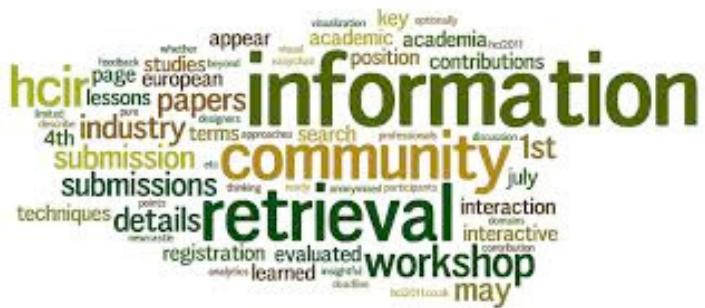




# Cours de Recherche d'Information

# 2<sup>ème</sup> année Ingénieurs en Informatique et Multimédia



# Chiraz Latiri



## Autour de trois parties :

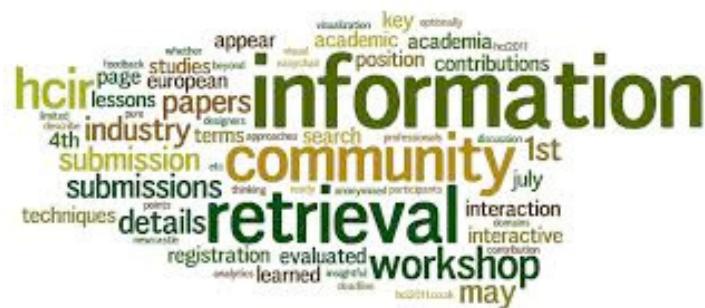
- I. Introduction à la RI
- II. Fondements de la RI
- III. RI multimédia
- IV. RI Sociale

**Chiraz Latiri**

[Chiraz.latiri@gnet.tn](mailto:Chiraz.latiri@gnet.tn)



# Partie I : Introduction à la RI



# Chiraz Latiri

# Recherche d'Information (RI) / Information Retrieval (IR)



# *Recherche d'Information*

Ensemble des méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la **sélection d'information pertinente pour un utilisateur**

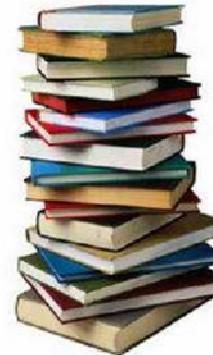




# *Les acteurs de la Recherche d'Information*

**Collection :**

un ensemble de documents



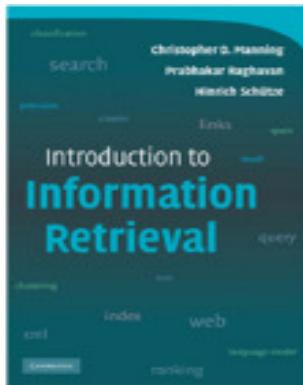
**Utilisateur :**

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

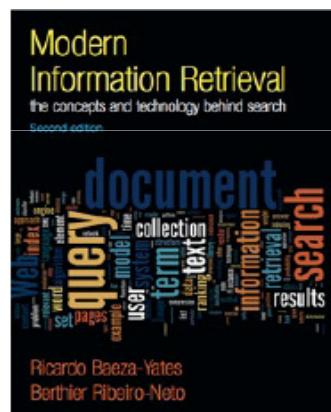


**Système de RI :** l'outil qui doit retrouver les documents **pertinents** pour le besoin de l'utilisateur

# *RI: définitions*



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.



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

# *RI: définitions*

- 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.**

# RI : une requête classique



# *RI: qu'est ce qu'on cherche?*

- *Où se trouve la librairie la plus proche de chez moi ?*
- *Qui est actuellement en tête du Top 14 de rugby ?*
- *Quels sont les titres mentionnés à la une du journal Le Monde d'aujourd'hui ?*
- *Que rapporte la une du Monde d'aujourd'hui sur les candidats à l'élection présidentielle ?*
- *Quels sont les films qui passent ce soir sur la TNT ?*
- *Dans quels films Jean Rochefort et Philippe Noiret ont-ils joué ensemble ?*
- *Quels sont les logiciels d'installation de logiciels sous Linux/Debian ?*
- *Comment peut-on installer des logiciels sous Linux/Debian ?*
- *Quelle est la traduction du mot anglais "ice" en français ?*
- *Qui était Claude Bernard ?*

## Questions

- Quelle est la nature des résultats attendus ?
- Comment évalue-t-on la pertinence des résultats ?
- Sous quelle forme doit-on formuler ses requêtes ?
- ...

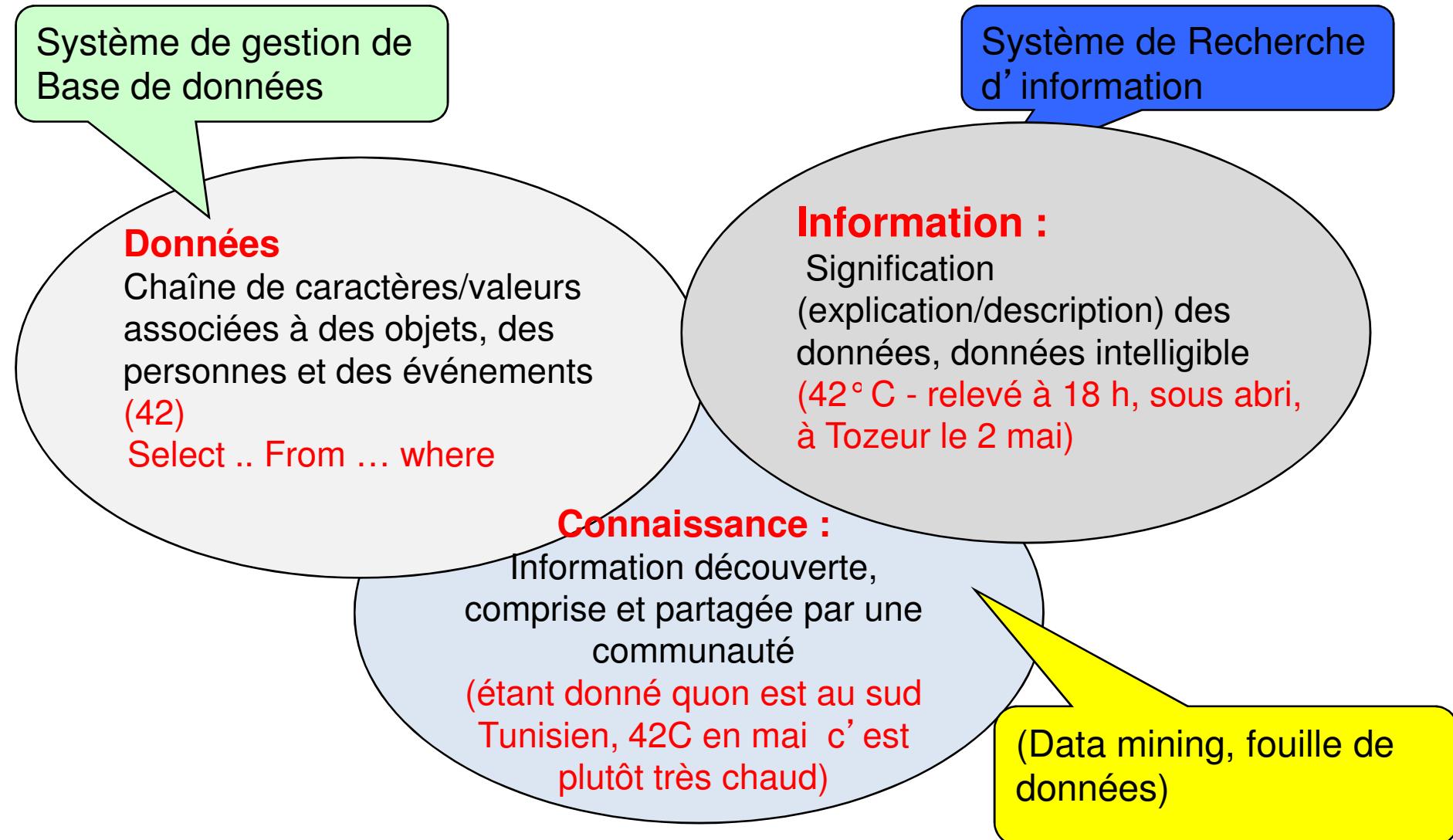
# *Information vs. données*

- Les **données** sont reçues, stockées et retrouvées par un endosystème. Les données sont impersonnelles ; elles sont disponibles pour tout utilisateur du système.
- L'**information**, en revanche, est un ensemble de données qui correspond à un besoin particulier. Le concept d'information a des composantes personnelles et temporelles absentes du concept de donnée.

*(R. R. Korfhage, 1997)*

- La **connaissance** est la sélection, l'appropriation et l'interprétation des informations, souvent implicite et utile pour l'utilisateur.

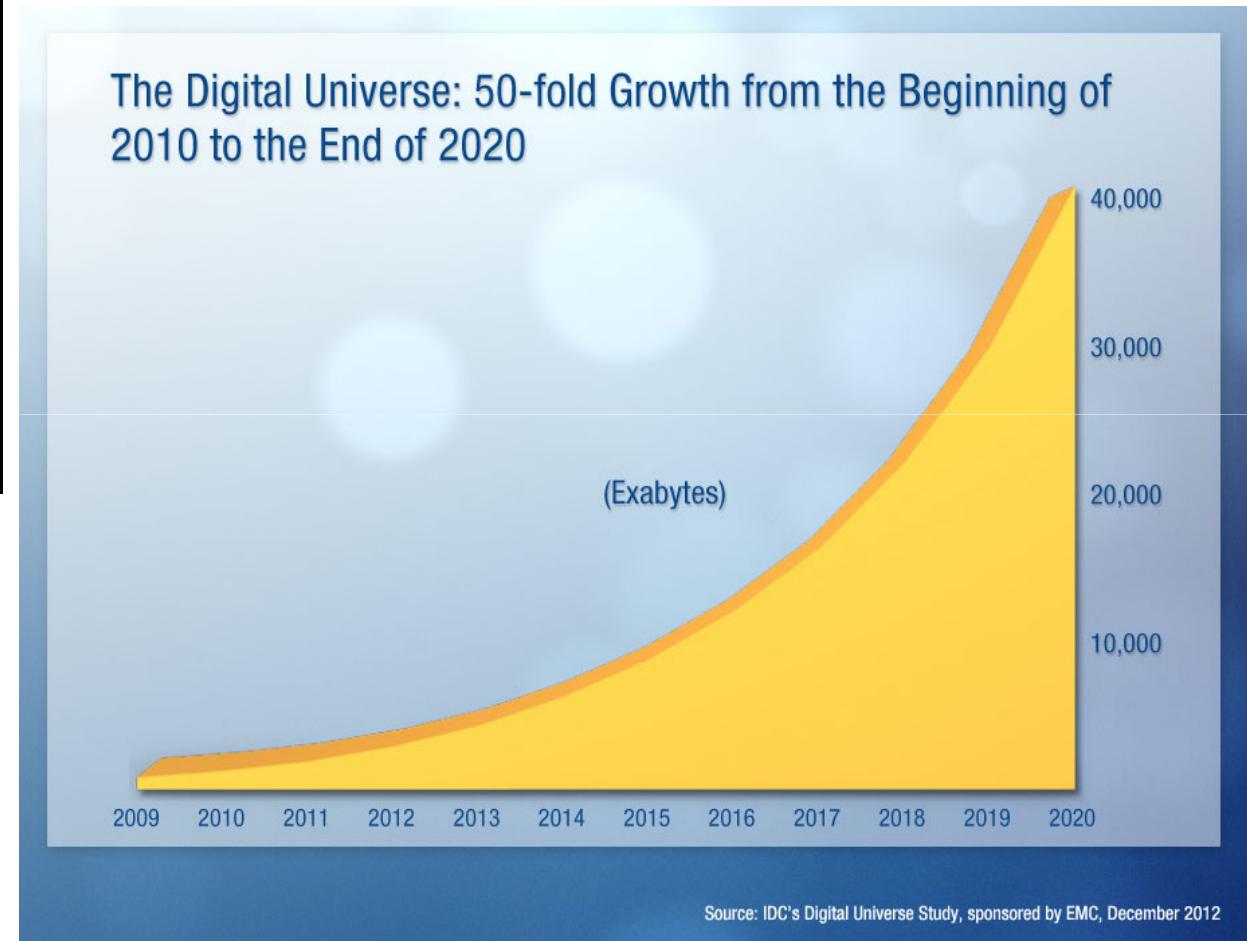
# *Données-Information-Connaissances*



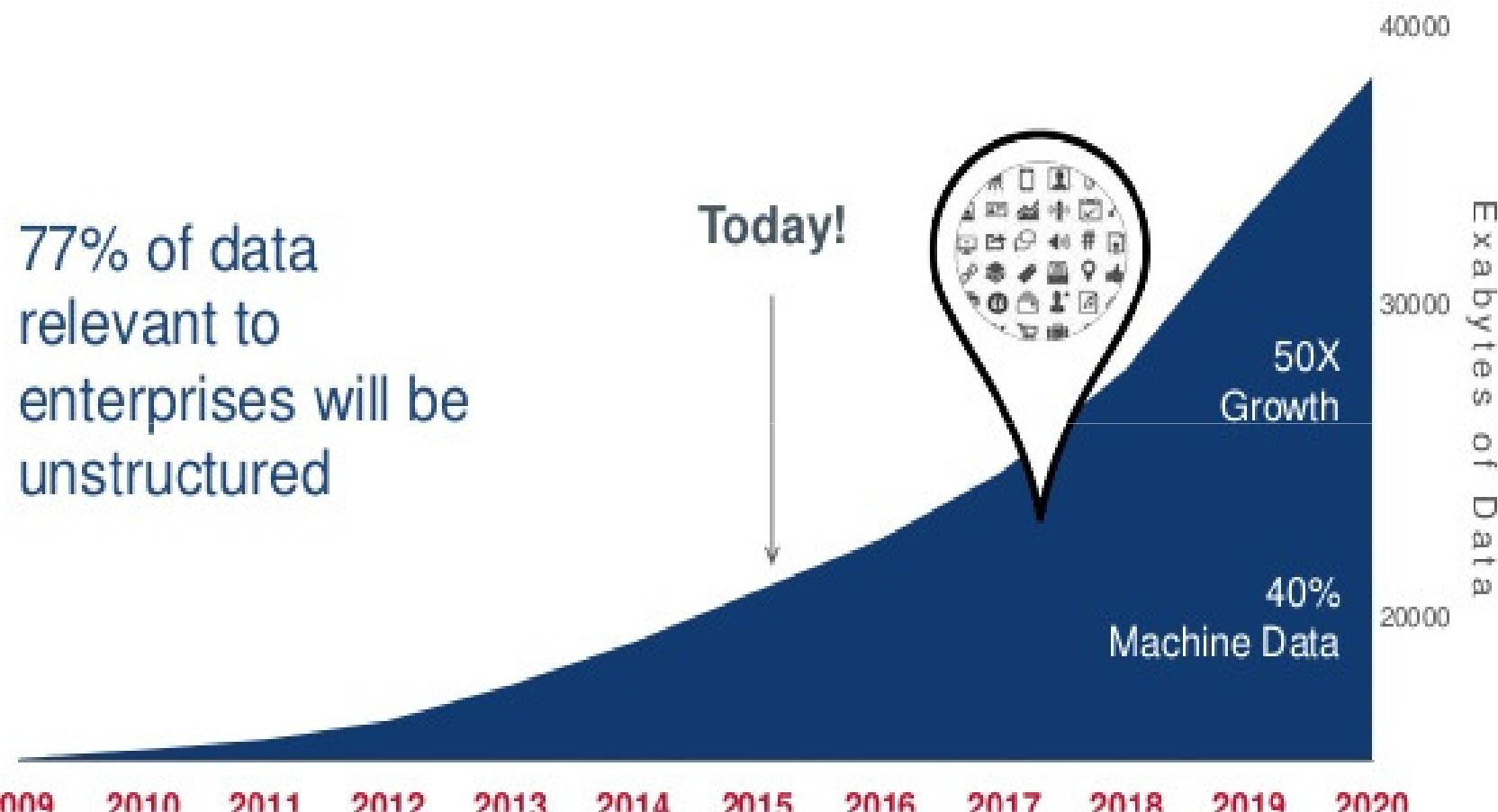
# *L'information est partout sous de gros volumes*

<i>KiloOctets</i>	$10^3$
<i>MegaOctets</i>	$10^6$
<i>GigaOctets</i>	$10^9$
<i>TeraOctets</i>	$10^{12}$
<i>PetaOctets</i>	$10^{15}$
<i>ExaOctets</i>	$10^{18}$
<i>ZettaOctets</i>	$10^{21}$

Facteur de 10 en 5 ans!



# *Big data...*



3

Source: IDC's Digital Universe Study, sponsored by EMC, April 2014

© 2015, Pentaho. All rights reserved. pentaho.com Worldwide +1 (866) 660-7555

# Informations hétérogènes

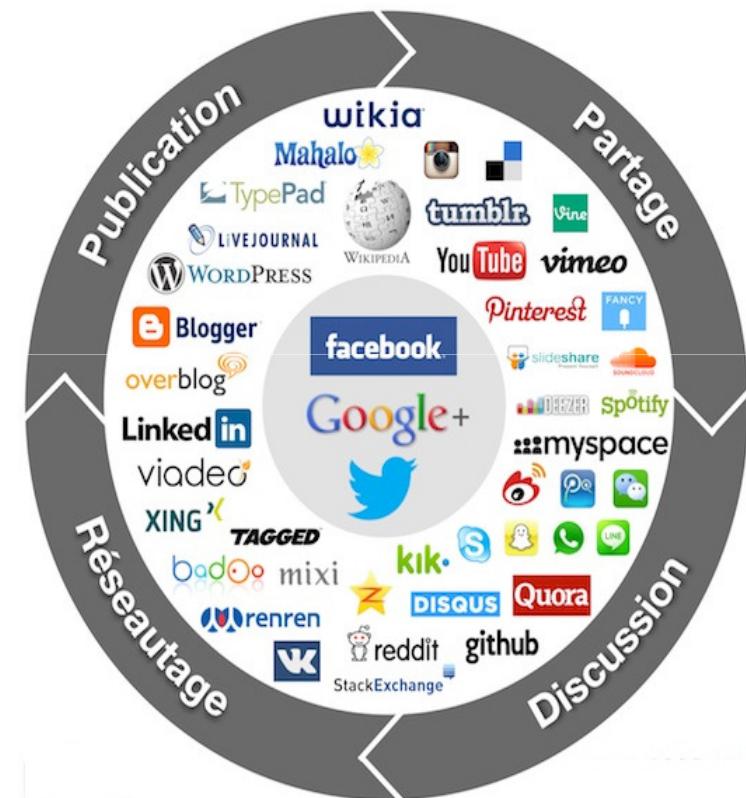
Origines de l'information numérique....diverses et nombreuses

## Twitter (2015)

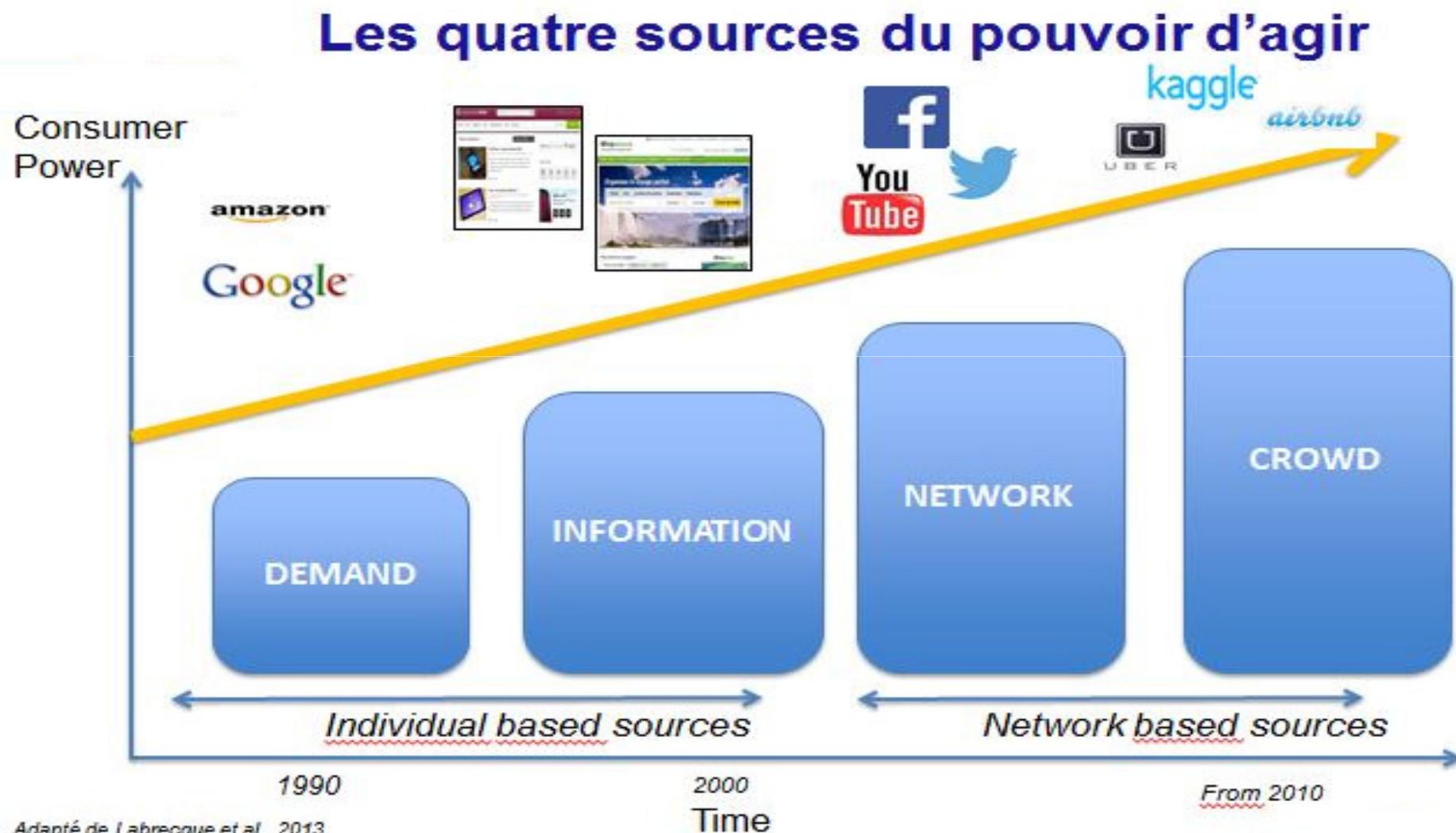
- Les tweets contenant une photo sont deux fois plus partagés que la moyenne.
- 500 millions de tweets sont envoyés chaque jour.
- 300 milliards de tweets ont été envoyés depuis le 21 mars 2006.

## Facebook (2015)

- Facebook (2015) Statuts partagés chaque seconde : 4100
- Likes distribués chaque minute : 1,8 million
- Likes distribués chaque jour : 4,5 milliards
- Contenus partagés chaque jour : 4,75 milliards
- Messages envoyés chaque jour : 10 milliards
- Données échangées chaque minute : 350 gigaoctets



*...et multi-sources*



Adapté de Labrecque et al., 2013

## *Premier Bilan*

Le problème n'est pas la disponibilité de l'information (**Bases de données, Fichiers annotés, Corpus de textes, Pages Web, Vidéos annotées, images,....**)



Sa sélection, son identification



Arriver à trouver au bon moment l'information utile et pertinente

# Diversité des besoins d'information (1/2)

- Recherche d'un **élément connu**
  - L'utilisateur sait exactement quels éléments il recherche.  
Il sait reconnaître les éléments désirés s'il les voit.
  - *Exemple* : recherche d'une citation bibliographique précise (Livre de Eric Gaussier sur la RI).
- Recherche d'une **information générale**
  - L'utilisateur recherche une information sur un sujet en général. Il existe de nombreuses façons de décrire le sujet.
  - Il est possible que l'information pertinente ne soit pas reconnue
  - Cette information peut ne pas satisfaire l'utilisateur que de façon partielle.
  - *Exemple* : Les réformes de l'enseignement secondaire en Tunisie

# *Diversité des besoins d'information (2/2)*

- Recherche d'une **information précise**
    - L'utilisateur recherche une information spécifique mais ignore sous quelle forme elle se présente.
    - Réponse partielle impossible
    - *Exemple : À quelle heure Chokri Belaid a-t-il été assassiné ?*
  - **Exploration**
    - Le but n'est pas de répondre à une question en particulier, mais de parcourir l'ensemble des données pour découvrir quels types d'informations concernant un sujet ou un domaine sont présents.
- **Navigation**

# *Diversité des problèmes*

- Difficultés d'**accès, couverture, temps de traitement**
  - Les sources d'information sont **très grandes**, réparties sur de **nombreux sites** dans des **localisations différentes**.
- Difficultés de définition de la **pertinence**
  - Comment un document remplit-il le **besoin informationnel** d'une personne donnée ?
  - Quelle est sa **pertinence** ? Comment la mesure-t-on ?
- Difficulté d'**exploitation**
  - Les documents pertinents ne sont pas nécessairement dans la **langue** de la requête.
  - L'information recherchée n'est pas nécessairement clairement identifiable dans un document.

