2008)
RankCosine (IP&M 2007)
Supervised Rank Aggregation (WWW 2007)
Relational ranking (WWW 2008)
Learning to order things (NIPS 1998)

What factors are playing a role?



- Identifier les critères (Cleverdon 66)
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats
  - Capacité d'un système à sélectionner des documents pertinents.

Rappel : capacité d'un système à sélectionner tous les documents pertinents de la collection (R= Nb pert sel/total pert)
Précision : capacité d'un système à sélectionner que des documents pertinents (P= Nb pert sel / total selec.)

R-Précision, MAP, P@X, RR (Reciprocal Rank) NDGC, BPREF, E-mesure, Coverage, Novelty,

Démarche d'évaluation

- Démarche Analytique (formelle) :
  - Difficile pour les SRI, car plusieurs facteurs : pertinence, distribution des termes, etc. sont difficiles à formaliser mathématiquement
- Démarche Expérimentale (lab-based evaluation) (Cranfield Paradigm)
  - « benchmarking ».
  - Evaluation effectuée sur des collections de tests
  - Collection de test : un ensemble de documents, un ensemble de requêtes et des pertinences (réponses positives pour chaque requête)
- User studies evaluation
  - RI interactive, comportement de l'utilisateur

Campagnes d'évaluation

- TREC - Text REtrieval Conference
  - Évaluation des approches RI (beaucoup de tâches sont évaluées dans cette campagne)
- CLEF - Cross Language Evaluation Forum
  - Évaluation des approches de croisement de langues (multilinguisme)
- INEX - Initiative for the Evaluation of XML Retrieval
  - Évaluation de la RI sur des documents de type XML
- NTCIR- NII Testbeds and community for information access Research

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
- Conclusion

SIGIR 2015 topics

- **Document Representation and Content Analysis** (text representation, document structure, linguistic analysis, NLP for IR, cross- and multi-lingual IR, information extraction, sentiment analysis, clustering, classification, topic models, facets, text streams)
- **Queries and Query Analysis** (query intent, query suggestion and prediction, query representation and reformulation, query log analysis, conversational search and dialogue, spoken queries, summarization, question answering)
- **Retrieval Models and Ranking** (IR theory, language models, probabilistic retrieval models, learning to rank, combining searches, diversity and agregated search)
- **Search Engine Architectures and Scalability** (indexing, compression, distributed IR, P2P IR, mobile IR, cloud IR)
- **Users and Interactive IR** (e.g., user studies, user and task models, interaction analysis, session analysis, exploratory search, personalized search, context-based search, social and collaborative search, search interface, whole session support)
- **Filtering and Recommending** (content-based filtering, collaborative filtering, recommender systems)

SIGIR 2015 topics

- **Evaluation** (test collections, experimental design, effectiveness measures, session-based evaluation, simulation)
- **Web IR and Social Media Search** (link analysis, click models/behavioral modeling, social tagging, social network analysis, blog and microblog search, forum search, community-based QA, adversarial IR and spam, vertical and local search, expert finding)
- **IR and Structured Data** (XML search, ranking in databases, desktop search, entity search)
- **Multimedia IR** (e.g., image search, video search, speech/audio search, music search)
- Search applied to the Internet of Things (billions of devices, sensors, and actuators are now connected to the Web, which will affect how people search and browse the Web)
- **Other Applications** (e.g., digital libraries, enterprise search, genomics IR, legal IR, Medical search, patent search, text reuse, geographic IR, new retrieval problems)

Quelques thématiques de recherche

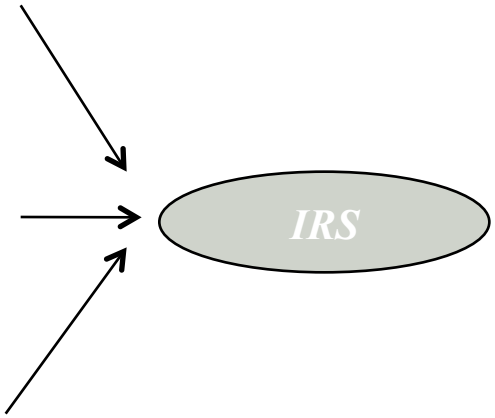
- Recherche d'information personnalisée/contextuelle
  - Prendre en compte le contexte de l'utilisateur (ses centres d'intérêts, sa tâche, sa géolocalisation,)
- Recherche d'information sociale
  - Recherche d'information dans des contenus sociaux (blog, microblog, ...)
  - Exploitation des contenus sociaux
- Recherche d'opinions
  - Identifier les pages de type opinion puis classer les opinions (+) (-)
- Recherche agrégée
  - Construire la réponse à la question en s'appuyant sur plusieurs sources