# *Verrous de la RI*

- 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

# *Diversité des Tâches de RI (1/4)*

- **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)

## *Diversité des Tâches de RI (2/4)*

- Classification-Catégorisation/Clustering (partitionnement)
  - 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 ? »

## *Diversité des Tâches de RI (3/4)*

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

## *Diversité des Tâches de RI (4/4)*

- 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

# *Recherche d'information sur le Web*

- Sur Internet : utilisation massive par des **utilisateurs non experts**
  - Domaine d'une importance économique majeure
  - La requête typique est constituée d'au plus quelques mots clés
  - Les utilisateurs s'adaptent aux outils
- Une partie du web n'est pas directement **accessible** (web invisible, dont pages à accès restreint et pages dynamiques)
- L'information présente est fortement **multilingue** : les documents répondant aux requêtes peuvent être dans des langues différentes
- L'information présente n'est pas toujours **fiable**
- La **visualisation** de l'information est particulièrement importante : classement des résultats, présentation d'extraits, extraction de segments pertinents, etc.

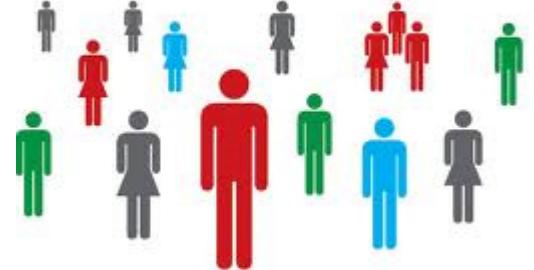
# Moteurs de recherche



- Distinguer d'abord :
  - **Outils propres au web** : moteurs de recherche, métamoteurs, moteurs de blogs...(Google, yahoo les plus populaires).
  - **Outils accessibles par le web** : bases de données, catalogues...
- Deux critères essentiels :
  - **Offre des ressources** : outil généraliste / spécialisé
  - **Mode d'indexation** : outil humain / automatisé
- **Moteurs généralistes** : *Google, Yahoo, Ask, Bing...*
- **Moteurs spécialisés** : géographique : [Google.tn](#), selon le contenu des ressources indexées ([Google Scholar](#)), presse ([Google News](#)), disciplinaire [SciencesDirect](#) en Sciences exactes, par domaine (PubMed), forums ([Google Groups](#)), listes de diffusion (info-IC) ; blogs , selon les supports : images, vidéos (Google ou Yahoo), fichiers son.

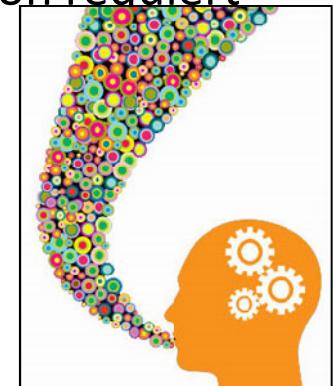
# *Les difficultés de la RI : le facteur humain*

- Le besoin d'information de l'utilisateur est parfois **vague** et toujours **subjectif**.
  - La **perte d'information** entre la réalité du besoin d'information et son expression peut être importante.
  - La pertinence d'un document pour une requête est une notion variable et très complexe à définir.
- Il ne peut pas exister de système de recherche d'information parfait
- L'évaluation d'un système dépasse les aspects habituels de performance informatique
- L'humain est subjectif, versatile, et il utilise un **langage "naturel"** !



# *Les difficultés de la RI : le facteur "langage"*

- À la différence des langages artificiels, le **langage "naturel"** est :
    - **Implicite** : tout n'est pas dit dans les textes et leur compréhension requiert une importante connaissance sur le contexte et sur le monde
    - **Redondant** : la langue offre de nombreuses façons de formuler le même contenu
    - **Ambigu** : un même énoncé peut souvent être interprété de différentes façons
  - La recherche d'information est encore compliquée par le fait que :
    - Les mots peuvent jouer des rôles différents dans les textes
    - Les atomes de sens peuvent être des mots ou des groupes de mots (termes)
- Il est compliqué de **formuler son besoin d'information**  
(perte d'information entre besoin et requête)



# *Caractère implicite de la langue*

- Connaissance du langage et des **conventions langagières**

*Q : Le voisin est-il chez lui ?*

*R : Sa voiture est devant le portail*

(implicature conversationnelle)

- Connaissance du **contexte**

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

(Sport ? Courrier ? Accident ?)

- Connaissance du **monde**

*La Nouvelle-Zélande va tailler la France en pièces.*

(métonymie + langage figuré + actualité du rugby)

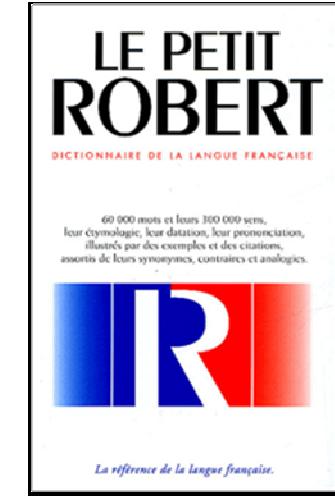
- Déduction** (présupposition)

*Ravaillac a assassiné Henri IV en 1610.*

⇒ Henri IV est mort en 1610.

# Caractère redondant de la langue

- Au niveau lexical
  - **Synonymie** : vélo et bicyclette
  - **Hyperonymie** et **hyponymie** : véhicule / vélo / VTT
  - **Méronymie** et **holonymie** : pédale / pédalier / vélo
- Abréviations et sigles
  - *S'il-vous-plaît* et SVP, VTT et Vélo Tout Terrain
- Entre mots et expressions
  - **Périphrases** : lave-vaisselle et machine à laver la vaisselle
  - **Définitions** : selle et petit siège, le plus souvent de cuir, d'un cycle ou d'un véhicule à deux roues à moteur
- Glissements de sens (synonymie contextuelle)
  - Il a écrit un **papier/article** sur la recherche d'information



# Caractère ambigu de la langue

- Les **homonymes** sont des mots qui ont une même graphie mais des sens différents

Article Discussion Lire Modifier Afficher l'histo

## Noyau

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

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

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

# *Mots composés*

- Les **mots composés** sont beaucoup moins polysémiques
- Les rechercher ensemble dans les textes est bénéfique (mais compliqué)
- Ils ont un sens qui n'est pas la composition des sens des atomes
  - *Homme-grenouille*
  - *Pomme de terre*
  - *Traitements de texte*

© M. Heinrich, J. Negra

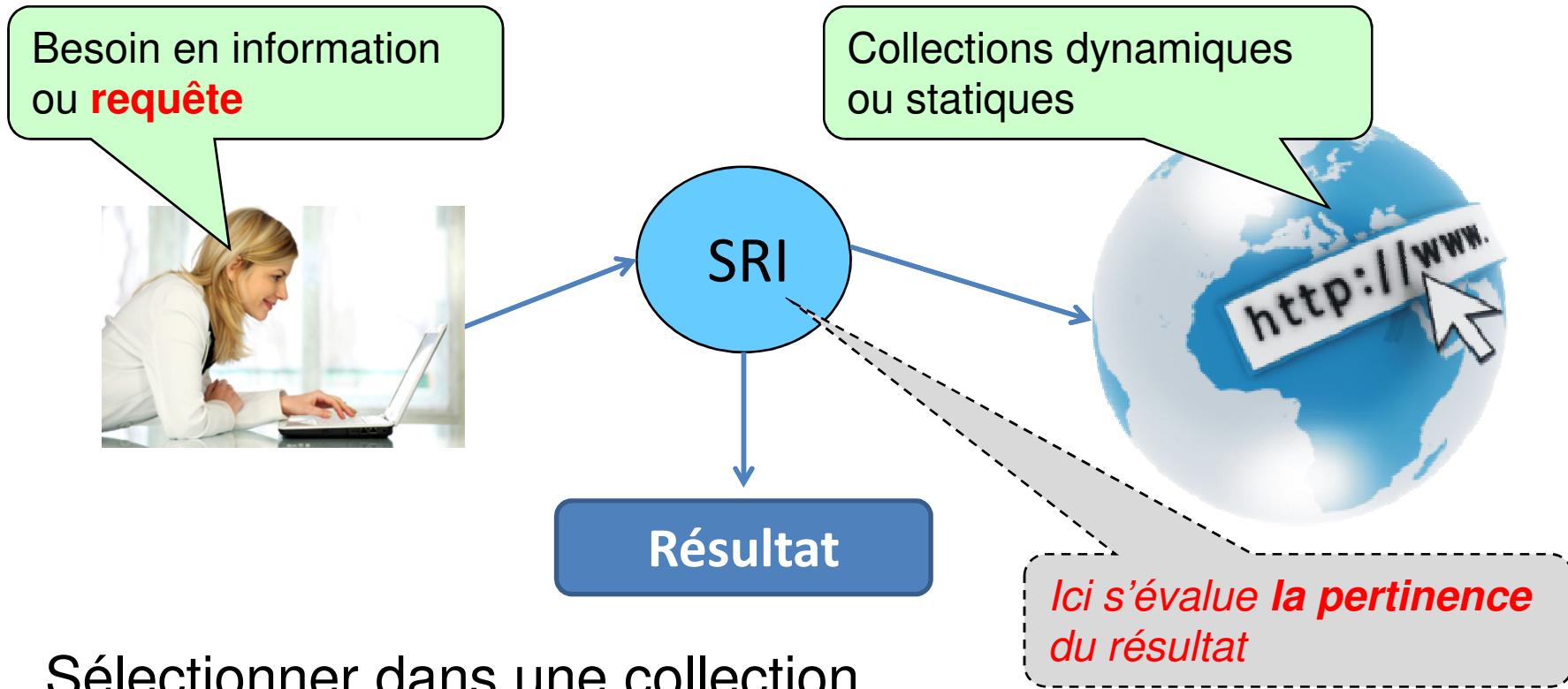


# Partie II : Fondements de la RI



# Chiraz Latiri

# Problématique de la RI



## Sélectionner dans une collection

- Résultats: les informations ( documents, images, vidéos...)
- ... pertinentes répondant aux
- ... besoins en information des utilisateurs (Requêtes)

# Besoin en information ou requête

# *Besoin en information?...Requête (1/2)*

- 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 - transmédia)
  - Social (réseaux sociaux, microblogs)

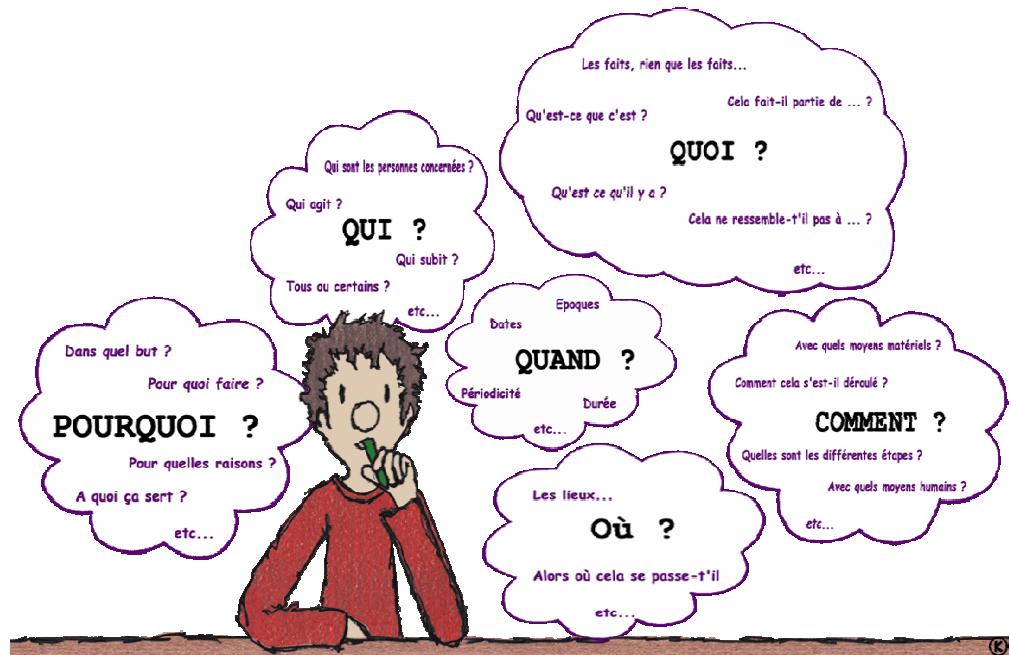


- Comprendre le contenu vs. l'interpréter → Ambiguïté du langage naturel (polysémie, synonymie, ...)
- Définir la granularité respectivement au besoin (document, paragraphe, phrase, titre, unité linguistique...)

# *Besoin en information?...Requête (2/2)*

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

Comment capturer le besoin de l'utilisateur ?



# *Besoin d'information : Intention de la requête*

- Besoin = requête
- Exemple : Apple



# *Résumé des notions clés de la RI*

- Au cœur de tout système de RI (SRI)
  - Relation entre le document et ... la requête ou le besoin de l'utilisateur (Besoin en information)?
- 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.’

# **Notion de pertinence en RI**

# *Pertinence: notion clé de la RI (1/2)*

**Rappel :** Un système de recherche d'information doit satisfaire un besoin d'information d'un utilisateur

- Pertinence : capacité d'un système à répondre exactement à la requête demandée par l'utilisateur
- pertinence utilisateur vs pertinence système
- Pertinence utilisateur : satisfaction de l'utilisateur
- Pertinence système : estimation du système

**Pertinence utilisateur :** représente la façon dont l'utilisateur évalue les documents retrouvés par le SRI en fonction de son besoin d'information (on parle de jugements de pertinence).

- ✓ C'est une évaluation subjective : dépend de l'utilisateur et varie au cours du temps,
- ✓ Pas possible de la mesurer de manière automatique

## *Pertinence: notion clé de la RI (2/2)*

**Pertinence système :** mesurée automatiquement par les SRI en comparant les représentations des documents et celles des requêtes  
→ diffère d'un SRI à un autre en fonction des méthodes utilisées pour comparer les documents et la requête.

**Attention :** un document considéré comme pertinent par le système ne l'est pas nécessairement par l'utilisateur (et inversement).

- L'enjeu de la RI : rapprocher tant que possible **la pertinence système** de **la pertinence utilisateur**.
- **Plusieurs pertinences système**
  - **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

# Processus RI

Besoin en information ou **requête**



Requête

Langage de requêtes

Traitement

Liste de mots

SRI

Appariement  
Ranking

Traitement =  
Indexation

Index (mots clés)

Fichier  
inverse

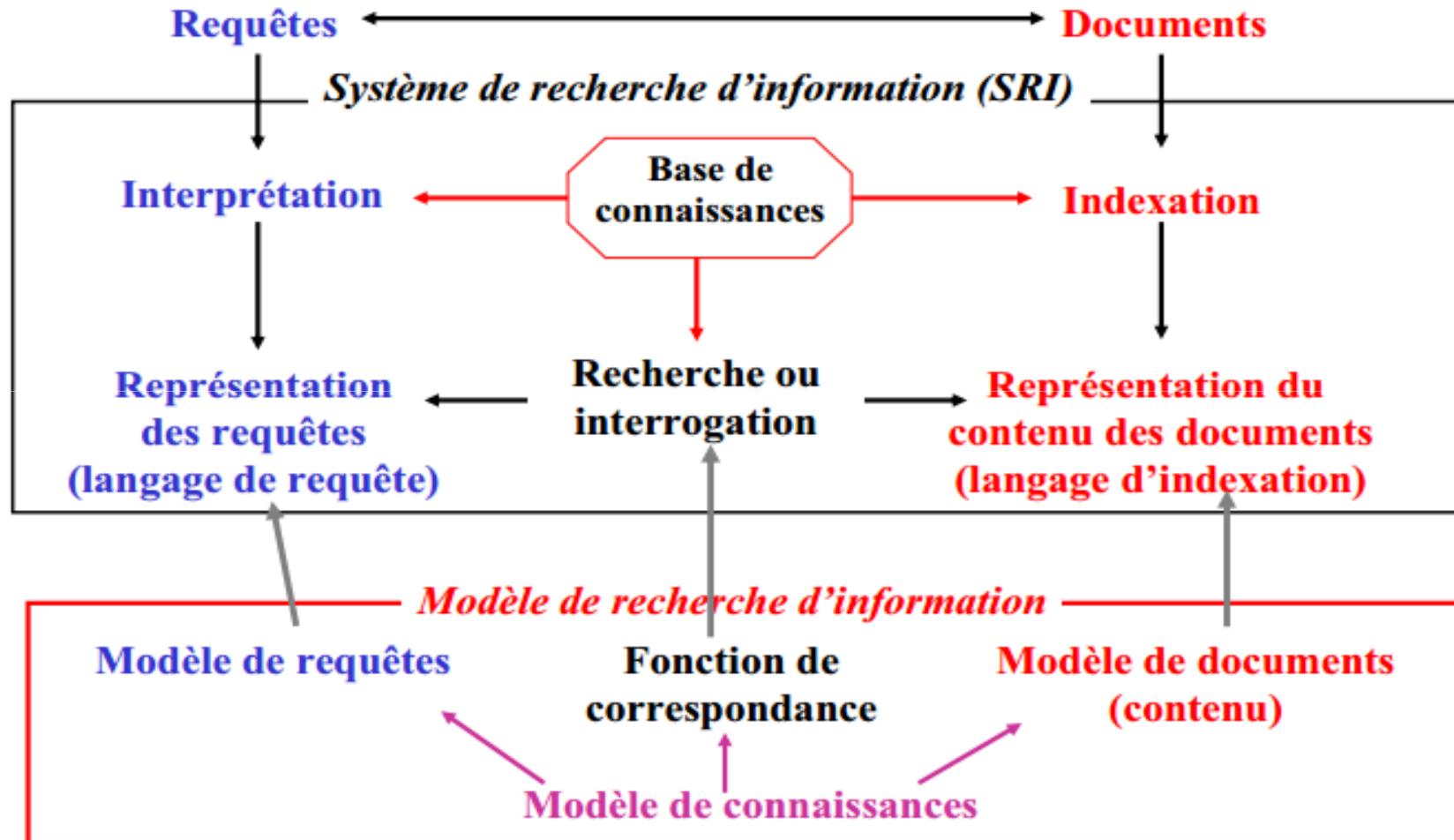
Visualisation

Modèles de RI :  
Vectoriel, probabiliste  
Ranking dans le web



Indexation et organisation physique

# *SRI et modèle de RI*



# *Pour récapituler...(1/2)*

## **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 ?

## *Pour récapituler...(2/2)*

- Déterminer si la représentation d'un document correspond à celle de la requête => développer un processus d'évaluation.
- Un bon système de RI doit donner une évaluation de correspondance qui reflète bien la pertinence du système, qui à son tour, correspond bien au jugement de pertinence de l'utilisateur.

# Métriques de base d'évaluation en RI

## *Evaluation d'un SRI*

- Le but de la RI est de trouver **des documents pertinents** à une requête, et donc utiles pour l'utilisateur.
- La qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir.
- Plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système.

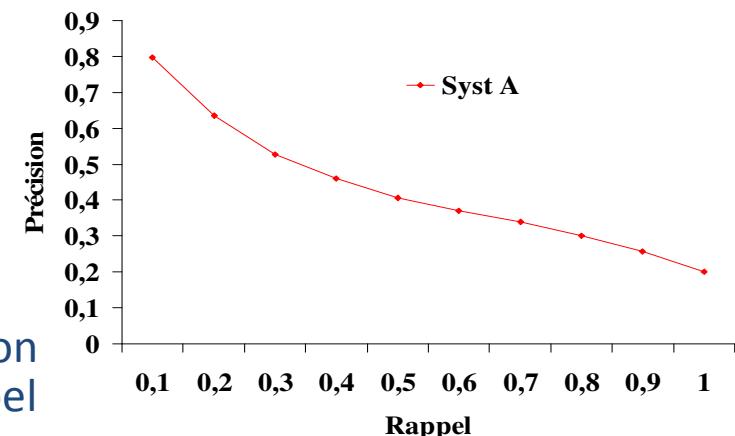
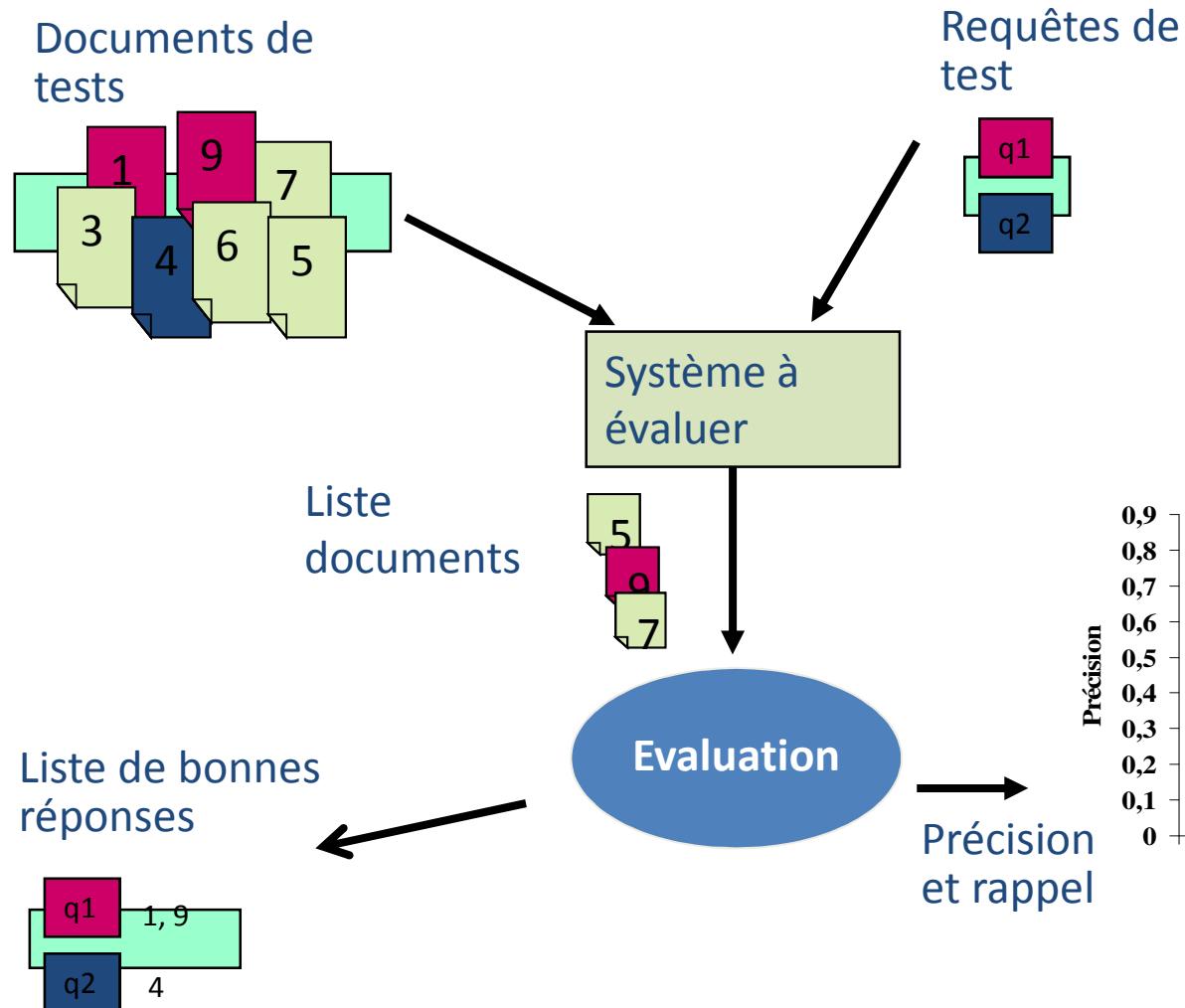
## *Collections de test*

- Pour arriver à une telle évaluation, on doit connaître d'abord les réponses idéales de l'utilisateur.
- Une collection de test (ou de référence) :
  - un ensemble de documents;
  - un ensemble de requêtes;
  - la liste de documents pertinents pour chaque requête.

# *Démarche d'évaluation en RI (1/2)*

- 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** » : collection de test.
  - 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)

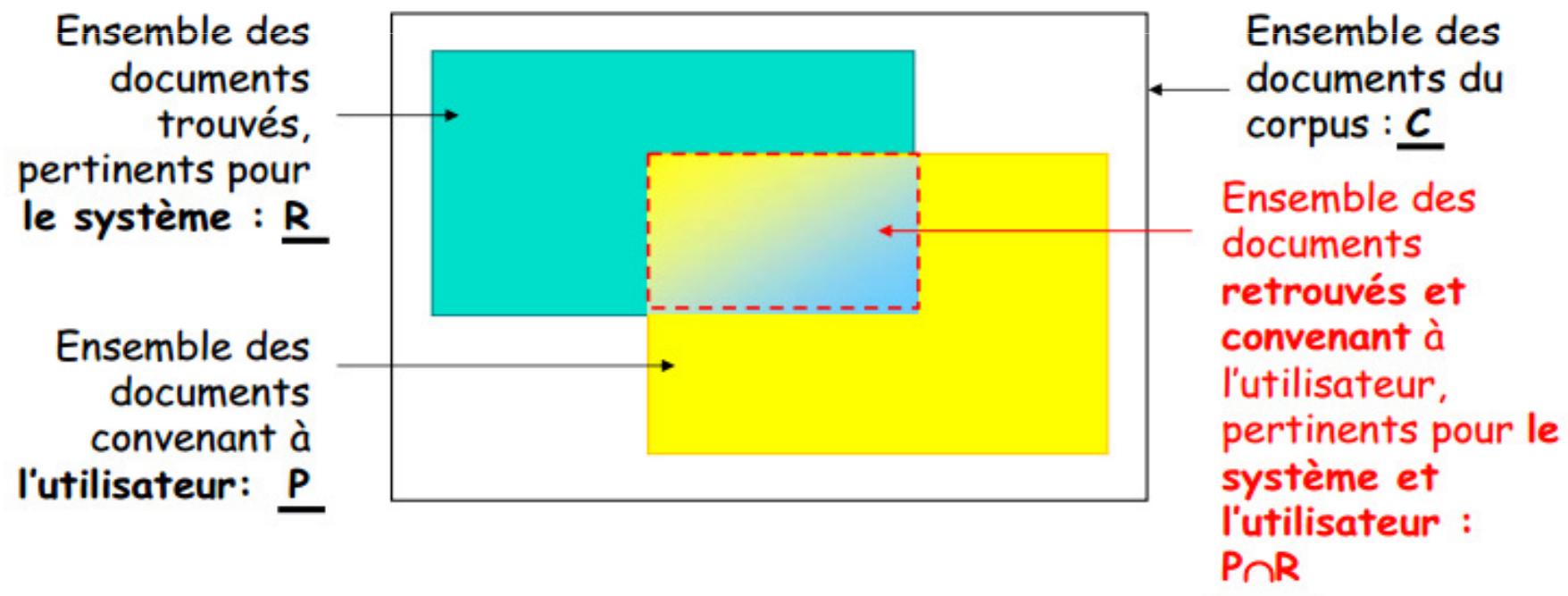
# Démarche d'évaluation en RI (2/2)



# Evaluation d'un SRI

Qu'est ce  
qu'un  
modèle RI?

- Comparer les SRI de manière théorique (via leur **modèle**) est un problème non-résolu, on passe donc par des évaluations de type «boîte noire » en testant les résultats d'un système par rapport aux réponses idéales attendues.
- **Objectif :** Rapprocher pertinence système et utilisateur



# *Rappel vs Précision (1/3)*

Les critères essentiels sont :

- **Le rappel** : capacité du système à fournir en réponse tous les documents pertinents
- **La précision** : capacité du système à ne fournir que des documents pertinents en réponse.

Ces deux critères sont antagonistes dans la réalité

## Rappel vs Précision (2/3)

- **Le rappel** est le rapport du nombre de documents trouvés par le système et convenant à l'utilisateur au nombre total de documents convenant à l'utilisateur.


$$\text{rappel} = \frac{|P \cap R|}{|P|} \in [0,1]$$

- **La précision** est le rapport du nombre de documents trouvés par le système et convenant à l'utilisateur au nombre de documents retrouvés par le système.


$$\text{précision} = \frac{|P \cap R|}{|R|} \in [0,1]$$

## *Rappel vs Précision (3/3)*

Pour une requête et un système : 2 valeurs réelles

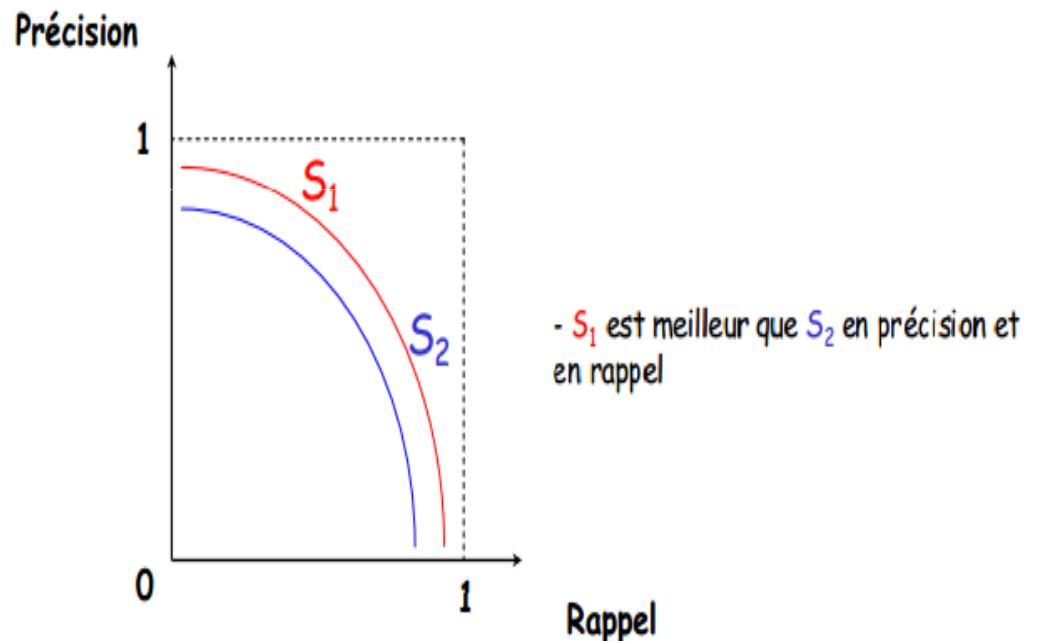
**Exemple :** un système retourne 5 documents, parmi lesquels 3 sont pertinents, sachant qu'il y a 10 documents pertinents dans le corpus, soit :

$$\text{Rappel} = 3 / 10$$

$$\text{Précision} = 3 / 5$$

Il faut des analyses plus fines  
des résultats :

**Courbes de rappel/précision**



# *Relation entre Rappel et Précision*

- Les deux métriques :
  - Ne sont pas indépendantes : quand l'une augmente, l'autre diminue.
  - Ne sont pas statiques : un système n'a pas qu'une mesure de précision et de rappel.
- Le comportement d'un système peut varier en faveur de précision ou en faveur de rappel (en détriment de l'autre métrique).

## *Idéalement...*

- Idéalement, on voudrait qu'un système donne de bons taux de précision et de rappel en même temps. Un système qui aurait 100% pour la précision et pour le rappel signifie qu'il trouve tous les documents pertinents, et rien que les documents pertinents.
- les réponses du système à chaque requête sont constituées de tous et seulement les documents idéaux que l'utilisateur a identifiés.
- En pratique, cette situation n'arrive pas. On peut obtenir un taux de précision et de rappel aux alentours de 30%.

# *Courbe rappel/Précision*

## *Courbes de rappel/précision*

Représente l'évolution de la précision et du rappel avec des résultats triés.

### Méthode :

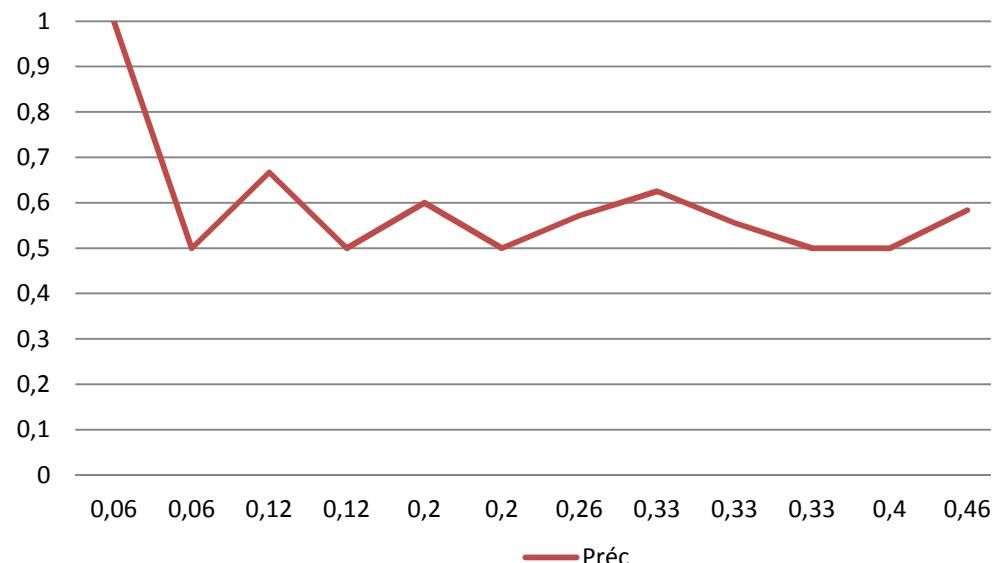
Pour chaque document retrouvé, on calcule la précision et le rappel obtenus en considérant seulement le premier document comme réponse, puis les deux premiers, puis les trois premiers etc., jusqu'à la réponse totale du système.

Doc	P(O/N)	Rap	Préc
D23	Oui	1/15=0,06	1/1=1
D12	Non	0,06	½=0,5
D55	Oui	2/15=0,12	2/3=0,66
D67	Non	0,12	2/4=0,5
D14	Oui	3/15=0,2	3/5=0,6
D22	Non	3/15=0,2	3/6=0,5
D40	Oui	4/15=0,26	4/7=0,57
D11	Oui	5/15=0,33	=5/8=0,62
D9	Non	5/15=0,33	= 5/9 =0,55
D44	Non	5/15=0,33	=5/10= 0,5
D87	Oui	6/15=0,4	= 6/11=0,54
D18	Oui	7/15=0,46	= 7/12=0,58

## QUESTION Rappel/précision

Un SRI retourne 12 documents, parmi lesquels 7 sont pertinents, sachant qu'il y a 15 documents pertinents dans le corpus

### Précision/Rappel



## *Taille d'un corpus*

- Pour qu'un corpus de test soit significatif, il faut qu'il possède un nombre de documents assez élevé.
- Les premiers corpus de test développés dans les années 1970 renferment quelques milliers de documents.
- Les corpus de test plus récents (par exemple, ceux de TREC) contiennent en général plus 100 000 documents (considérés maintenant comme un corpus de taille moyenne), voir des millions de documents (corpus de grande taille).

# *Courbe rappel/Précision (1/2)*

## **Courbes de rappel/précision**

Pour traiter l'évaluation sur plusieurs requêtes, on calcule la moyenne aux points de rappels standards.

### **Précision moyenne à x documents (P@x)**

- Il est également courant de calculer le taux de précision après un nombre de documents fixés pour une requête, puis de faire la moyenne sur toutes les requêtes.
- Il existe des programmes qui génèrent les tableaux pour les courbes de rappel/précision et les précisions moyennes à 5, 10, 20, 50 et 100 documents (Terrier, Lucene et TrecEval).

# *Comparaison de systèmes et Précision moyenne*

- Pour comparer deux systèmes de RI, il faut les tester avec le même corpus de test (ou plusieurs corpus de test) :
  - On utilise la *précision moyenne* comme une mesure de performance.  
La précision moyenne est une moyenne de précision sur un ensemble de points de rappel.
  - L'amélioration relative qui est calculée en tant Gain comme suit :

$$Gain = \left( \frac{Perf(SRI 2) - Perf(SRI 1)}{Perf(SRI 1)} \right)$$

## *Campagnes d'évaluation*

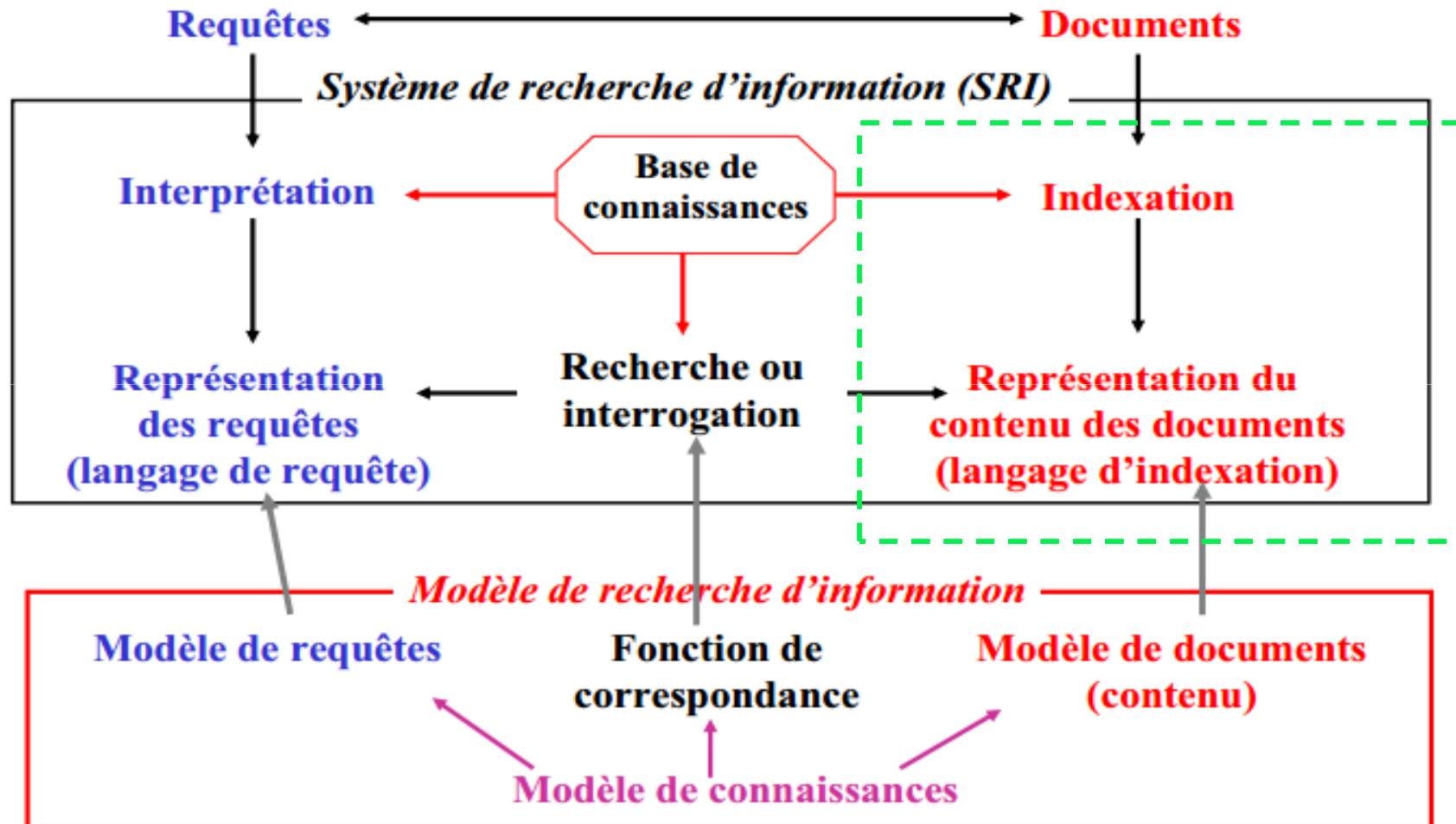
- TREC - Text REtrieval Conference
  - Évaluation des approches RI (beaucoup de tâches sont évaluées dans cette campagne)
- CLEF - Cross Language Evaluation Forum
  - Évaluation des approches de croisement de langues (multilinguisme)
- INEX - Initiative for the Evaluation of XML Retrieval
  - Évaluation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

# *Autres métriques en RI*

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

# **Indexation en R**

# *Indexation*



# Indexation: exemple

The screenshot shows a web browser window with the URL [https://en.wikipedia.org/wiki/Information\\_retrieval](https://en.wikipedia.org/wiki/Information_retrieval). The page title is "Information retrieval". A red dashed circle highlights the first paragraph of the main content area, which defines information retrieval as the activity of obtaining resources relevant to an information need from a collection of information resources. The paragraph also mentions automated systems used to reduce "information overload". Below this, a red dashed box highlights the "Contents" section and its sub-sections.

From Wikipedia, the free encyclopedia

**Information retrieval** (IR) is the activity of obtaining [information](#) resources relevant to an information need from a collection of information resources. Searches can be based on [metadata](#) or on [full-text](#) (or other content-based) indexing.

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 [hide]

- 1 Overview
- 2 History
- 3 Model types
  - 3.1 First dimension: mathematical basis
  - 3.2 Second dimension: properties of the model

# *Indexation: exemple*

- Décomposer le texte (Parsing)
- Décomposer les mots (Tokenizing)
- Supprimer les mots communs (Stop word removal)
  - Basé sur une “short list” “the”, “and”, “or”
- Raciniser les mots (Stemming)
- Regrouper les mots

<Title>: Information retrieval  
(Corps du texte) : Information retrieval (IR) is the activity of obtaining information resources...

Information, retrieval, information, retrieval, IR, is, the activity, of ,obtaining, information ...

Information, retrieval, information, retrieval, IR, activity, obtaining,

Information, retrieval, information, retrieval, IR, activity, obtain

Information 2, retrieval 2, IR 1, activity1, obtain 1

Un sac de mots (BOW)

# *Indexation : pourquoi ?*

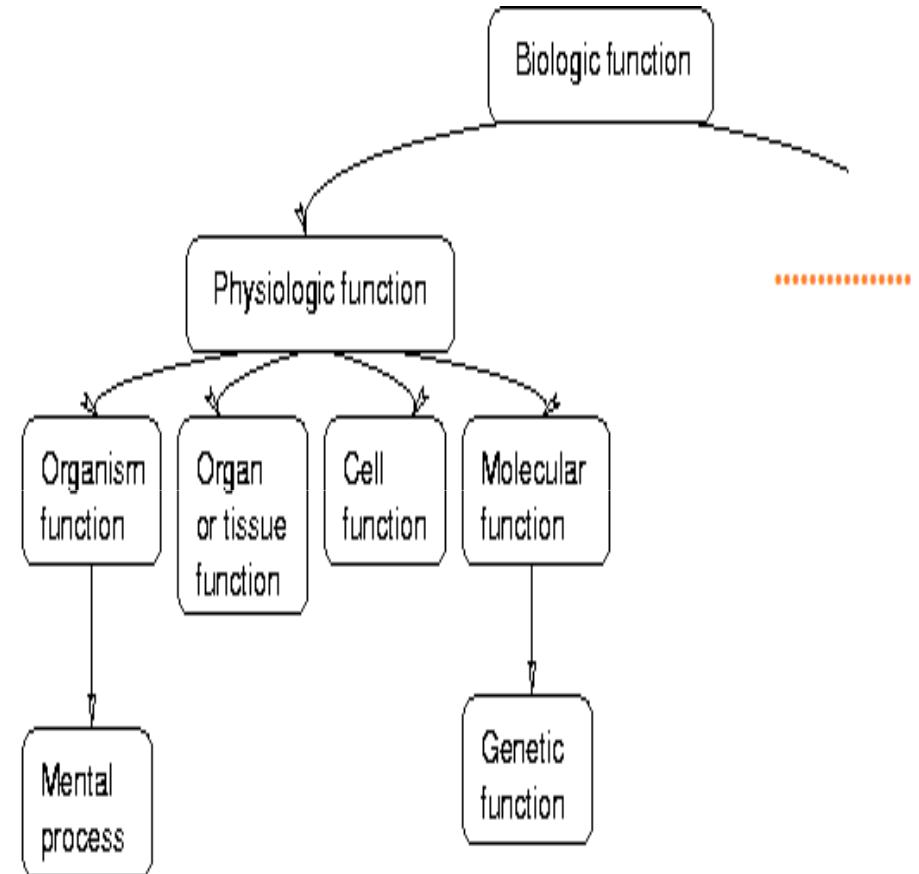
- L'idée principale du moteur de recherche est de retrouver les documents qui « **parlent de** » la requête.
- On utilise ce qu'on a sous la main : les **mots**
  - Qu'est-ce qu'un mot ?
  - Que faire lorsqu'un mot est « proche » d'un mot de la requête ?
- Le **parcours complet** de l'ensemble des documents avec les termes d'une requête est impossible : trop de documents et temps de réponse prohibitif.
- On passe par un traitement préalable : *l'indexation* :  
Le but de l'indexation automatique : "*transformer des documents en substituts capables de représenter le contenu de ces documents*"  
(Salton et McGill, 1983)

# *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 libre et contrôlée*

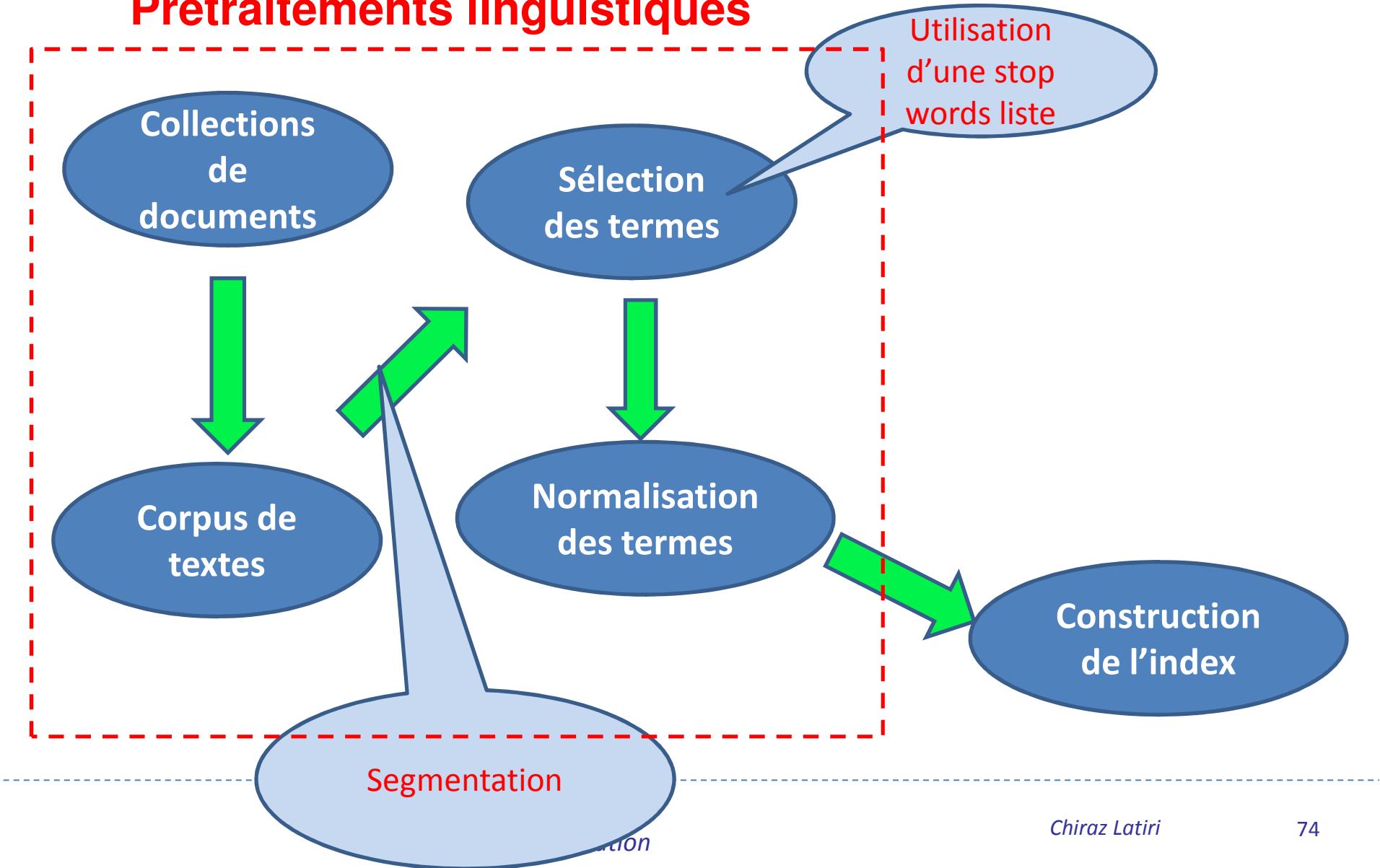
- Indexation **libre** :
  - Mots, termes des documents
- Indexation **contrôlée**
  - Listes de termes **prédéfinies**
  - Vocabulaire **contrôlé** (évite polysémie, synonymie et problèmes de granularité)
  - Thésaurus



*exemple : thésaurus UMLS*

# Approche générale de construction d'index

## Prétraitements linguistiques



# *Dans quels documents cherche-t-on ?*

- **Formats :**

- HTML (menus, tableaux, publicité, rendu)
- Texte brut (structure ?)
- pdf (problèmes d'encodage, rendu)
- Word (format propriétaire, structure)
- Excel (gestion des tableaux)
- OpenOffice (XML)
- ...