RI personnalisée/contextuelle

Java magazine

Java magazine

Java magazine



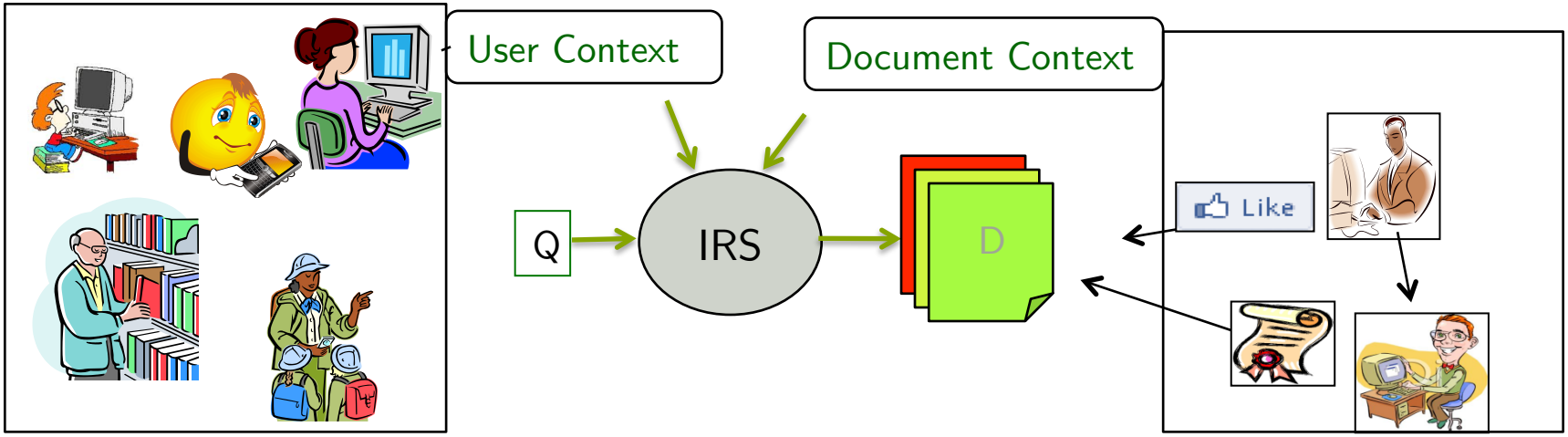
Ranked list of documents

Java SUN developers
Dev Journal: Top 10 Java guides
Java Coffee
...
Holidays in Java island

A same query by distinct users produces the same results

One size fits all (Lawrence 2000)

RI personnalisée/contextuelle



- Integrate the context in the IR process

A shift from query-centered IR to context-centered IR

→ information need = query +????

→ Relevance = a shift from topical relevance to contextual (situational) relevance

Social search paradigm

Definition

Social search : how social interactions (3) and social data (2) can enhance existing information-seeking experiences, as well as enable new information retrieval scenarios (1)

Three models

① Social data as new information to be searched
② Harnessing social data to augment search

③ Use social interaction and collaboration as part of the search process→ village paradigm

Social data as new content to be searched

Blog search

MicroBlog search

Expert search

Opinion search

Conversation search (forum …)

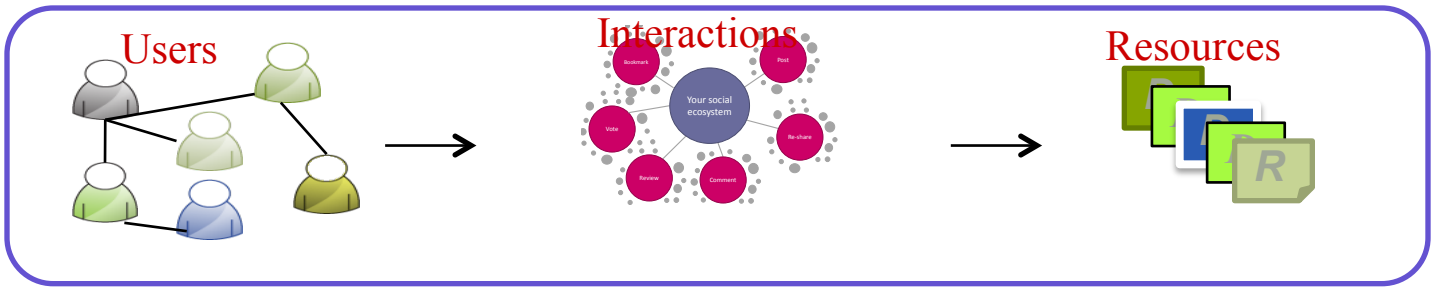
Social tagging

New research issues :
Information representation/indexing
Relevance model

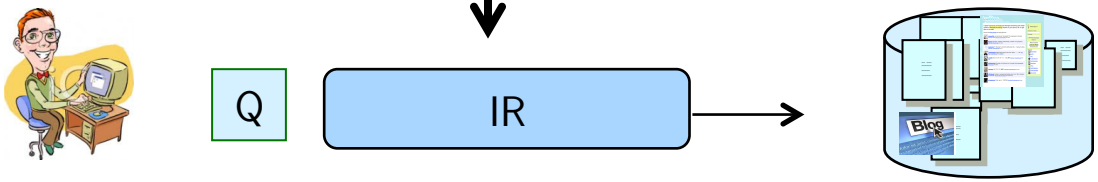
Microblog search

- Tweet characteristics (Recap.)
  - Shortness (composed of single sentence)
  - Specific syntax (RTre-tweets, #hashtags, @mentions, URL)
  - No context
  - Wide variety of topics
  - Net-speak language
  - Spams : the popularity of Twitter microblogging service makes him a target of spamming attacks.
  - Time sensitive
- Things we used to exploit in IR
  - Term frequency
  - Anchor text
  - Click throughdata
  - Field weight
  - Stable relevance judgement

Harnessing social data to augment search



$(U \times I \times R \times T)$



User side

Understand user intent (user profile)

Resource (document) side

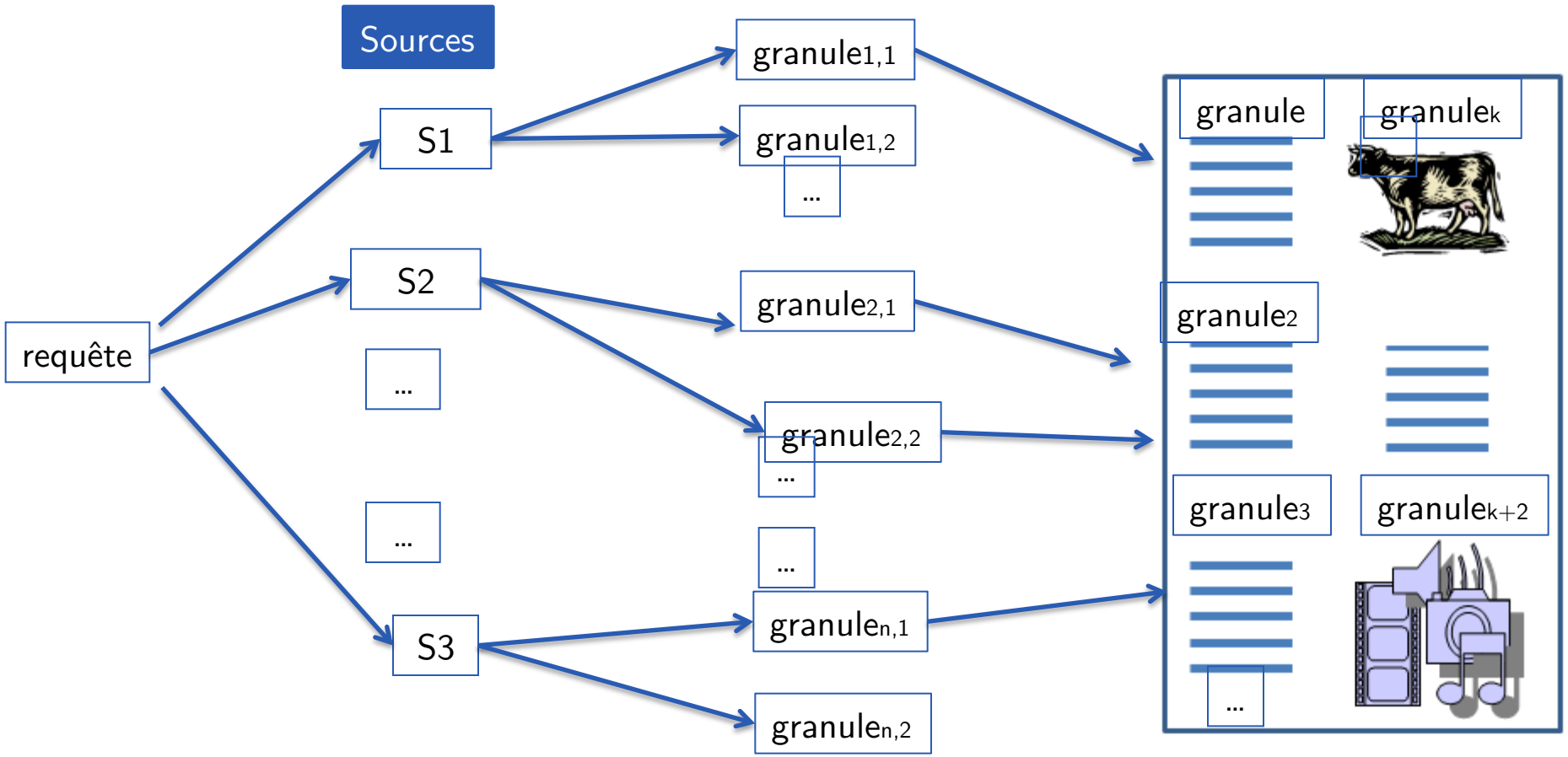
Better represent a (document)