- Il est assez simple de détecter le **type** d'un document
- Des **heuristiques** spécifiques à chaque format pour extraire le texte
- Les moteurs de recherche utilisent très rarement la structure des documents

# *Dans quels documents cherche-t-on ?*

- **Langues**

- Identification de langues, un problème difficile
- Des documents multilingues
- De la recherche d'information multilingue

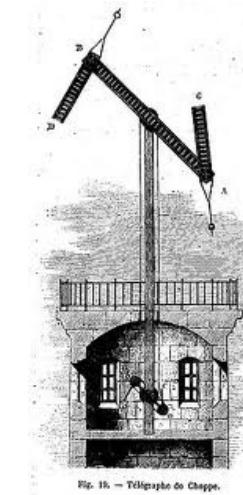


- **Encodages**

- Des erreurs dans la gestion de l'encodage peuvent conduire à des résultats erronés

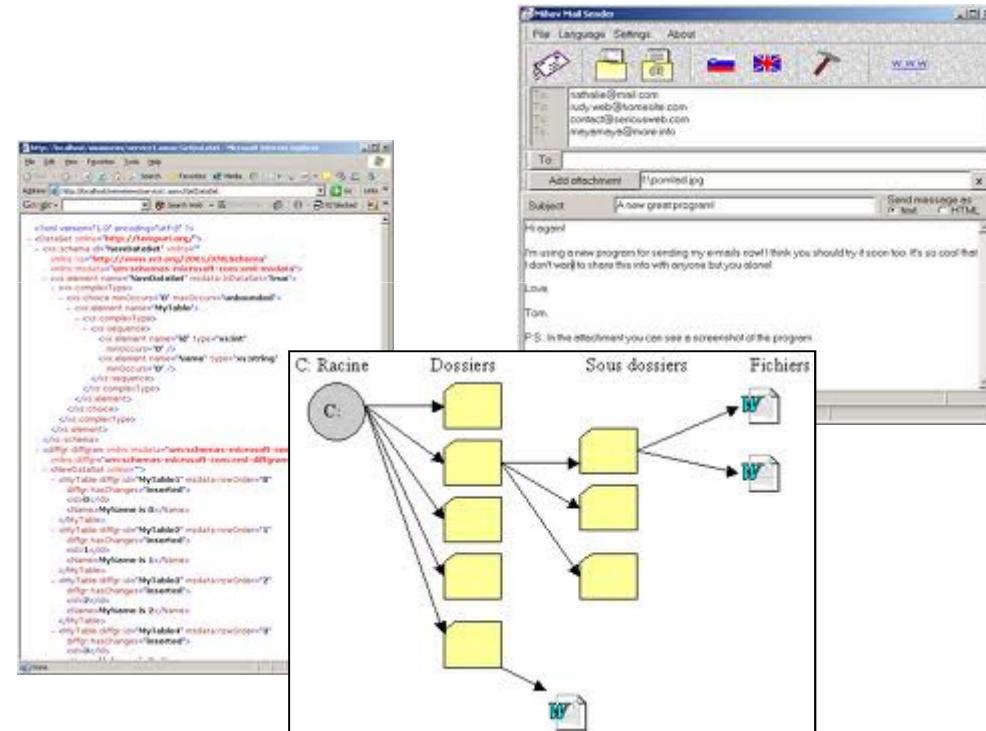
« prÃ©sident du PÃ©rou »

F	E	D	C	B	A
M	L	K	I	H	G
S	R	O	P	O	N
Y	X	W	V	U	T
+	3	2	1	&	Z
10	9	8	7	6	5



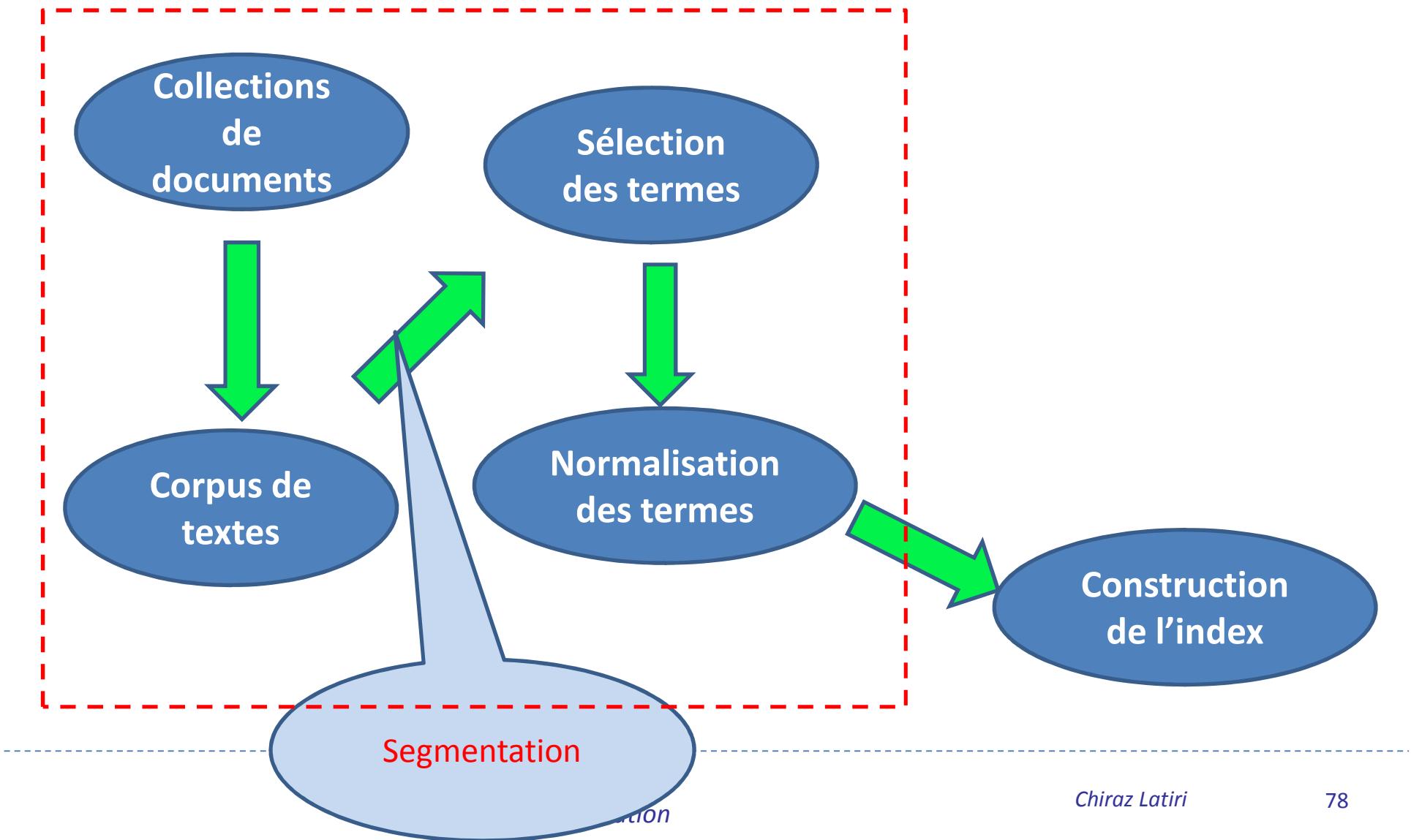
# Dans quels documents cherche-t-on ?

- « Unité » document
  - Un fichier ?
  - Un e-mail ?
    - Avec ses entêtes ?
    - Avec ses attachments ?
  - Un groupe de fichiers ?
    - Site Web
    - Document en plusieurs fichiers
  - Etc.



# Approche générale de construction d'index

## Prétraitements linguistiques



# *La segmentation (Tokenisation)*

- Identification des unités élémentaires (phonèmes, morphèmes, mots, etc.).  
Pour l'écrit, des mots et des phrases.
- Un problème très complexe dans certaines langues (chinois...)
- L'étape initiale indispensable pour tout travail sur le texte
- On obtient des *mots*, ou des *termes*, ou des *tokens*
- Ces unités seront les candidats à l'indexation et à la recherche dans une requête

# *La segmentation*

- Dans les langues « latines » :
  - Les **délimiteurs** de mots et de phrases peuvent être ambigus
    - *etc.*                    *T.A.L.*                    *21.3*                    *www.sncf.com*
    - *l'illusion*              *aujourd'hui*              *jusqu'à*
    - *Jean-Louis*             *donne-t-il*                *1914-1918*            *06-13-23-33-12*
  - Les mots (**noms propres** en particulier) peuvent avoir des variantes :
    - *Etats-Unis*            *États-Unis*
    - *France Inter*          *France-Inter*

# *La segmentation*

- Dans les langues "européennes" :
    - Les **nombres**, les **dates**
      - 14/07/1789 Mardi 12 mars
      - B-52
      - (+33) 6 45 65 13 95
      - Les anciens systèmes de RI retiraient tout simplement les nombres
      - Toujours source de beaucoup d'erreurs dans les systèmes de RI modernes

# *La segmentation*

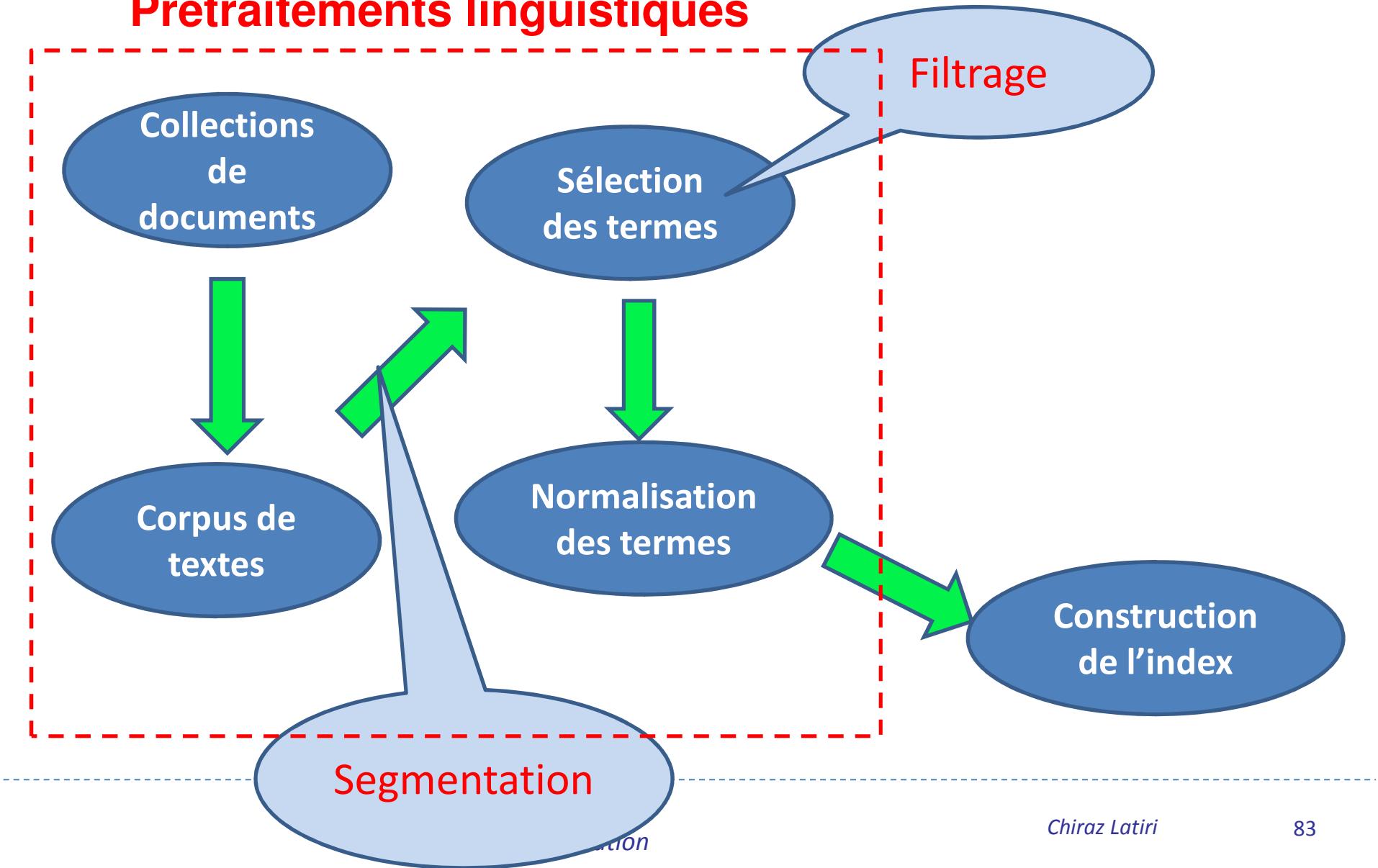
- En Japonais, Chinois, etc. il n'y a **pas d'espace entre les mots**
  - 學而不思則罔, 思而不學則殆。
  - La segmentation n'est pas toujours unique
- En Japonais, Coréen, on manipule **plusieurs types d'alphabets !**

東京、マドリード、イスタンブール(トルコ)が争う2020年夏季五輪の開催地は7日(日本時間8日)、ブエノスアイレスでの国際オリンピック委員会(IOC)総会で、IOC委員約100人の投票で決まる。
- En Arabe ou en Hébreu, on écrit **de droite à gauche**, mais certains éléments sont écrits de gauche à droite

يرتقب توزيع ما مجموعه 3026 وحدة سكنية جديدة موجهة لامتصاص السكن الهش عبر ولاية عنابة وذلك قبل نهاية السنة الجارية 2013

# *Approche générale de construction d'index*

## **Prétraitements linguistiques**



# *Filtrage des mots vides (Stopword list)*

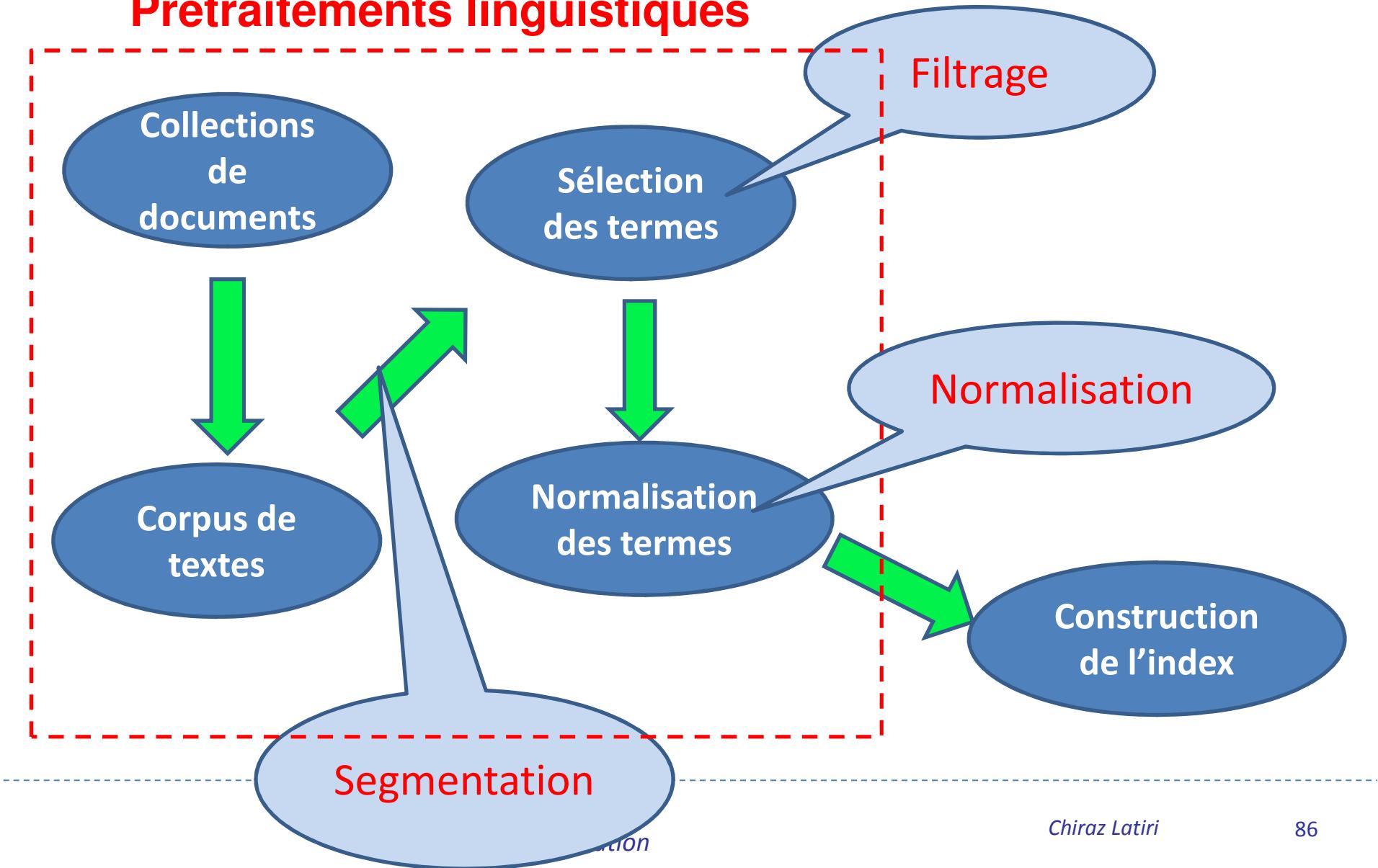
- Les mots « **outils** » n’apportent pas de sens au texte  
déterminants : « le », « la », pronoms : « je », « nous »,  
prépositions : « sur », « contre », ...
- Ce sont les mots les plus **fréquents** de la langue
  - Les 30 mots les plus fréquents représentent 30 % des occurrences de mots
  - Les supprimer permet d’économiser beaucoup de place dans l’index
- Mais :
  - On en a besoin pour des **requêtes multi-termes**  
« pomme de terre », « les Chevaliers du Zodiaque »
  - Ils sont parfois **porteurs de sens** dans des cas particuliers  
« Let it be », « The Who », « ça », « être ou ne pas être »
  - La **compression** permet finalement de conserver les mots vides dans peu d’espace

# *Antidictionnaire (Stopword list) Smart (571)*

a	all	and	are
a's	allow	another	aren't
able	allows	any	around
about	almost	anybody	as
above	alone	anyhow	aside
according	along	anyone	ask
accordingly	already	anything	asking
across	also	anyway	associated
actually	although	anyways	at
after	always	anywhere	available
afterwards	am	apart	away
again	among	appear	awfully
against	amongst	appreciate	
ain't	an	appropriate	

# Approche générale de construction d'index

## Prétraitements linguistiques



# *Normalisation textuelle (1)*

- Dans les **documents** comme dans la **requête**
- On veut par exemple **normaliser** :
  - « U.S.A. » et « USA » → USA
  - « morpho-syntaxe » et « morphosyntaxe » → morphosyntaxe
  - « Tuebingen », « Tübingen » et « Tubingen » → Tubingen
  - « Gorbatchov » et « Gorbatchev » → Gorbatchev
- Mais pas :
  - « sur » et « sûr »,
  - « pêche » et « péché »
  - En allemand, « mit » (avec) et « MIT »
  - En anglais, « C.A.T. » (Caterpillar) et « cat »
- Sans oublier les fautes de **frappe / d'orthographe**

# *Normalisation textuelle (2)*

- **Les règles communément utilisées :**
  - *Les ponctuations :*
    - Enlever les points et les traits d'union apparaissant dans les mots.
  - *La casse :*
    - Réduction de certaines lettres en minuscules.
    - L'heuristique la plus souvent utilisée : convertir les 1ères lettres des mots se trouvant au début de chaque phrase.
  - *Les accents :*

ambigüe → ambigue; forêt → foret,
  - *Les dates et les valeurs monétaires :* année/mois/jour ou mois/jour/année

# *Normalisation linguistique*

- Ramener un mot fléchi sous sa forme **canonique**.
- 3 types de normalisation linguistique:
  - La *lemmatisation*
  - La *racinisation*
  - L'*étiquetage*

# Lemmatisation

- Obtention de la **forme canonique** (le *lemme*) à partir du mot :
  - Pour un **verbe** : sa forme à l'infinitif (sans les flexions)  
*montrer, montreras, montraient* → montrer  
*am, are, is* → *be*
  - Pour un **nom, adjetif, article**, ... : sa forme au masculin singulier  
*vert, vertes, verts* → *vert*  
*computing* → *compute*  
*car, cars, car's, cars'* → *car*
- La lemmatisation demande des **ressources** et un **traitement linguistique**
  - En particulier pour les nombreuses exceptions
  - Long et donc difficile à mettre en œuvre pour des grandes collections
  - Dépendant de la langue

# Racinisation (stemming)

- Obtention de la **racine**, une forme tronquée du mot, commune à toutes les variantes morphologiques (troncature)
  - Suppression des **flexions**
  - Suppression des **suffixes**
  - Ex : *cheval, chevaux, chevalier, chevalerie, chevaucher*  
→ "cheva"(mais pas "cavalier")
  - **automate(s), automatic, automation** → **automat**
- La racinisation est généralement à base de règles
  - Rapide
  - Dépendant de la langue
- Elle agrège beaucoup plus que la lemmatisation
  - Index plus petit

# *Racinement : algorithme de Porter*

- 5 phases de **réduction** par **règles** (pour l'anglais, adapté ensuite au français)
- Si deux règles de réduction s'appliquent, on choisit celle qui supprime le plus long suffixe
  - *sses* → *ss*
  - *ies* → *i*
  - *ational* → *ate*
  - *tional* → *tion*
  - *Si  $m > 1$  alors cement* → ""  
*replacement* → *replac*  
*cement* → *cement*

# *Algorithme de Porter*

(Porter, M.F., 1980, *An algorithm for suffix stripping*, Program, 14(3) :130-137)

- Step 1: pluriels et participes passés
  - SSES -> SS caresses -> caress
  - (\*v\*) ING -> motoring -> motor
- Step 2: adj->n, n->v, n->adj, ...
  - (m>0) OUSNESS -> OUS callousness -> callous
  - (m>0) ATIONAL -> ATE relational -> relate
- Step 3:
  - (m>0) ICATE -> IC triplicate -> triplic
- Step 4:
  - (m>1) AL -> retrieval -> retrieiv
  - (m>1) ANCE -> allowance -> allow
- Step 5:
  - (m>1) E -> probate -> probat
  - (m > 1 and \*d and \*L) -> single letter controll -> control

# *Autres stemmers*

- Lovins stemmer  
<http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm>
- Krovetz stemmer (R. Krovetz, 1993: "Viewing morphology as an inference process," in R. Korfhage et al., *Proc. 16th ACM SIGIR Conference*, Pittsburgh, June 27-July 1, 1993; pp. 191-202.)

# *Exemples d'outils TALN*

*<http://www.atala.org/-Outils-pour-le-TAL-#type>*

<a href="#">ESSENTIAL (Résumeur de texte MULTILINGUE en ligne en 20 langues)</a>	Étiquetage, Traitement de corpus, Gestion de lexique, Autre (préciser)
<a href="#">Humanistique</a>	Étiquetage, Traitement de corpus, Gestion de lexique
<a href="#">DOSLoX &amp; WinLoX</a>	Traitement de corpus, Extraction de termes
<a href="#">CooLoX</a>	Traitement de corpus
<a href="#">SIAC</a>	Autre (préciser)
<a href="#">LIA PHON</a>	Étiquetage, Autre (préciser)
<a href="#">SyntEtiq</a>	Étiquetage, Analyse syntaxique
<a href="#">LEXTER</a>	Analyse syntaxique, Extraction de termes
<a href="#">MultAna</a>	Étiquetage
<a href="#">XLFG</a>	Analyse syntaxique
<a href="#">Lexed</a>	Gestion de lexique
<a href="#">TerminologyExtractor</a>	Extraction de termes
<a href="#">ACABIT</a>	Extraction de termes
<a href="#">DyALog</a>	Analyse syntaxique, Autre (préciser)
<a href="#">Sémiographe : fonctions syntaxiques</a>	Étiquetage, Analyse syntaxique, Autre (préciser)
<a href="#">Sémiographe : fonctions sémantiques</a>	Autre (préciser)
<a href="#">ANA</a>	Extraction de termes
<a href="#">Class4U</a>	Étiquetage, Traitement de corpus
<a href="#">Analyseur syntaxique du GREYC</a>	Analyse syntaxique, Traitement de corpus
<a href="#">CorTeCs</a>	Étiquetage
<a href="#">FASTER</a>	Extraction de termes
<a href="#">CORDIAL Universités ou Cordial Analyseur</a>	Étiquetage, Analyse syntaxique, Traitement de corpus, Extraction de termes

## *Exemple (tiré du livre de Croft et al.,)*

- Originale

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, ...

- Porter

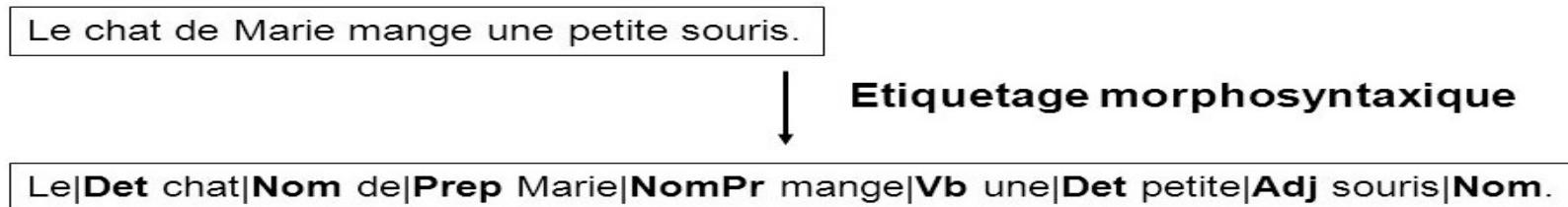
document describ market strategi carri compani agricultur chemic ...

- Krovetz

document describe marcketing strategy carry company agriculture chemical ...

# Étiquetage

- Associer aux mots leur **catégorie morphosyntaxique** (nom, verbe, adjectif, etc.)



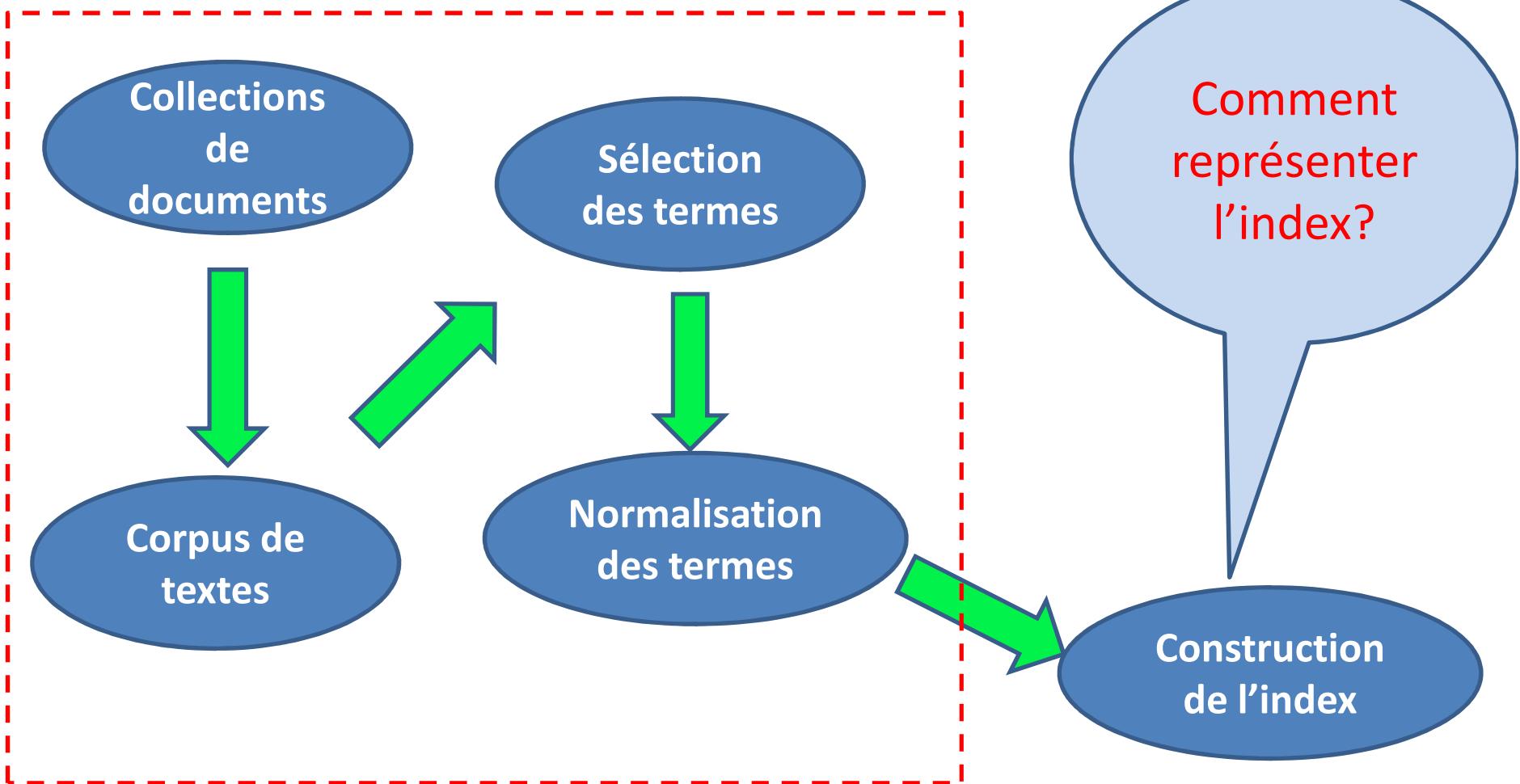
- Peut être utile en **recherche d'information** pour :
  - Supprimer les mots inutiles
  - Opérer des regroupements en termes complexes
  - Rechercher des mots ambigus avec plus de précision (*vers, or, pouvoir...*)
- L'outil le plus populaire
  - TreeTagger** : analyseur morpho-syntaxique
  - [http://txm.sourceforge.net/installtreekeeper\\_fr.html](http://txm.sourceforge.net/installtreekeeper_fr.html)

# *Récapitulation : La normalisation*

- Des analyses différentes pour des besoins différents :
  - **Lemmatisation** : pour rechercher/extraire de l'information, accéder au sens d'un lemme en faisant abstraction des flexions.
  - **Racinement (stemming)** : pour agréger les dérivations morphologiques à peu de frais, sans souci de la perte du sens et des lemmes initiaux.
  - **Étiquetage** : pour appliquer des techniques de TAL sur les catégories grammaticales plutôt que sur les mots eux-mêmes.
  - Types de **flexions**, de **dérivations** : pour appliquer des traitements plus fins en vue d'une analyse syntaxique et/ou sémantique.
- Des techniques assez bien maîtrisées : un pourcentage d'erreurs faible mais difficilement compressible.

# *Approche générale de construction d'index*

## **Prétraitements linguistiques**



# *Construction de l'index : le fichier inverse*

- Notion "classique" de l'index
- Un **fichier inverse** associe des index aux documents qui les contiennent. Chaque document possède un identifiant unique.
  - a ► d1, d2, d3, d4, d5...
  - à ► d1, d2, d3, d4, d5...
  - abaissa ► d3, d4...
  - abaissable ► d5
  - abandon ► d1, d5
  - abandonna ► d2
  - abasourdi ► d1
  - ...

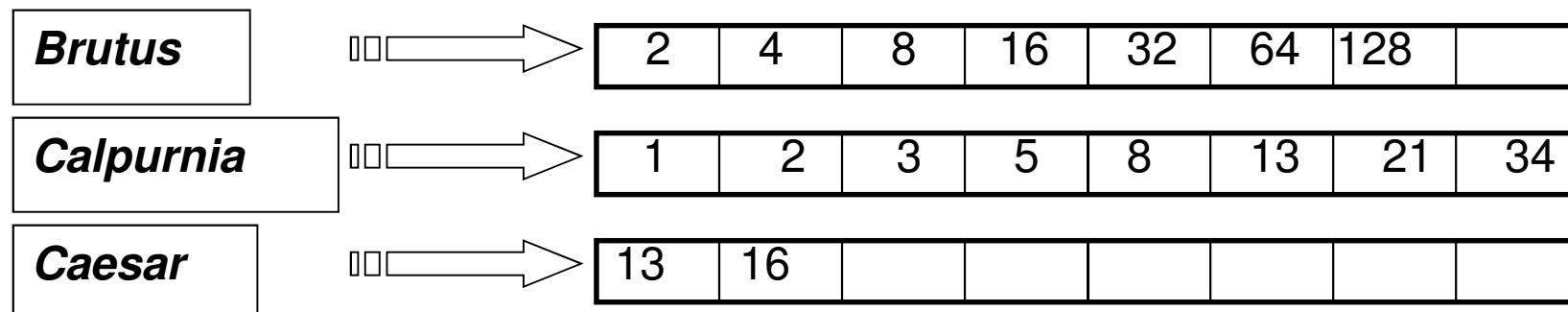
# Sac de mots

- Modèles « **sac de mots** » pour l'indexation et la recherche :
  - On oublie l'**ordre des mots**  
    (« *Jean est plus rapide que Marie* » = « *Marie est plus rapide que Jean* »)
  - On raisonne en termes de **présence / absence** des termes dans un document,  
ou en terme de **fréquence** de ces termes



# *Index inversé*

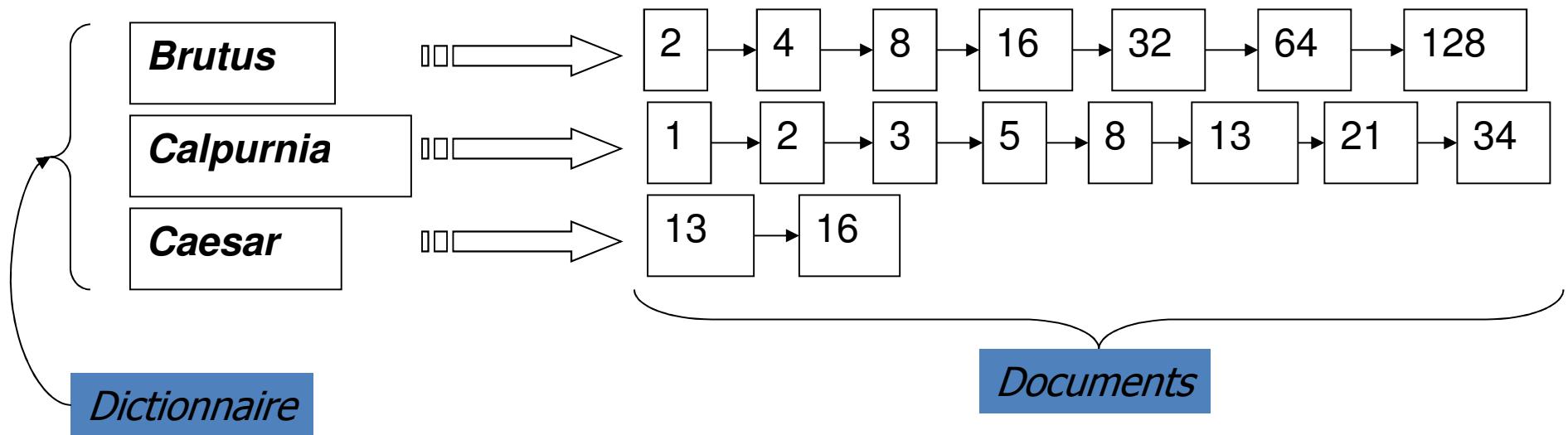
- Pour chaque terme  $T$ , on doit stocker la liste de tous les documents qui contiennent  $T$ .
- Que doit-on utiliser: un tableau ou une liste?



# *Index inversé*

Les listes chainées sont généralement plus utilisées que les tableaux :

- Allocation dynamique de l'espace mémoire



# *Index inversé:*

## *1/ Extraire les mots de chaque document*

- Extraire les termes de chaque document dans un fichier (1 fichier par document) ou un fichier pour plusieurs documents)

Ici un fichier pour les deux documents



Doc 1

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

Doc 2

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

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambition	2

# *Index Inversé:*

## *2/ Trier le fichier termes-documents*

Trier le fichier par  
ordre alphabétique des  
termes et par document



Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

*Chiraz Latiri*

105

Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

# Index Inversé:

## 3/ Trier le fichier termes-documents

- Pour chaque terme,
  - on dispose de la liste de documents qui le contient
  - Le nombre de documents comportant ce terme



Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

Term	Doc #	Freq
ambitious	2	1
be	2	1
brutus	1	1
brutus	2	1
capitol	1	1
caesar	1	1
caesar	2	2
did	1	1
enact	1	1
hath	2	1
I	1	2
i'	1	1
it	2	1
julius	1	1
killed	1	2
let	2	1
me	1	1
noble	2	1
so	2	1
the	1	1
the	2	1
told	2	1
you	2	1
was	1	1
was	2	1
with	2	1

# Construction de l'index

Terme	Id. Doc
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
....	....

Fichier inverse  
(dictionnaire)

En RI,  
"fréquence" =  
"nb d'occurrences"



Terme	Fréquence	Liste
ambitious	1	→ 2
be	1	→ 2
brutus	2	→ 1 → 2
capitol	1	→ 1
caesar	2	→ 1 → 2
did	1	→ 1
enact	1	→ 1
hath	1	→ 2
i	1	→ 1
i'	1	→ 1
it	1	→ 2
julius	1	→ 1
killed	1	→ 1
let	1	→ 2
me	1	→ 1

# Construction de l'index

Terme	Fréquence	Liste
ambitious	1	→ 2
be	1	→ 2
brutus	2	→ 1 → 2
capitol	1	→ 1
caesar	2	→ 1 → 2
did	1	→ 1
enact	1	→ 1
hath	1	→ 2
i	1	→ 1
i'	1	→ 1
it	1	→ 2
julius	1	→ 1
killed	1	→ 1
let	1	→ 2
me	1	→ 1

Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



Term	Doc #	Freq
ambitious	2	1
be	2	1
brutus	1	1
brutus	2	1
capitol	1	1
caesar	1	1
caesar	2	2
did	1	1
enact	1	1
hath	2	1
I	1	2
i'	1	1
it	2	1
julius	1	1
killed	1	2
let	2	1
me	1	1
noble	2	1
so	2	1
the	1	1
the	2	1
told	2	1
you	2	1
was	1	1
was	2	1
with	2	1

## *Index Inversé:*

## 4/ Construire le dictionnaire et le « posting »

## Fichier documents

# *Recherche de groupes de mots*

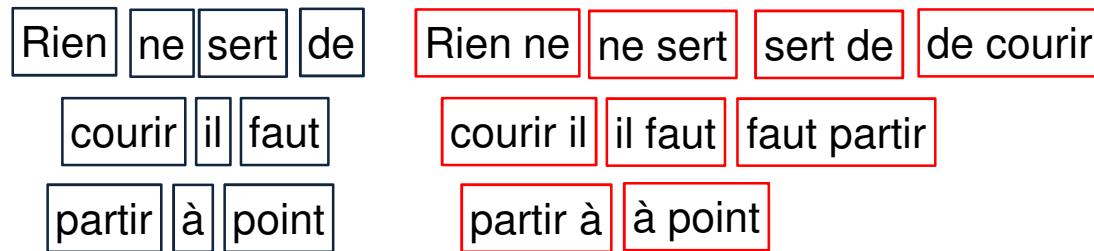
- Recherche sur **Nicolas Sarkozy**.
- On veut obtenir des textes contenant « **Nicolas Sarkozy** », et non par exemple :  
*« Nicolas Bedos après son sketch sur DSK et Carla Bruni-Sarkozy. »*
- De nombreuses requêtes sont implicitement des recherches de **groupes de mots**, au moins en partie :  
*« Nicolas Sarkozy » Disneyland*
- Nos index inversés <terme : documents> ne suffisent plus

# *La notion de n-grammes*

- **n-gramme** : une sous-séquence de  $n$  éléments extraite d'une séquence donnée. (cf. modèles de Markov)
- *Ici, n-grammes de mots*
  - uni-gramme : tous les mots
  - bi-gramme : une sous-séquence de 2 éléments
  - etc.
- Différent du groupe de mots d'un point de vue linguistique
- Pour la génération des n-grammes à partir d'un corpus de textes, on utilise des outils de modélisation statistique de langage tel que **SRILM** le plus utilisé (<http://www.speech.sri.com/projects/srilm/>)

# *Index de bi-grammes (1/4)*

- Indexer (en plus des mots simples) toutes les paires de termes du texte.



Comment éviter d'indexer toutes les paires ?

- On considère donc chaque bi-gramme comme un terme du dictionnaire
- Une requête sur un bi-gramme est immédiate

## *Index de bi-grammes (2/4)*

- Les requêtes plus longues ( $n$ -grammes,  $n > 2$ )

« pommes de terre »



« pommes de » AND « de terre »

et ainsi de suite pour des requêtes encore plus longues...

Risque de faux positifs, pourquoi ?

## *Index de bi-grammes (3/4)*

- Autre solution plus économique, on supprime les « mots vides » dans l'index et dans la requête

« pommes de terre »



« pommes terre »

Encore un risque de faux positifs, pourquoi ?

- Ça ne suffit pas pour

« Université Paris-Sud 11 » ou

« Centre National de la Recherche Scientifique »

## *Index de bi-grammes (4/4)*

- Conduit à des **faux positifs**
- Dictionnaire **beaucoup plus gros** et index vite ingérable
- **Impraticable** pour  $n > 2$

• si on a 200 000 termes uniques  
• et si on considère les n-grammes de  $n = 1$  à  
• on obtient un dictionnaire de  $3,2 \times 10^{26}$  entrées !

- On peut utiliser des index bi-grammes dans certaines situations ou pour certains groupes de mots, mais ce n'est pas la solution standard pour la recherche de groupes de mots.

⇒ Index de positions

# **Deux lois statistiques au cœur de la RI**

# *Les statistiques au cœur de la RI*

## 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).

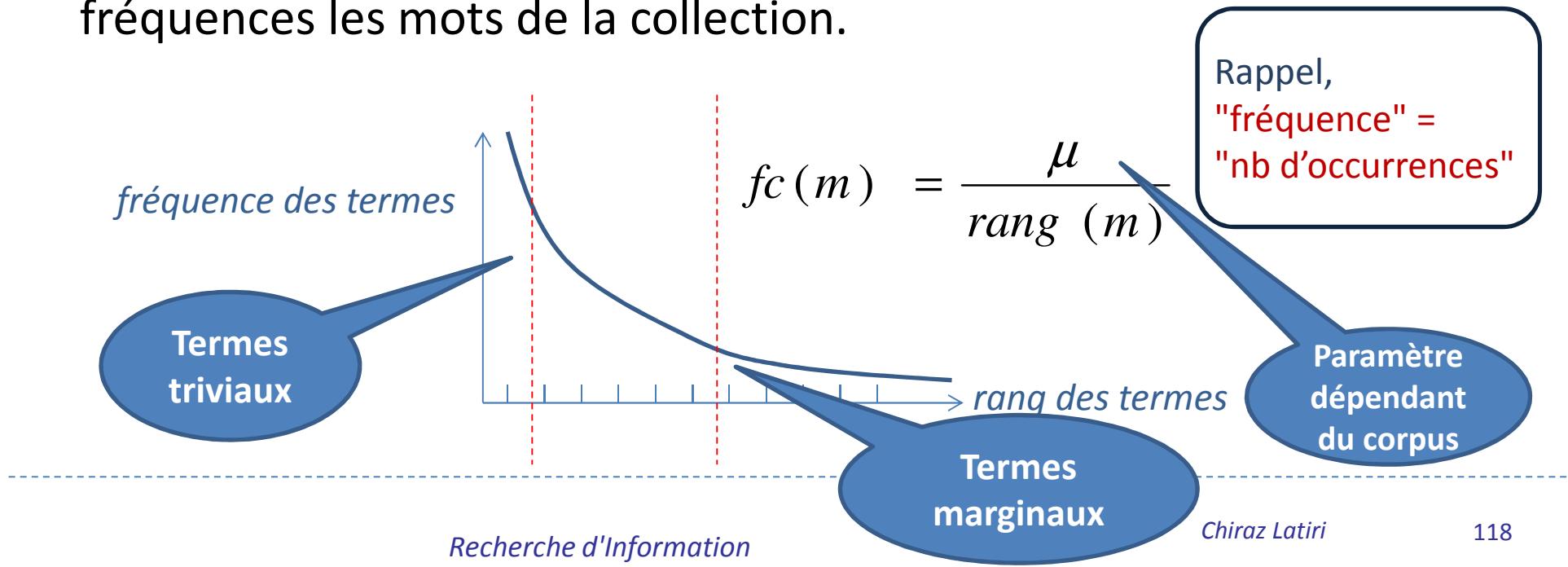
LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, **2**, 159-165 (1958).

# *Loi de zipf : fréquence des termes*

- **Constat intuitif :** Beaucoup de mots fréquents, et peu de mots rares.

**Loi de Zipf :** la fréquence d'occurrence  $fc(m)$  d'un mot  $m$  dans une collection de documents est inversement proportionnelle à son rang.

- Le rang est obtenu lorsque l'on trie par ordre décroissant des fréquences les mots de la collection.



## *Tf : term frequency*

- $tf$  (fréquence des termes) : La fréquence d'un terme t est un indicateur de son importance
- Intuitivement, plus un document contient d'occurrences d'un terme t, plus il sera pertinent pour une requête contenant ce terme t.

$tf_{t,d}$  = Nombre d'occurrences du terme  $t$  dans le document  $d$

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

## *Idf : inverse document frequency*

- IDF (Inverse Document Frequency) la fréquence inverse du terme dans la collection, plus un terme est fréquent moins il est discriminant.
- Les termes très fréquents dans tous les documents ne sont pas si importants (ils sont moins **discriminants**)
- On compense donc la fréquence des termes dans les documents (*tf*) en prenant en compte leur fréquence dans la collection (*df*)

$$idf(t) = \log \left( \frac{N = Nb\_docs\_dans\_la\_collection}{df(t) = Nb\_docs\_contenant\_t} \right) = \log \left( \frac{N}{df(t)} \right)$$

## *Pondération $tf \times idf$ : poids d'un terme*

- Le **poids**  $w_{i,j}$  d'un terme  $t_i$  dans un document  $d_j$  est la combinaison de ces deux métriques ( $tf \times idf$ ) pour rendre compte du caractère discriminant d'un terme dans un document.
- Le poids  $w_{i,j}$  d'un terme  $t_i$  dans un document  $d_j$  :
  - augmente avec sa **fréquence dans le document**
  - augmente avec sa **rareté dans la collection**

$$w_{i,j} = tf(t_i, d_j) \times idf(t_i, d_j)$$

## *Loi de Heaps pour la caractérisation du vocabulaire*

- La taille du **vocabulaire (V)**, appelé aussi **Lexique**, croît exponentiellement en fonction du nombre de mots présents dans un **corpus (M)**.
- **Loi de Heaps** : Caractérise le nombre de mots distincts dans un corpus

$$V = K \times M^\beta \text{ avec } 0 < \beta < 1$$

- $\beta$  dépend de la **langue** de  $M$
- et  $K$  dépend des **prétraitements** appliqués à  $M$  pour extraire les termes.

*exemple sur  
un corpus de  
~1M mots:*

Taille du lexique



# Modèles de RI

# Processus RI

Besoin en information ou **requête**



Requête

Langage de requêtes

Traitement

Liste de mots

SRI

Appariement  
Ranking

Traitement =  
Indexation

Index (mots clés)