Trust a document

Recherche agrégée (Aggregated search)

- Terminology
  - Federated search
  - Meta search
  - Data fusion
  - Data integration
  - Vertical search
  - Aggregated search
  - Content aggregation
  - Mutidocument summarization
  - Document generation
  - ...

Recherche agrégée (Aggregated search)



Recherche agrégée (Aggregated search)

Calendrier Abonnement

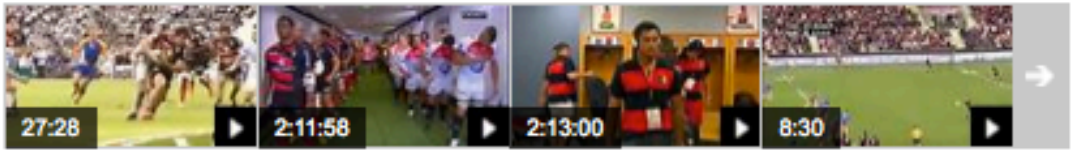
Page D'accueil Le Cop

Musée

Calendrier Stade Toulousain
Rugbyrama
Stade **Toulouse Transferts**
Stade **Français**

Stade Toulousain - Page d'accueil Translate this page
www.**stadetoulousain**.fr/index2.php
Saracens / **Stade Toulousain** - Interview de Maxime MÉDARD Election du stadiste de la saison. Le **Stade** dans les Médias . Suivre ...