Fichier  
inverse

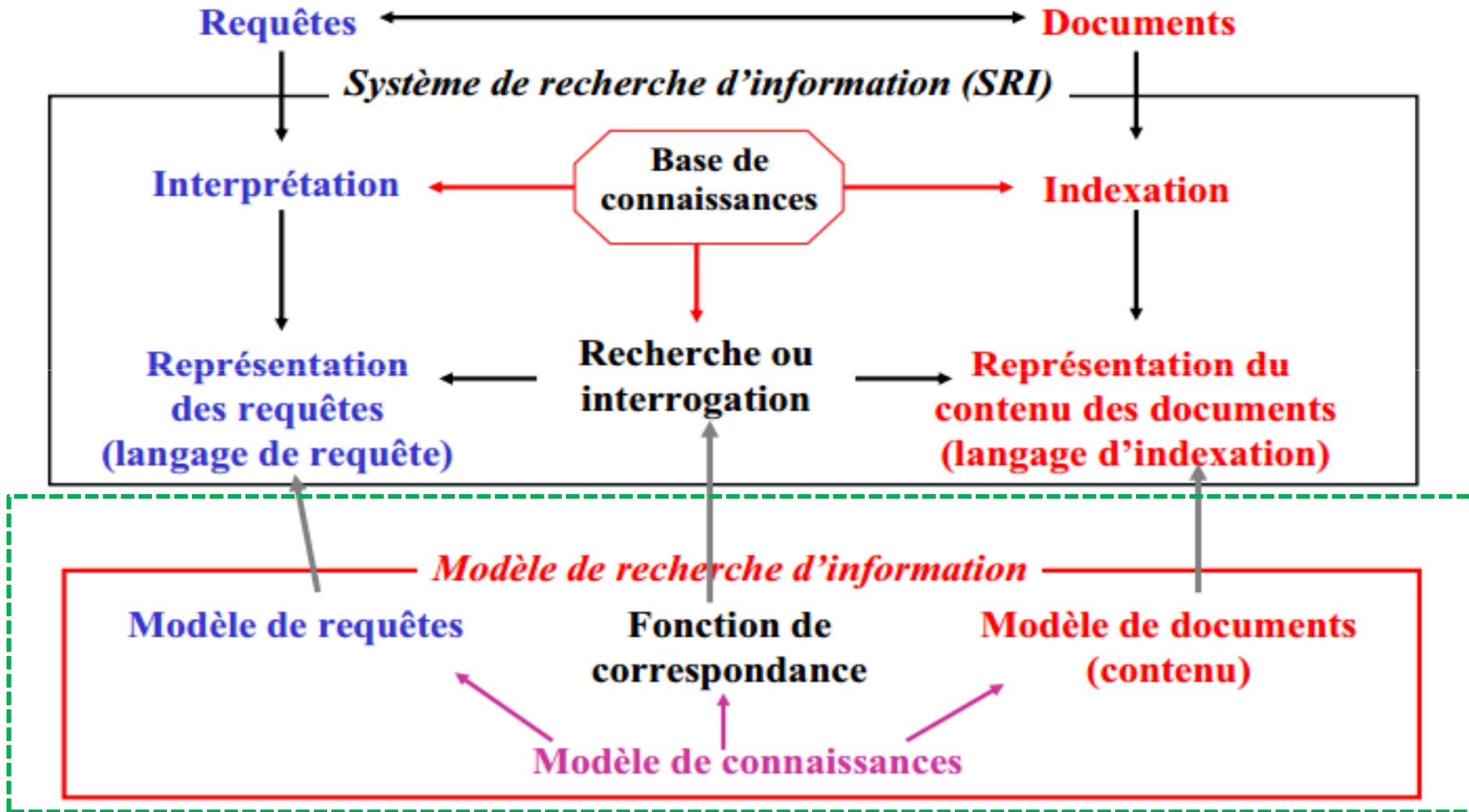
Visualisation

Modèles de RI :  
Vectoriel, probabiliste  
Ranking dans le web

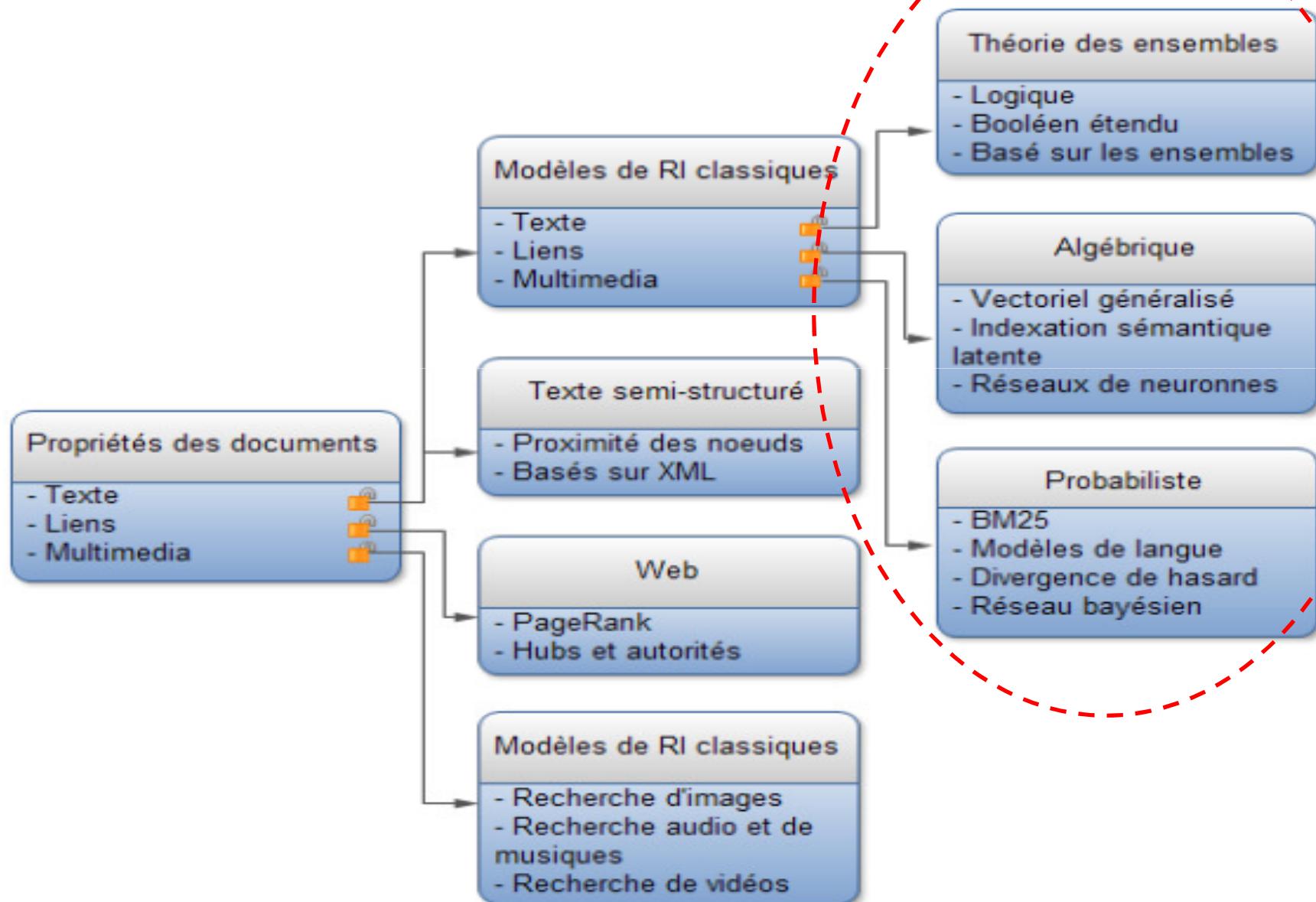


Indexation et organisation physique

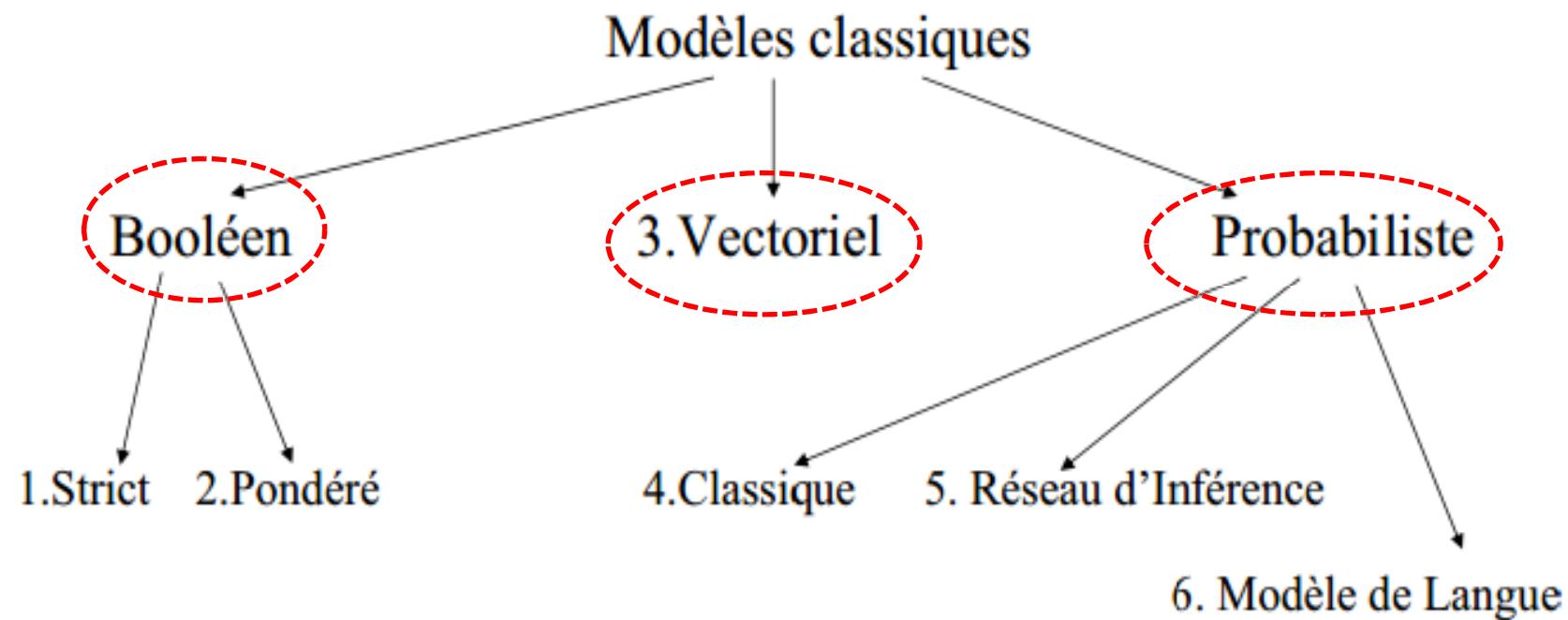
# Modèle de RI : au cœur du SRI



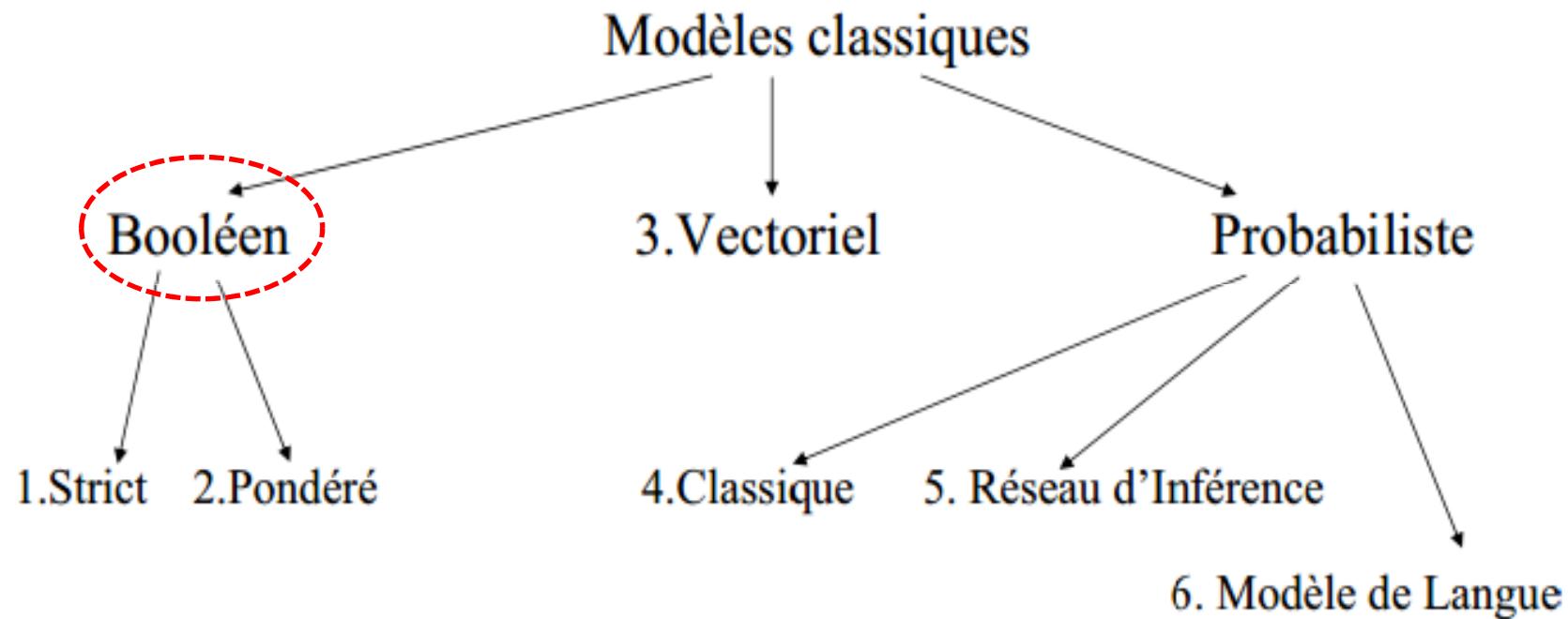
# Taxonomie des modèles en RI (1/2)



## *Taxonomie des modèles en RI (2/2)*



# *Modèle Booléen*



# *Le modèle booléen (1/2)*

- Le premier et le plus simple des modèles
- Basé sur la théorie des ensembles et l'**algèbre de Boole**
- Les termes de la requête sont soit présents soit absents
  - **Poids binaire** des termes, 0 ou 1
- Un document est soit pertinent soit non pertinent
  - **Pertinence binaire**, et jamais partielle (**modèle exact**)
- La requête s'exprime avec des **opérateurs logiques**
  - AND, OR, NOT
  - (cyclisme OR natation) AND NOT dopage
  - le document est pertinent si et seulement si son contenu respecte la formule logique demandée

## *Le modèle booléen (1/3)*

- Modèle de connaissances :  $T = \{t_i\}$ ,  $i [1, .. N]$ 
  - Termes  $t_i$  qui indexent les documents
- Le modèle de documents (contenu) est une expression booléenne dans la logique des propositions avec les  $t_i$  considérés comme des propositions :
- Un document D1 est représenté par une formule  $\mathcal{D}1$ :

$$\mathcal{D}1 = t_1 \wedge t_3 \wedge t_{250} \wedge t_{254}$$

Une requête Q est représentée par une formule logique  $\mathcal{Q}$ :

$$Q = (t_1 \wedge t_3) \vee (t_{25} \wedge t_{1045} \wedge \neg t_{134})$$

## *Le modèle booléen (2/3)*

- **Document D** = Conjonction logique de termes (non pondérés)
- **Requête Q** = expression booléenne de terme
- $R(D, Q) = D \rightarrow Q$

### Exemple

$$D_1 = t_1 \wedge t_2 \wedge t_3 \quad (\text{les 3 termes apparaissent dans } D)$$

$$D_2 = t_2 \wedge t_3 \wedge t_4 \wedge t_5$$

$$Q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$$

$D_1 \rightarrow Q$ , alors  $R(D_1, Q) = 1$ .

mais  $D_2 \not\rightarrow Q$ , alors  $R(D_2, Q) = 0$ .

## *Le modèle booléen (3/3)*

La fonction de correspondance est basée sur l'implication logique en logique des propositions :

Un document D répond à une requête Q si et seulement si

$$D \supset Q$$

Utilisation de déduction par :

Axiomes :  $(a \wedge b) \supset a$ ,  $(a \wedge b) \supset b$ ,  $a \supset (a \vee b)$ ,  $b \supset (a \vee b)$ , ...

modus ponens (MP) : si  $a$  et  $a \supset b$  alors  $b$



Permet l'expression de requêtes complexes

## *Le modèle booléen : exemple*

- $D = t_1 \wedge t_3$
- $Q = t_1 \vee t_4$

**Déduction :**

**x:**  $t_1 \wedge t_3 \supset t_1$  (équivalent à  $D \supset t_1$ )

$\text{MP}(x) : t_1$

**y:**  $t_1 \supset t_1 \vee t_4$  (équivalent à  $t_1 \supset Q$ )

$\text{MP}(y) : Q$

**Conclusion**

Q est dérivable à partir de D donc le document D répond à la requête.

# Modèle booléen : exemple

Requête  $Q$  : (cyclisme OR natation) AND NOT dopage

Le document contient					Pertinence du document
cyclisme	natation	cyclisme OR natation	dopage	NOT dopage	
0	0	0	0	1	0
0	0	0	1	0	0
0	1	1	0	1	1
0	1	1	1	0	0
1	0	1	0	1	1
1	0	1	1	0	0
1	1	1	0	1	1
1	1	1	1	0	0

## *Remarques sur le modèle booléen*

- Correspondance stricte : Oui/Non
    - $Q = t_1 \wedge t_3 \wedge t_4$
    - $D1 = t_1 \wedge t_4,$   
 $Q \not\subseteq D1$
    - Le document  $D1$  (représenté par  $D1$ ) n'est pas pertinent pour la requête  $Q$  (représentée par  $Q$ ) d'après le modèle, alors qu'il contient une description « proche » de la requête.
  - Pas de distinction entre les documents pertinents
    - $Q = t_1 \wedge t_4$
    - $D2 = t_1 \wedge t_4,$
    - $D3 = t_1 \wedge t_3 \wedge t_4 \wedge t_5 \wedge t_6 \wedge t_7$   
 $D1 \supset Q$  et  $D3 \supset Q$
- Le document  $D2$  (représenté par  $D2$ ) est-il plus ou moins pertinent que  $D3$  (représenté par  $D3$ ) pour la requête  $D$  (représentée par  $Q$ ) ?**

# *Modèle booléen : avantages et inconvénients*

- **Avantages :**

- Le modèle est **transparent** et **simple** à comprendre pour l'utilisateur :
  - Pas de paramètres "cachés"
  - Raison de sélection d'un document claire : il répond à une formule logique
- Adapté pour les spécialistes (**vocabulaire constraint**)

- **Inconvénients :**

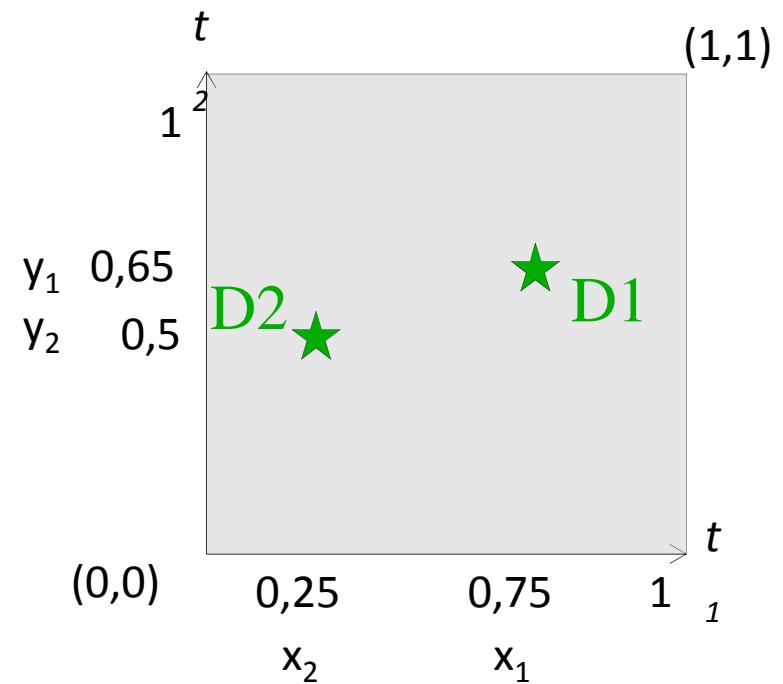
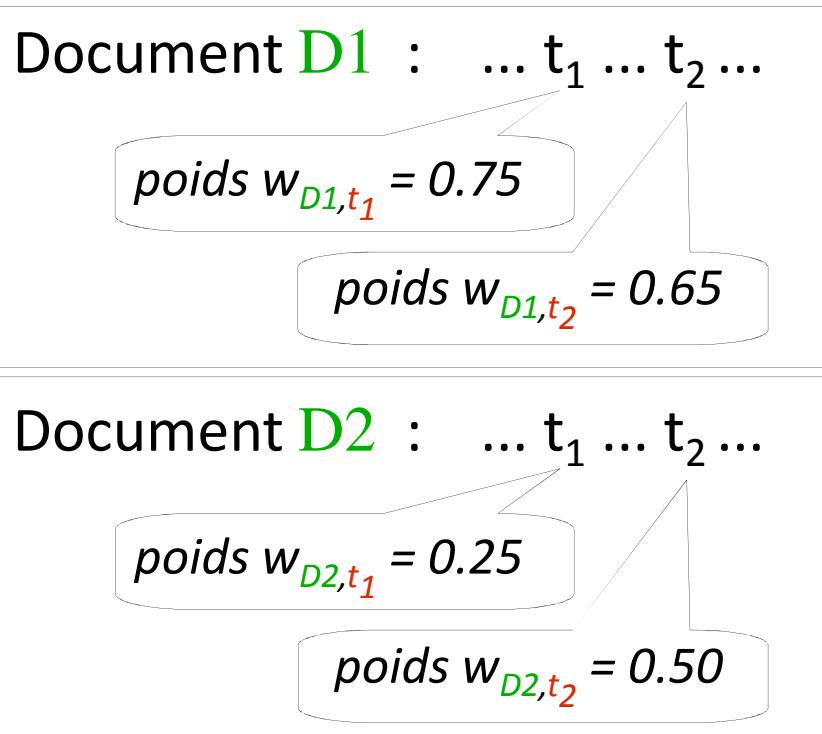
- Il est difficile d'exprimer des requêtes longues sous forme booléenne
- Le **critère binaire** peu efficace
  - Il est admis que la pondération des termes améliore les résultats
  - cf. modèle booléen étendu
- Il est impossible d'**ordonner** les résultats
  - Tous les documents retournés sont sur le même plan
  - L'utilisateur préfère un classement lorsque la liste est grande

# *Modèle booléen étendu*

- Idée : permettre l'utilisation des **opérateurs logiques** tout en proposant une **pertinence graduée**
- **Combinaison des modèles booléen et vectoriel**
- Utilisation de la **pondération** des termes dans un document (tf.idf)
- Comme dans le modèle vectoriel, positionnement des documents dans un espace euclidien dont les axes sont les termes de la requête
- Calcul de la distance entre les coordonnées du document et :
  - les coordonnées idéales (requête ET)
  - les coordonnées nulles (requête OU)

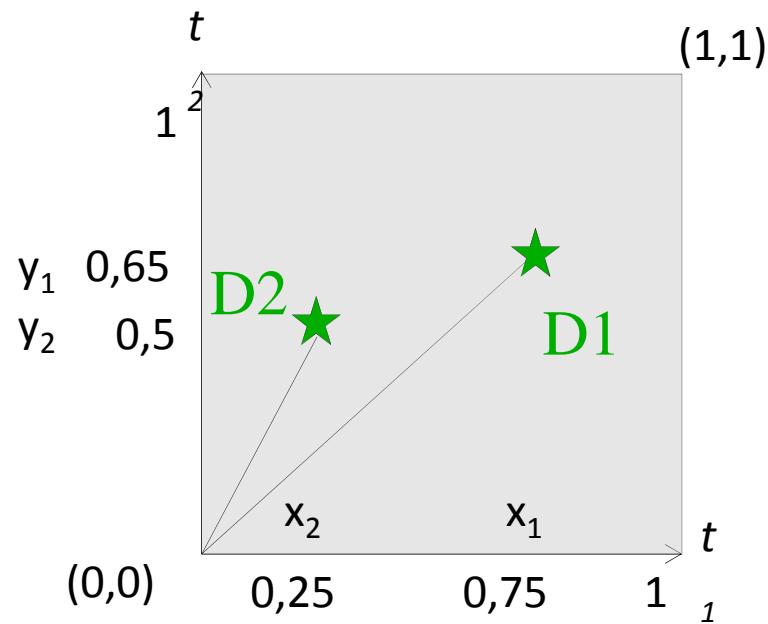
# Modèle booléen étendu : exemple (1/2)

Requête  $Q$  :  $t_1 \text{ AND/OR } t_2$

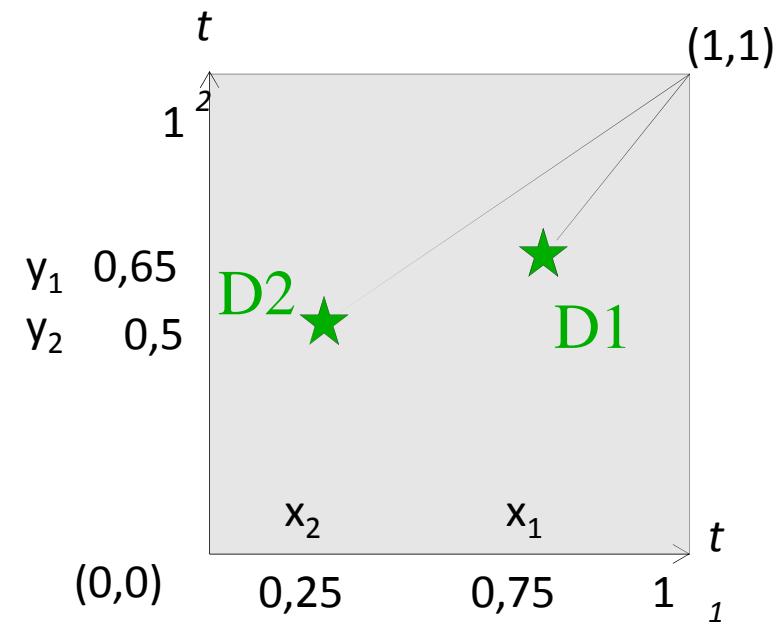


## Modèle booléen étendu : exemple (2/2)

$t_1 \text{ OR } t_2$



$t_1 \text{ AND } t_2$



$$RSV(\vec{D}, \overrightarrow{Q_{OR}}) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$RSV(\vec{D}, \overrightarrow{Q_{AND}}) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

# Modèle booléen étendu : formule finale

$$RSV(\vec{D}, \overrightarrow{Q_{OR}}) = \sqrt[p]{\frac{\sum_{i=1..m} c_m^p}{m}}$$

$$RSV(\vec{D}, \overrightarrow{Q_{AND}}) = 1 - \sqrt[p]{\frac{\sum_{i=1..m} (1 - c)_m^p}{m}}$$

avec :

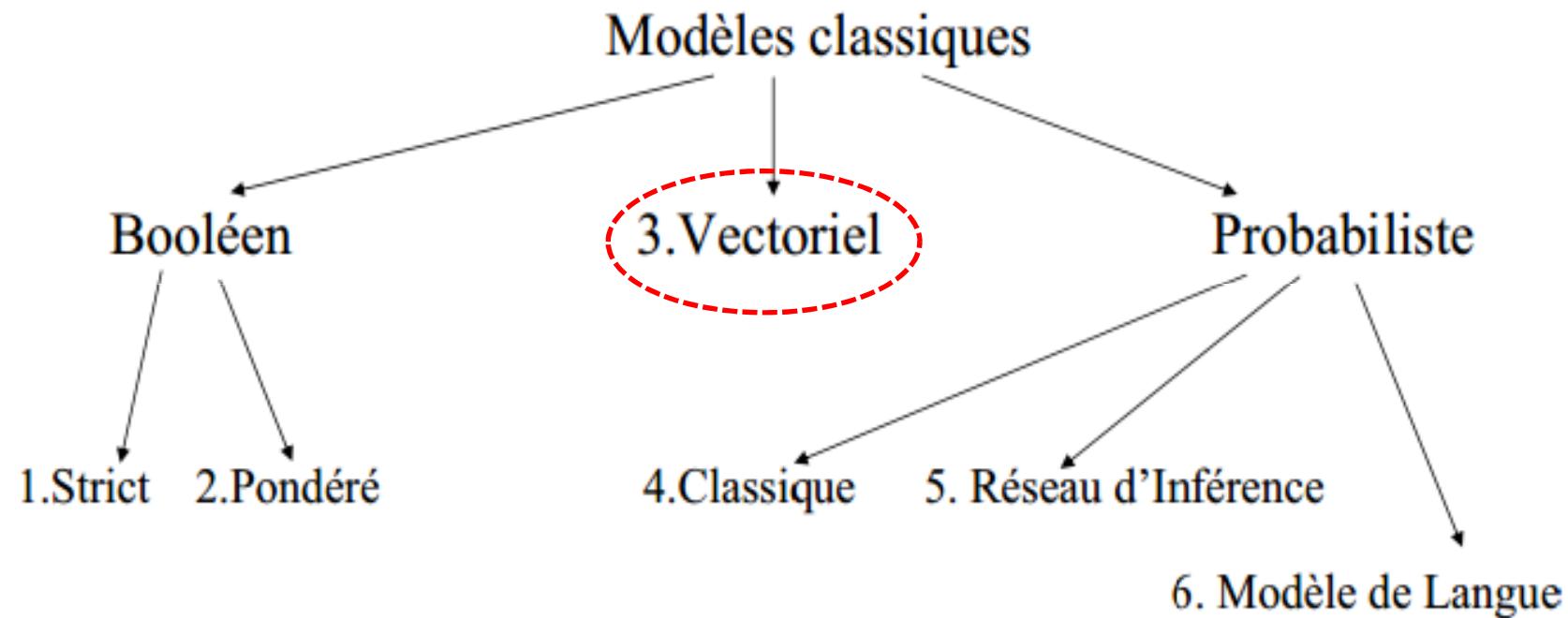
- $c$  les coordonnées des mots
- $m$  le nombre de termes de la requête
- $1 \leq p \leq \infty$

$\left. \begin{array}{l} p = 1 \rightarrow \text{modèle booléen classique} \\ p = 2 \rightarrow \text{exemple précédent} \end{array} \right\}$

## *Le modèle booléen pondéré : Bilan*

- Extension du modèle booléen en intégrant des pondérations (dénotant la représentativité d'un terme pour un document).
- Modèle de connaissances :  $T = \{t_i\}$ ,  $i \in [1, .. N]$   
Termes  $t_i$  qui indexent les documents
- Un document D est représenté par :
  - Une formule logique D (idem modèle booléen)
  - Une fonction  $W_D : T \rightarrow [0,1]$ , qui pour chaque terme de T donne le poids de ce terme dans D. Le poids vaut 0 pour un terme non présent dans le document.
  - Fonction de correspondance non binaire basée sur une similarité inspiré de la logique floue
  - Limitation : on ne tient pas compte dans la réponse de tous les termes de la requête.

# *Modèles algébriques*



## Modèle vectoriel (1/4)

- Modèle de connaissances :  $V = \{t_i\}$ ,  $i \in [1, .. N]$
- Tous les documents sont décrits suivant ce vocabulaire  $V$
- Un document  $D_i$  est représenté par un vecteur décrit dans l'espace vectoriel  $\mathbb{R}^N$  défini par  $V$ :

$$\vec{D}_i = \langle w_{i,1}, w_{i,2}, \dots, w_{i,j}, \dots, w_{i,N} \rangle$$

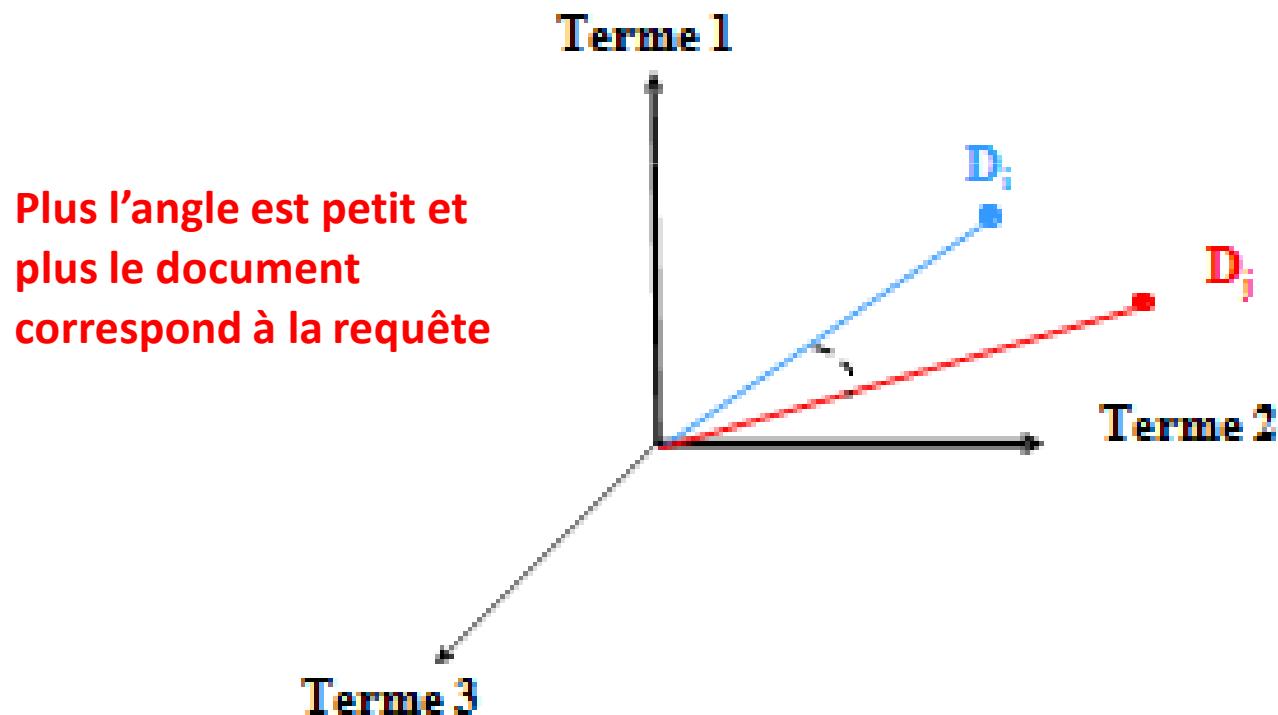
avec  $w_{i,j}$  le poids d'un terme dans un document

Une requête  $Q$  est représentée par un vecteur décrit dans l'espace vectoriel  $\mathbb{R}^N$  défini par  $V$ :

$$\vec{Q} = \langle w_{Q,1}, w_{Q,2}, \dots, w_{Q,j}, \dots, w_{Q,N} \rangle$$

## Modèle vectoriel (2/4)

Plus les vecteurs représentant les documents sont « proches », plus les documents sont similaires :



## *Modèle vectoriel (3/4)*

Comment trouver les poids des termes pour les documents ?

- Un document

- « Un violon est issu de bois précieux comme l’érable, palissandre, l’ébène... »

- Pour indexer, la première idée est de compter les mots les plus fréquents excepté les termes non significatifs comme « de », « avec », « comme »...

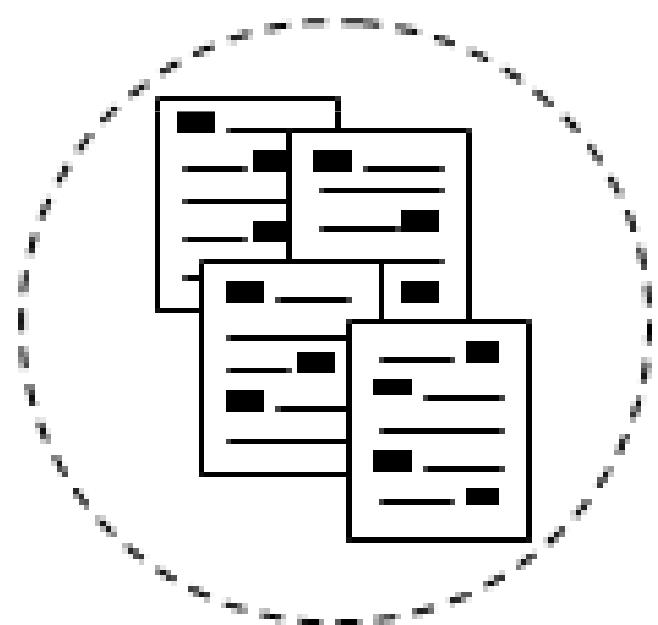
- « Un violon est composé de bois précieux comme l’érable, le palissandre, l’ébène... »

- Les termes en rouge sont ceux qui sont sélectionnés

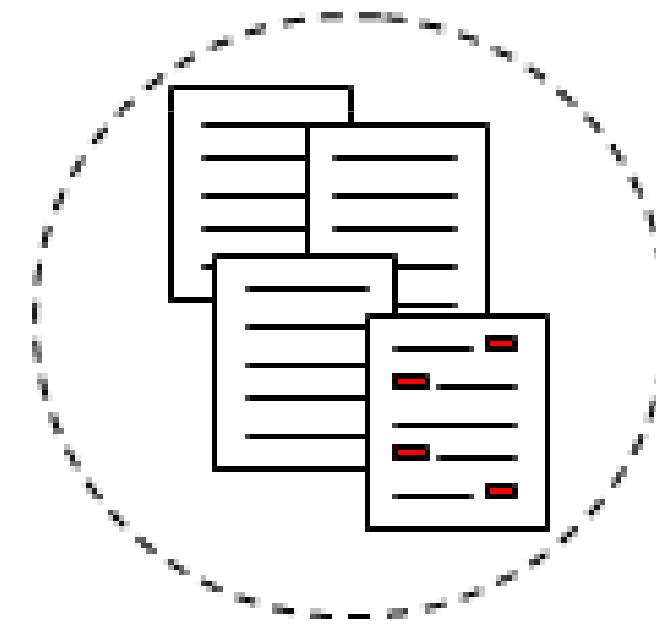
## *Modèle vectoriel (4/4)*

On tient compte du corpus (base de documents) entier, un terme qui apparaît beaucoup ne discrimine pas nécessairement les documents :

**Terme fréquent dans tout le corpus**



**Terme fréquent dans un seul document**



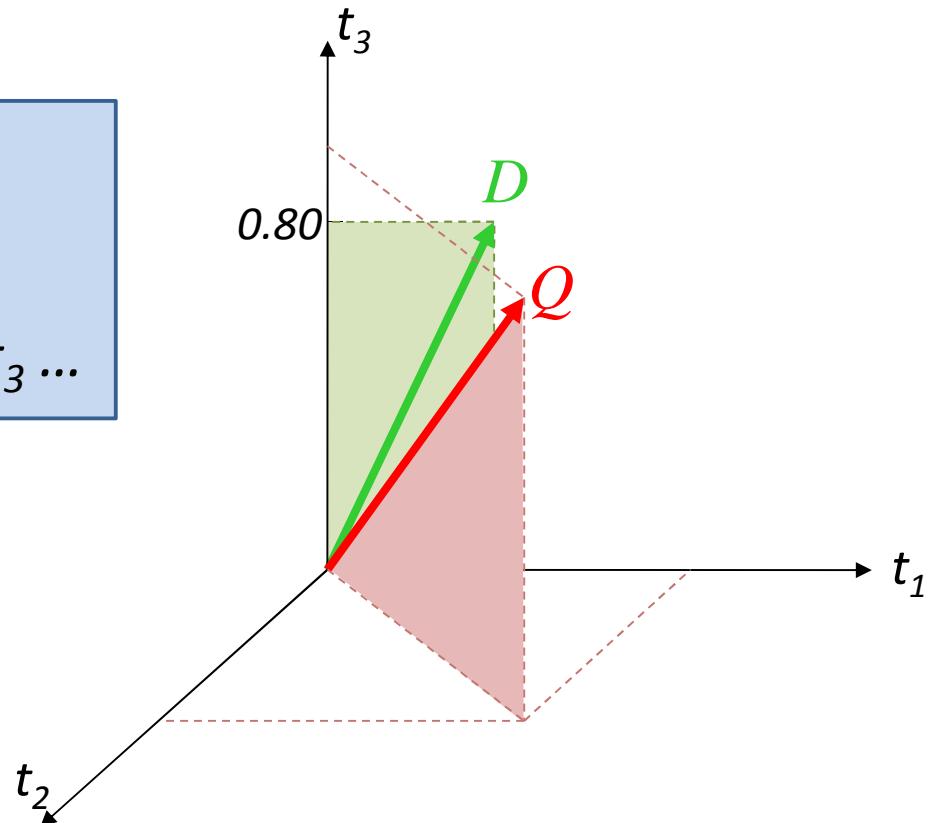
# Représentation

Requête  $Q$  :  $t_1 t_2 t_3$

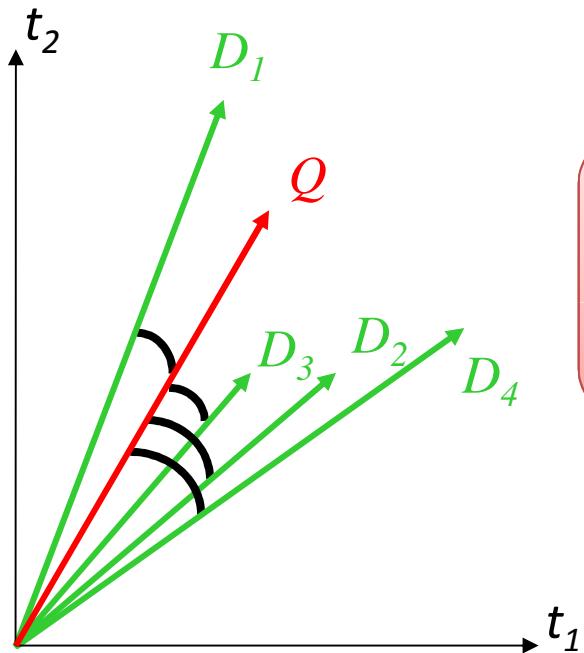
Document  $D$  :  $\dots t_1 \dots t_3 \dots$

Poids  $w_{D,t1} = 0.45$

Poids  $w_{D,t3} = 0.80$



# Quelle mesure de similarité ? (1/2)



**Cosinus**

$$\text{sim}(\vec{Q}, \vec{D}) = \frac{\vec{Q} \cdot \vec{D}}{|\vec{Q}| \times |\vec{D}|} = \frac{\sum_{i=1}^n w_{i,Q} \times w_{i,D}}{\sqrt{\sum w_{i,Q}^2} \times \sqrt{\sum w_{i,D}^2}}$$

(Le produit scalaire avec  
normalisation de la longueur  
des vecteurs)

## *Quelle mesure de similarité ? (2/2)*

- Autres mesures :

- Dice

$$RSV(\vec{Q}, \vec{D}) = \frac{2 \sum w_{iQ} \times w_{iD}}{\sum w_{iQ} + \sum w_{iD}}$$
$$\frac{2 |A \cap B|}{|A| + |B|}$$

- Jaccard

$$RSV(\vec{Q}, \vec{D}) = \frac{\sum w_{iQ} \times w_{iD}}{\sum w_{iQ} + \sum w_{iD} - \sum w_{iQ} \times w_{iD}}$$
$$\frac{|A \cap B|}{|A \cup B|}$$

- Overlap

$$RSV(\vec{Q}, \vec{D}) = \frac{\sum w_{iQ} \times w_{iD}}{\min(\sum w_{iD}, \sum w_{iQ})}$$
$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

## *Modèle vectoriel : Bilan*

- On représente la **requête** comme un **vecteur** (quelle pondération ?)
- On représente chaque **document** comme un **vecteur pondéré**
- On calcule la **similarité** (cosinus par exemple) entre chaque vecteur document et le vecteur requête
- On **ordonne** les résultats dans l'ordre inverse des scores obtenus
- On fournit les  **$k$  premiers résultats** à l'utilisateur

# *Modèle vectoriel : avantages et inconvénients*

- **Avantages :**

- Le langage de requête est plus **simple** (liste de mot-clés)
- Les performances sont meilleures grâce à la **pondération** des termes
- Le renvoi de documents à **pertinence partielle** est possible
- La fonction d'appariement permet de **trier** les documents

- **Inconvénients :**

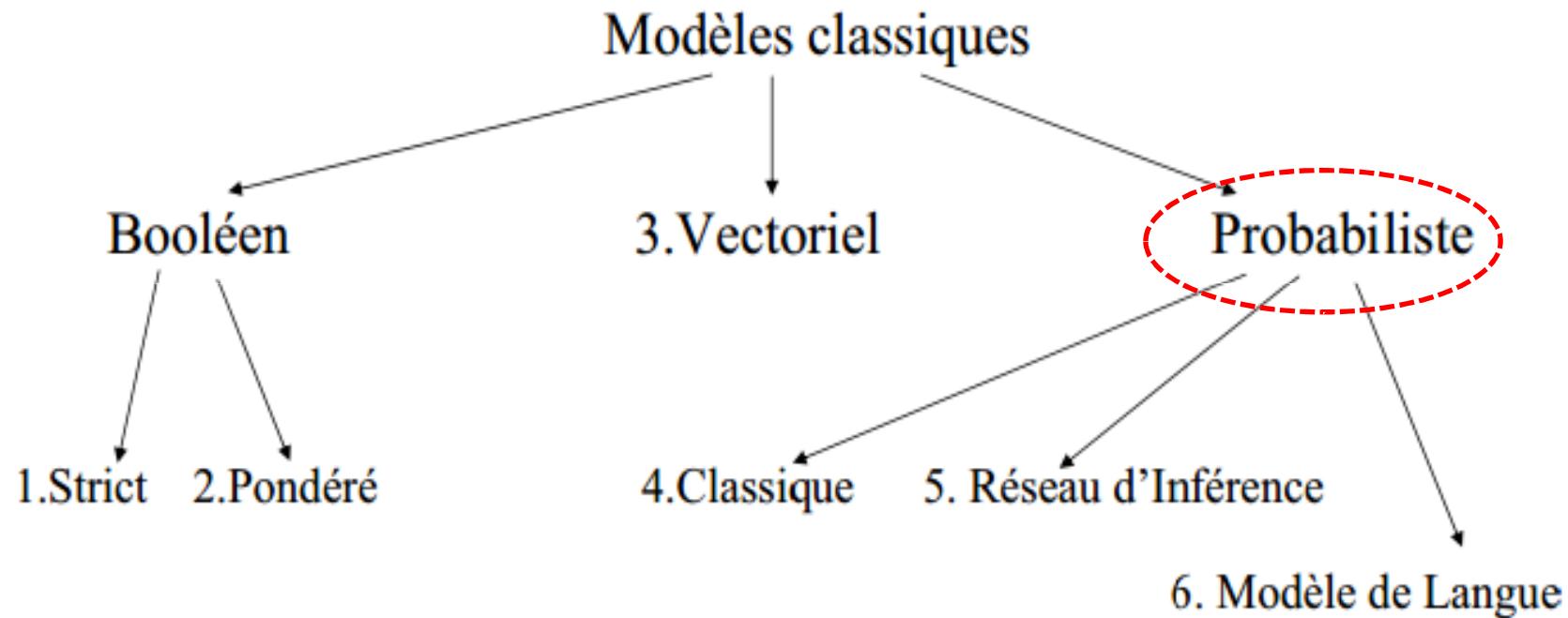
- Le modèle considère que tous les termes sont **indépendants** (inconvénient théorique)
- Le langage de requête est **moins expressif**
- L'utilisateur voit moins pourquoi un document lui est renvoyé

↳ Le modèle vectoriel est le plus populaire en RI

## *Autres modèles algébriques*

- Modèle vectoriel généralisé
  - Représente les dépendances entre termes
  - Théoriquement intéressant, mais efficacité non démontrée
- *Latent Semantic Indexing (LSI)*
  - Propose d'étudier les "concepts" plutôt que les termes, car ce sont eux qui relaient les idées d'un texte.
  - Lie les documents entre eux et avec la requête
  - Permet de renvoyer des documents ne contenant aucun mot de la requête
  - Moins de dimensions

# *Modèles probabilistes*



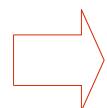
## *Modèle probabiliste classique*

- Le modèle probabiliste classique consiste à calculer la pertinence d'un document en fonction de pertinences connues pour d'autres documents.
- Ce calcul se fait en estimant la pertinence de chaque index pour un document et en utilisant le **Théorème de Bayes** et une règle de décision

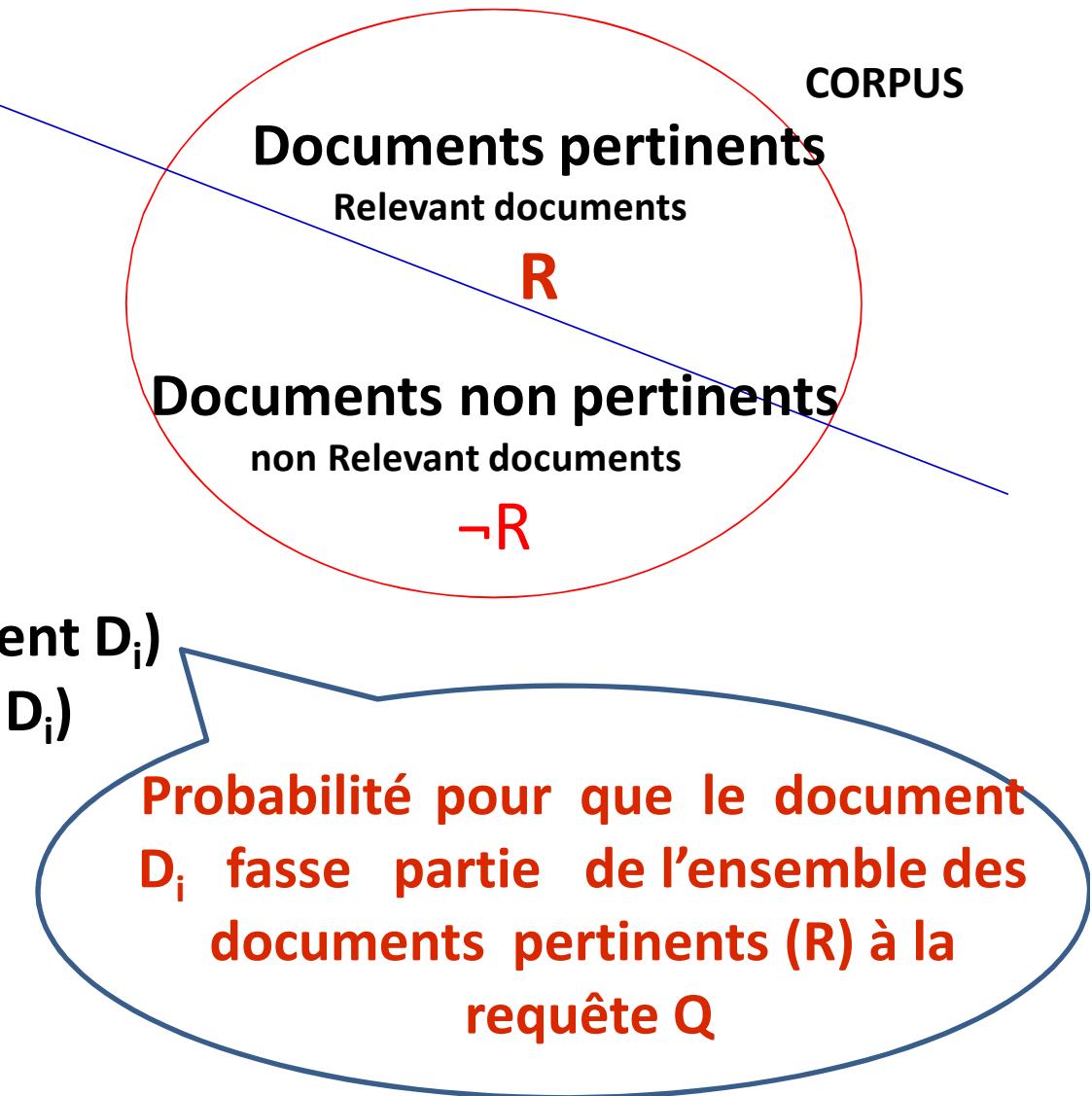
# Modèle probabiliste classique

$$\text{Corpus} = R \cup \neg R$$
$$\neg R \cap R = \emptyset$$

Pour une requête Q



$P(\text{pertinence}_Q / \text{document } D_i)$   
notée simplement  $P(R / D_i)$



# Modèle probabiliste (1/3)

- Estimation de la probabilité de pertinence d'un document par rapport à une requête
  - *Probability Ranking Principle (Robertson 77)*
  - $R : D$  est pertinent pour  $Q$
  - $\neg R : D$  n'est pas pertinent pour  $Q$
  - Le but : estimer
    - $P(R/D)$  : probabilité que le document  $D$  soit contienne de l'information pertinente pour  $Q$
    - $P(\neg R/D)$
- variables indépendantes,  
deux ensembles de  
documents séparés

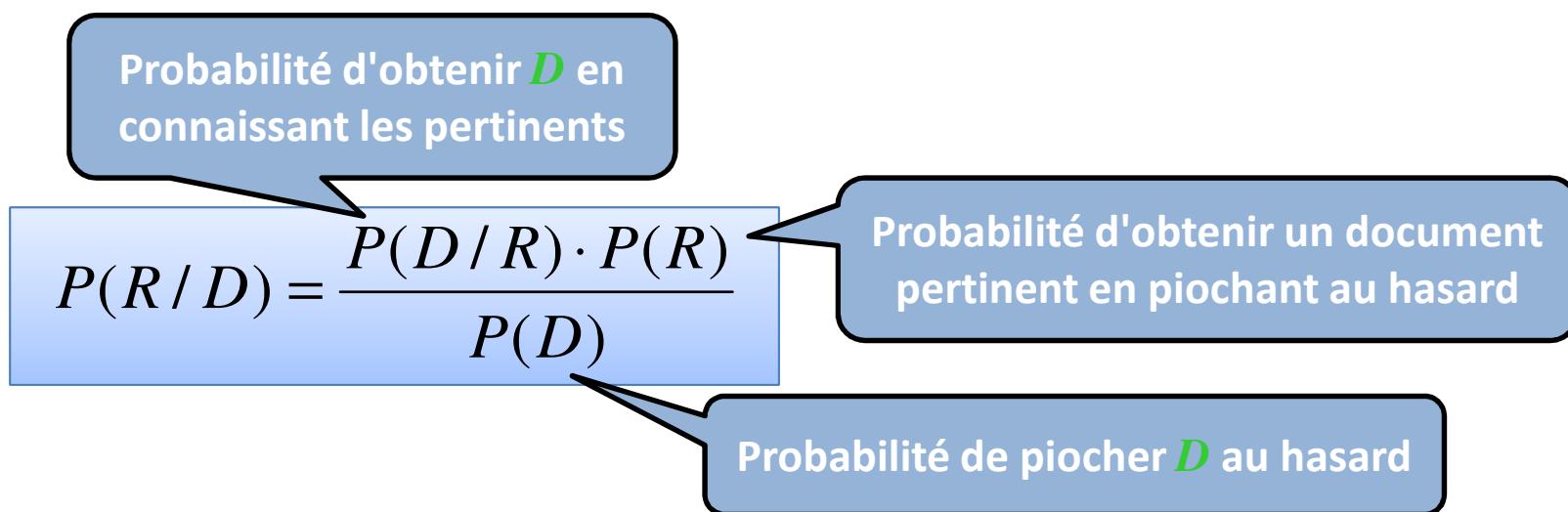
$$\text{si } \frac{P(R/D)}{P(\neg R/D)} > 1 \text{ ou si } \log \frac{P(R/D)}{P(\neg R/D)} > 0 \text{ alors } D \text{ est pertinent}$$