Videos of **stade Toulousain**
bing.com/videos



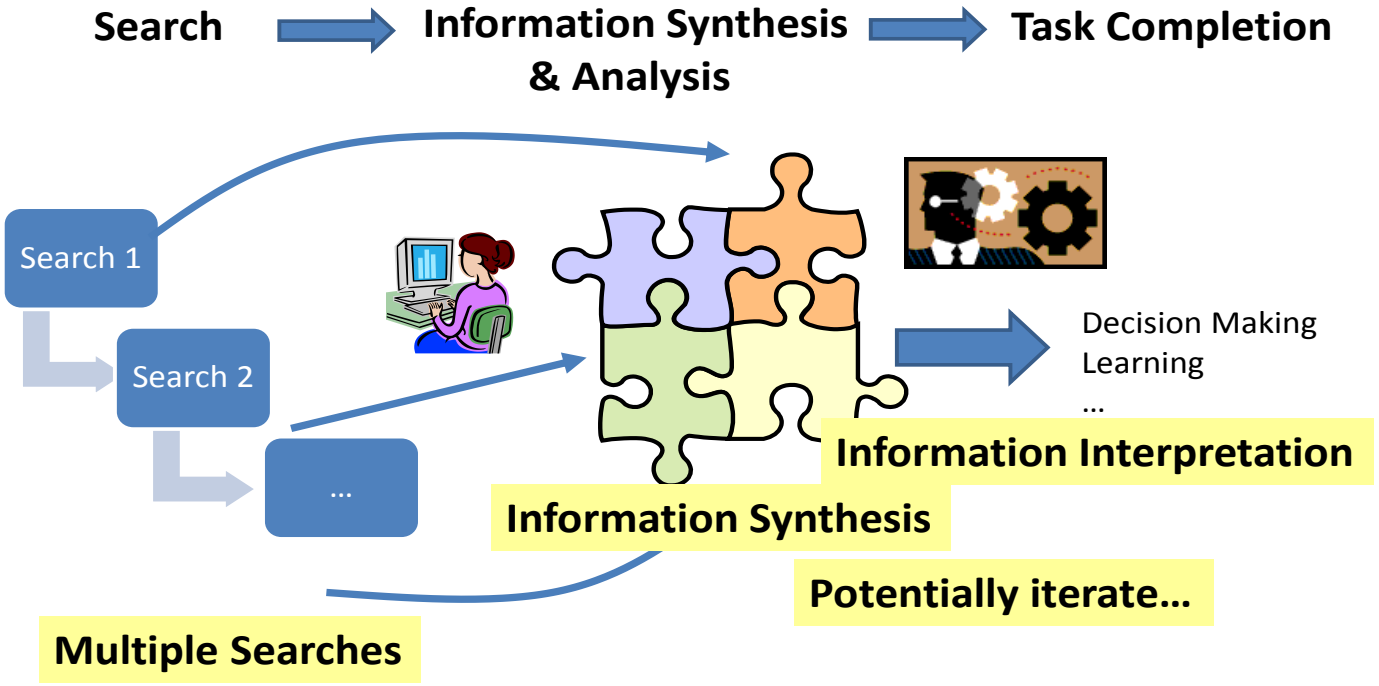
Compilation des essais du Stade ... YouTube
Stade Toulousain - RC Toulon [Final... YouTube
Stade Toulousain - Montpellier [Final... YouTube
stade toulousain compilation YouTube

Stade toulousain - Wikipédia Translate this page
fr.wikipedia.org/wiki/**Stade_toulousain**
Histoire · Palmarès · Les finales du **Stade** ... · Personnalités ...
Stade toulousain Généralités Fondation 1907 Statut professionnel depuis le 1 er février 1998 Couleurs rouge et noir **Stade Stade** Ernest-Wallon (19 500 places ...

stade toulousain fiche équipe - Rugby en... Translate this page

Recherche agrégée (Aggregated search)

- Task completion: Product purchases (compare customer Reviews, compare News articles, .., local activities (analysis of topic diffusion, ..), healthcare decisions
  - Tweet summarization de TREC



Plan

- Concepts de base de la Recherche d'information (RI)
  - Intérêt et contours
  - Tour d'horizon sur les techniques de RI
- Panorama RI - applications, scénarios
- Conclusions

Points de repère

- 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 de repère

- 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)
- EARIA 2006

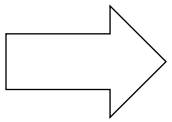
IR and Search Engines

Information Retrieval

Relevance
- -Effective ranking

Evaluation
- -Testing and measuring

Information needs
- -User interaction



Search Engines

Performance
- -Efficient search and indexing

Incorporating new data
- -Coverage and freshness

Scalability
- -Growing with data and users

Adaptability
- -Tuning for applications

Specific problems
- -e.g. Spam

États des lieux

- Plusieurs avancées :
  - Techniques de pondération (Luhn, salton, ... BM25, modèle de langue, théorie de l'info)
  - Plusieurs modèles
  - Reformulation de requêtes (relevance feedback, expansion ...)
  - Évaluation des performances (métriques, benchmark)
  - Architecture physique (inverted file)
  - ...

Hans-Peter Luhn

Gerard Salton

Constat : les techniques statistiques sont effectives en RI

États des lieux

- Les modèles statistiques (BOW) s'appuient sur deux hypothèses
  - Les termes importants sont redondants → LN est redondant
  - La cooccurrence entre les mots indique le sujet du document
  - L'analyse de textes en surface est capable de capturer ces phénomènes
  - Les statistiques « savent » mesurer ces phénomènes

États des lieux

- Plusieurs programmes d'évaluation :
  - TREC (Text REtrieval Conference)- DARPA,NIST
  - 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 tâches

  → Comparaisons riches

États des lieux

- Systèmes de RI open source
  - Smart (Cornell)
  - MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
  - Lemur (CMU/Univ. of Massachusetts)
  - Lucene (Nutch)
  - Terrier (Univ Glasgow)

- Permettent d'indexer et d'interroger différents types documents textuels (texte libre, html, xml, pdf, ...). Plusieurs modèles de RI sont programmés (vectoriel, probabiliste (BM25), modèle de langue)

Concusion

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

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

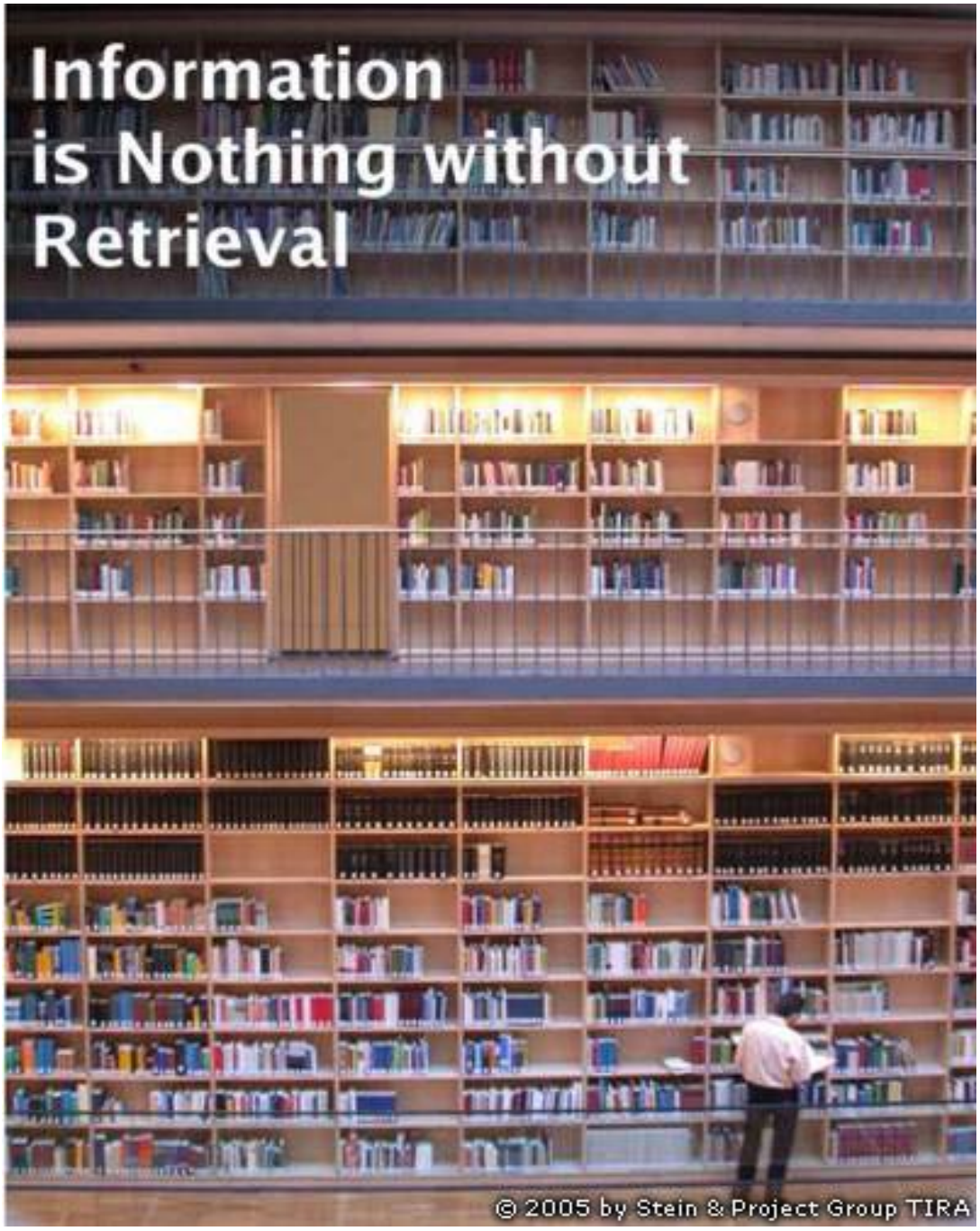
Ref. bibliographiques

- 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/

Merci


Information
is Nothing without
Retrieval
© 2005 by Stein & Project Group TIRA