## Modèle probabiliste (2/3)

- Rappel du théorème de **Bayes** :

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

- On ne sait pas calculer  $P(R/D)$ , mais on peut calculer  $P(D/R)$



## *Modèle probabiliste (3/3)*

- En utilisant l'hypothèse d'indépendance des termes :

$$P(D / R) = \prod_{i=1}^n P(t_i \in D / R)$$

- Pour estimer les probabilités sur les termes, on peut utiliser des requêtes déjà résolues (apprentissage) puis des pondérations
- Exemple (système Okapi) :
  - le *tf.idf*
  - la longueur du document
  - la longueur moyenne des documents

# *Modèle probabiliste : conclusion*

- Deux modèles phares :
  - 2-poisson
  - Okapi BM25
- Autres modèles de type probabiliste :
  - Réseaux bayésiens
  - Modèle de langue
- Conclusion :
  - Problème des probabilités initiales
  - Termes indépendants
  - Résultats comparables à ceux du modèle vectoriel

# *Quelques outils*

- lucy/zettair <http://www.seg.rmit.edu.au/zettair/>
- cheshire <http://cheshire.lib.berkeley.edu/>
- dataparksearch engine <http://www.dataparksearch.org/>
- lemur <http://www.lemurproject.org/>
- **lucene** (et **solr**) <http://jakarta.apache.org/lucene/docs/>
- **terrier** <http://ir.dcs.gla.ac.uk/terrier/>
- wumpus <http://www.wumpus-search.org/>
- xapian <http://www.xapian.org/>

*liste et liens sur <http://www.emse.fr/~mbeig/IR/tools.html>*

# Relevance feedback

# *Relevance feedback (1/2)*

- "Réinjection de la pertinence"
- Hypothèse : la requête initiale de l'utilisateur n'est pas la requête idéale pour obtenir les documents qu'il cherche
- But : **déplacer le vecteur de la requête** pour la rapprocher des documents pertinents



## *Relevance feedback (2/2)*

- "**Manuel explicite**" :

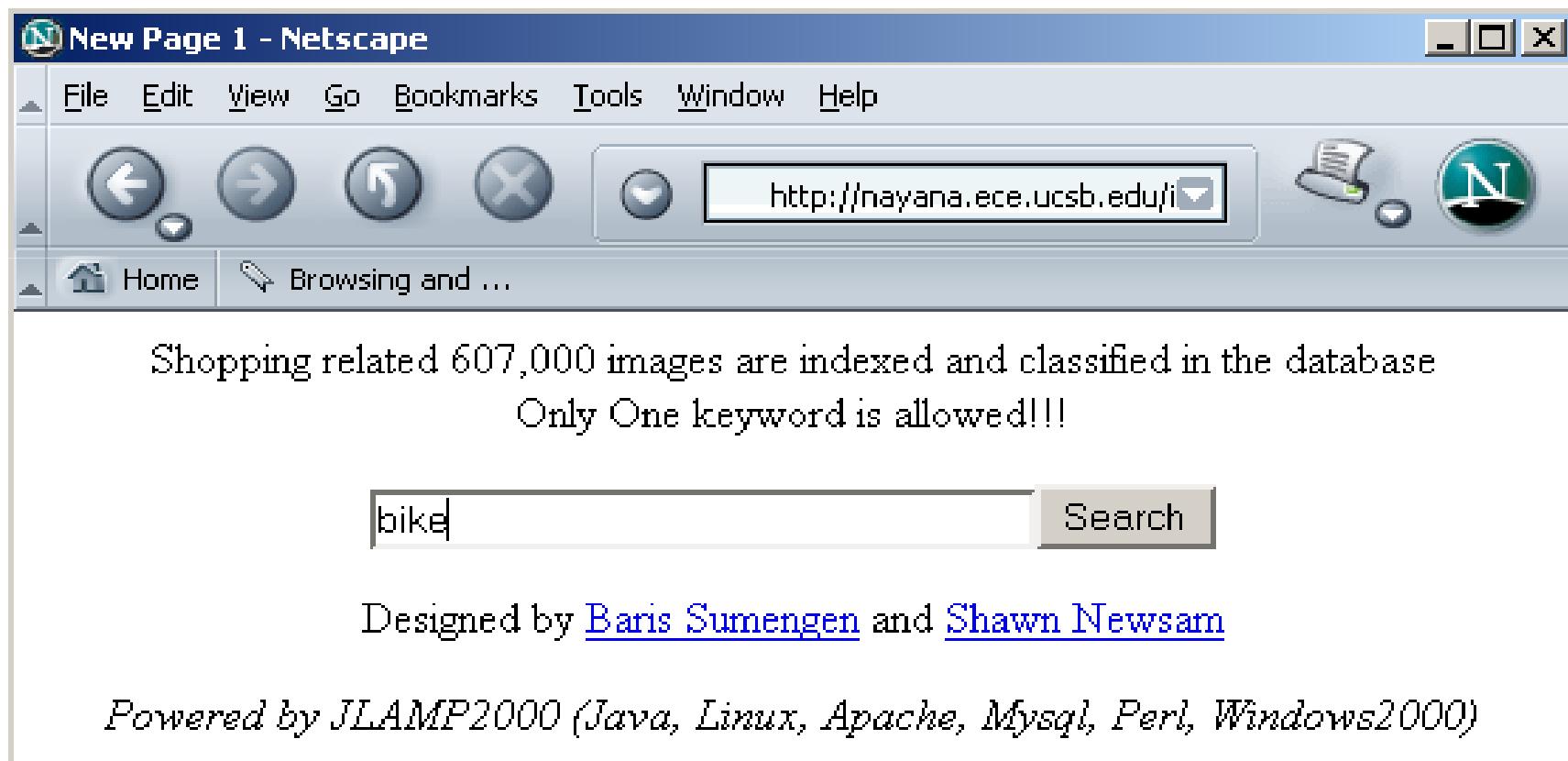
- L'utilisateur visualise les  $n$  premiers résultats
- Il estime la pertinence de chacun (0 ou 1)
- Nouvelle requête obtenue à partir des documents jugés pertinents et non pertinents

- **Automatique** (*blind relevance feedback*) :

- Les  $k$  premiers résultats du premier *run* sont supposés pertinents ( $k$  souvent fixé à 5, 10 ou 20)
- Même processus que pour le *relevance feedback* manuel (sans les documents non pertinents)

# Relevance Feedback: Exemple

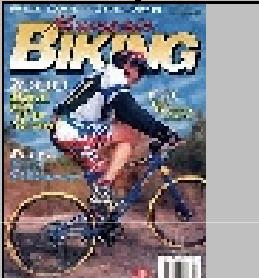
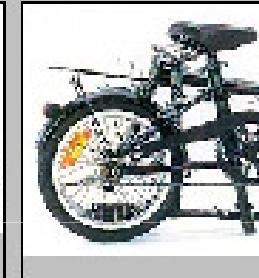
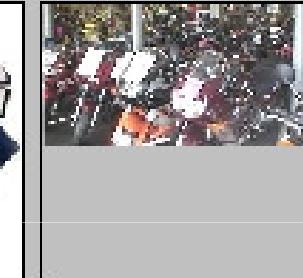
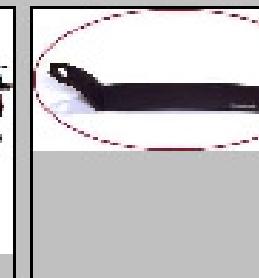
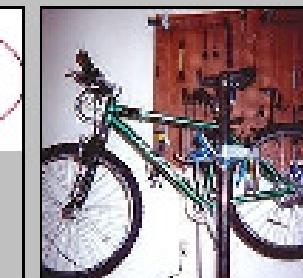
- <http://nayana.ece.ucsb.edu/imsearch/imsearch.html>



# Résultats de la requête initiale

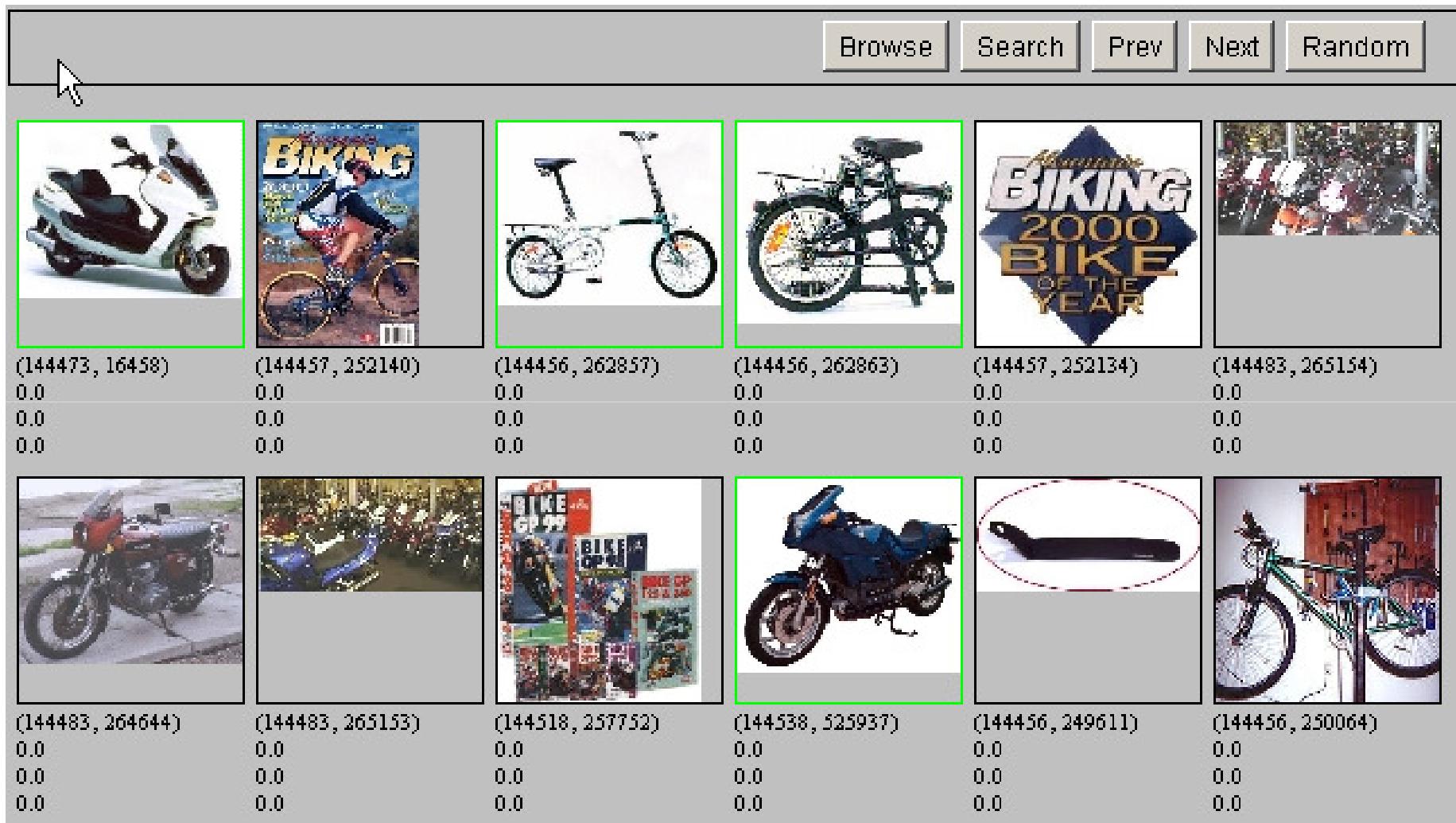
Browse   Search   Prev   Next   Random



					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0

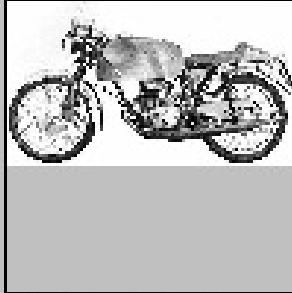
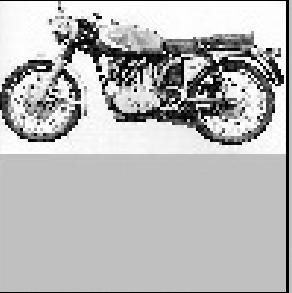
# Relevance Feedback

Browse   Search   Prev   Next   Random



(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0

# Résultats après Relevance Feedback

						Browse	Search	Prev	Next	Random	
						(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
						(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

# Rocchio 1971 Algorithm (SMART)

$$\vec{o'} = \alpha \vec{o} + \beta \vec{P} + \gamma \vec{NP}$$

moyenne des vecteurs  
des documents non pertinents  
valeur négative (ex : -0,25)

moyenne des vecteurs  
des documents pertinents  
valeur positive (ex : 0.5)

vecteur requête initial  
valeur positive supérieure aux autres (ex : 1)

nouveau vecteur requête

## Formule de Rocchio : exemple

$$\vec{o'} = \alpha \vec{o} + \beta \vec{P} + \gamma \overrightarrow{NP}$$

$$\left\{ \begin{array}{l} \vec{o} = (5,0,3,0,1) \\ \vec{P} = (2,1,2,0,0) = D_1 \\ \overrightarrow{NP} = (1,0,0,0,2) = D_2 \end{array} \right.$$

$$\begin{aligned}\vec{Q}' &= \vec{Q} + \frac{1}{2} \vec{P} - \frac{1}{4} \overrightarrow{NP} \\ \vec{Q}' &= (5.75, 0.5, 4, 0, 0.5)\end{aligned}$$

cosinus	D1	D2
Q1	0,90	0,53
Q2	0,95	0,43

# *Recherche multimédia*

- Texte et/ou image et/ou audio et/ou vidéo...
- Des collections très volumineuses :
  - ex : collection Wikipédia pour INEX
  - 4.6 Go en texte seul, 60 Go avec les images
- Documents structurés (MPEG-7...)
- Utilisation :
  - des métadonnées
  - du texte "environnant" les images (légende, point de référence...)
  - des caractéristiques propres des documents autres que le texte :
    - Analyse d'image
    - Speech-to-text
    - ...

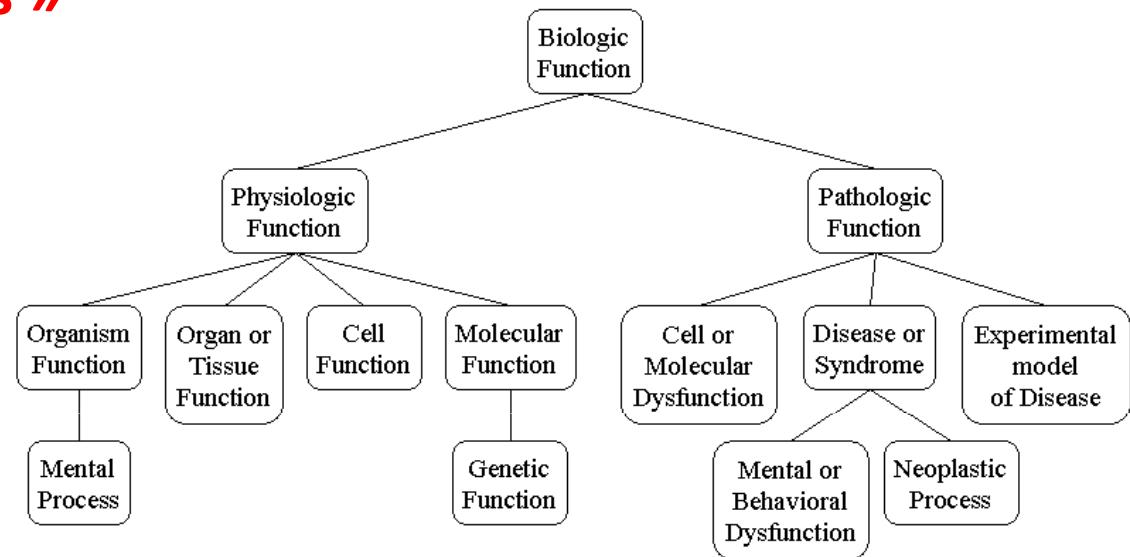
# *Expansion de la requête*

- Ajouter des mots pertinents à la requête initiale et les pondérer efficacement
- Méthodes pour palier les problèmes liés au langage naturel
  - « bateau » ne ramène pas le mot « navire »
  - « thermodynamique » ne ramène pas « chaleur »
  - « félin » ne ramène pas « chat »
  - ...
- Le *relevance feedback* sert aussi à ça (en partie)

Pourquoi ?

# *Expansion de la requête*

- Les **thesaurus « manuels »**



- Les **thesaurus automatiques** (voir page suivante)
- **L'analyse des logs** de requêtes

# *Génération automatique de thesaurus*

- Fondée sur la **similarité entre deux mots**
- **Co-occurrence de deux mots** : deux mots qui apparaissent fréquemment ensemble possèdent une relation sémantique entre eux
  - Ex: « location » et « appartement »
  - Conduit à des **relations sémantiques** non spécifiées
- **Co-occurrence des contextes** : deux mots sont similaires s'ils co-ocurrent avec des mots similaires
  - Ex: « bateau » et « navire », « chat » et « félin », mais aussi « chat » et « chien », « PS » et « UMP », etc.
  - Conduit plutôt à des relations lexicales de **synonymie** ou **hyperonymie**, mais peut également être plus large
  - Possibilité d'utiliser les **relations syntaxiques** également

# *Génération automatique de thesaurus*

- chat → animal de compagnie, siamois, client IRC, persan, chien, ...
- télévision → TV, séries, programme, radio, images, ...
- Expansion de requêtes à base de thesaurus :
  - Ajouter les mots jugés similaires à la requête
  - Éventuellement, donner des pondérations en fonction du niveau de similarité
- Quand s'arrête-t-on d'étendre la requête ?

Quels sont les effets de ces expansions de requêtes sur la précision et le rappel ?

**Le Web,  
qu'est-ce que ça change ?**

# Qu'est-ce que le Web ?

- Environ 250 millions de sites Web
- Plus de 1000 milliards de pages Web (selon Google)
- Bientôt 2 milliards d'humains connectés
- Beaucoup, beaucoup d'information

## WikiLeaks: 250'000 notes «confidentielles» de la diplomatie US

**WASHINGTON** | Le contenu de 250'000 câbles diplomatiques américains dévoilés par le site WikiLeaks a été publié dimanche par les grands titres de la presse mondiale, révélant les dessous de la diplomatie des Etats-Unis, comme lorsque Ryad appelait Washington à attaquer l'Iran.



operations is staggering. Rothschild said Facebook manages more than 25 terabytes of data per day in logging data, which he said was the equivalent of about 1,000 times the volume of mail delivered daily by the U.S. Postal Service.

## Wikipédia: Un million d'articles en français

Née en mars 2001, la version en français de l'encyclopédie en ligne Wikipédia a franchi dans la semaine du 20 au 27 septembre 2010 le seuil symbolique du **million d'articles**<sup>1</sup>. Après la version en anglais (qui compte actuellement 3,4 millions d'articles) et la version en allemand (1,1 million), la version en français est la troisième à franchir cette barre. Les prochaines versions qui compteront un million d'articles pourraient être celles en italien et en polonais.



## 25 Terabytes of Log Data – Daily

The amount of log data amassed in Facebook's operations is staggering. Rothschild said Facebook manages more than 25 terabytes of data per day in logging data, which he said was the equivalent of about 1,000 times the volume of mail delivered daily by the U.S. Postal Service.

# *Du contenu (un peu) structuré*

- Des métadonnées

```
<head>
  <title>Chiraz Latiri, page personnelle (home page)</title>
  <meta name="description" lang="fr" content="Page personnelle de
                                                Chiraz Latiri"/>
  <meta name="description" lang="en" content=<< Chiraz Latiri
                                                Home Page"/>
  <meta name="author" content=<< Chiraz Latiri "/>
  <meta name="keywords" lang="fr" content="informatique,
                                                enseignement, Recherche d'Information, Fouille de
                                                données, Ingénierie des connaissances"/>
...
...
```

# *Du contenu (un peu) structuré*

- Des **liens**
  - Un complément d'**information**
  - Une dimension **sociale**

The screenshot displays a user profile for "Chiraz Latiri" on a platform that integrates social networking features. At the top, there's a navigation bar with "Fresh Bookmarks", "Hotlist", and "Explore" buttons. Below the navigation, a message reads: "The freshest bookmarks that are flying like hotcakes on Delicious and beyond." A "See more recent bookmarks" link is also present.

On the right side of the interface, there is a network graph visualization. The graph consists of six nodes (A, B, C, D, E, F) represented by colored circles (blue, red, orange, green, yellow, purple). Node B (red) has the highest value of 38.4. Node C (orange) has a value of 34.3. Node E (yellow) has a value of 8.1. Node D (green) has a value of 3.9. Node F (purple) has a value of 3.9. Node A (blue) has a value of 3.3. Edges connect node B to nodes A, C, D, E, and F. Nodes D, E, and F are interconnected among themselves. Nodes A, B, and C are interconnected among themselves.

**OpenGraph protocol**

A diagram illustrating the OpenGraph protocol. It shows a Facebook "Like" button on a movie page for "The Rock". Dashed arrows indicate how the page's content is being shared or embedded via the protocol. To the right, a snippet of the movie's page on IMDb is shown, featuring a thumbnail for "The Rock", the title, the number of likes (12,932), and a comment from "Francis Luu" who liked it 5 minutes ago.

Below these examples, a YouTube video thumbnail for "Fix Your Windows Vista PC Problems in one click" is shown with a "via youtube.com" link and "SAVE | SHARE" buttons.

# Du contenu (un peu) structuré

- Des indications de **forme**

- Titres (h1, h2...)
- Caractères gras, italiques, soulignés
- Couleurs
- Listes...

## Composantes [\[modifier\]](#)

### Prétraitements [\[modifier\]](#)

La première étape en recherche d'information est d'établir ces techniques permettant de passer d'un **document** textuel à une représentation plus pratique, la manière d'indexer limitant ou favorisant les possibilités de recherche.

- Il faut extraire d'un texte un ensemble de **descripteurs**. Ceux-ci sont la plupart du temps (après suppression des mots grammaticaux) les termes qui apparaissent dans un document, souvent transformés (**lemmatisation**, ...)
- À l'aide de ce jeu de descripteurs, il est possible de représenter le **document** par un vecteur dans l'espace des termes. Il est alors possible de faire des comparaisons entre documents et de déterminer leur similitude.

### Recherche [\[modifier\]](#)

Une fois les documents transformés, il est possible de rechercher ceux qui répondent le mieux à une question d'un utilisateur. Plusieurs approches sont possibles :

- *L'approche ensembliste* qui considère que l'ensemble des documents s'obtient par une série d'opérations (intersection, union, différence, etc.). La requête SQL1 correspond à cette approche dite aussi de logique de premier niveau.
- *L'approche algébrique* (ou vectorielle) qui considère que les documents et les questions font partie d'un même **espace vectoriel**.
- *L'approche probabiliste* qui essaie de modéliser la notion de *pertinence*.

Il est enfin possible d'utiliser des modèles capables d'interagir avec l'utilisateur, afin d'améliorer petit à petit les réponses du système en fonction des réactions de l'utilisateur, et en indiquant à chaque fois les documents pertinents pour sa question. Ces indications peuvent aussi servir pour améliorer globalement les résultats de recherche.

# Du contenu (un peu) structuré

- Du **balisage "sémantique"** pour l'application de feuilles de style

```
<h1>La SNCF toujours à l'heure sur la hausse des prix</h1>
<h4>décryptage</h4>
<p class="sstitre">Poussé par le gouvernement, [...]
<p class="clear"><a href="http://www.liberation.fr/economie/01
sur-la-hausse-des-prix" class="reactions" title="Réaction à l'ar
hausse des prix">2 réactions</a></p>
<p class="clear authors">Par <strong>CATHERINE
<img class="" alt="Des TGV à la gare de Lyon, à Paris.">
src="http://q.liberation.fr/photo/id/214862/r/03/02/w/459/m/1291632570">
<p class="legende"> Des TGV à la gare de Lyon, à Pa
class="teaseTitle"> Le nouveau secrétaire d'Etat au
<div id="libe-access-block"><div id="bloc-
class="bloc-article-abonne-container"> <s>
cet article est réservée aux abonnés LIBE+. <a href="http://
class="color">Abonnez-vous</span></a> dès maintenant!
```

**ÉCONOMIE** 06/12/2010 À 00H00

## La SNCF toujours à l'heure sur la hausse des prix

DÉCRYPTAGE ➔ Poussé par le gouvernement, qui s'est dit ce week-end favorable à une augmentation «raisonnable» dès janvier, le groupe poursuit sa politique de grand écart tarifaire.

2 réactions

Par CATHERINE MAUSSION



Des TGV à la gare de Lyon, à Paris. (© AFP Loic Venance)

Le nouveau secrétaire d'Etat aux Transports, Thierry Mariani, a affirmé ce week-end être favorable à une hausse, dès janvier, «tout à fait raisonnable» du billet de train. Sera-t-elle aussi modérée qu'il le promet ? Pourquoi cette hausse ? Rituellement, chaque année, le prix du billet

La lecture de cet article est réservée aux abonnés LIBE+. **Abonnez-vous** dès maintenant, à partir de 6€ par mois.

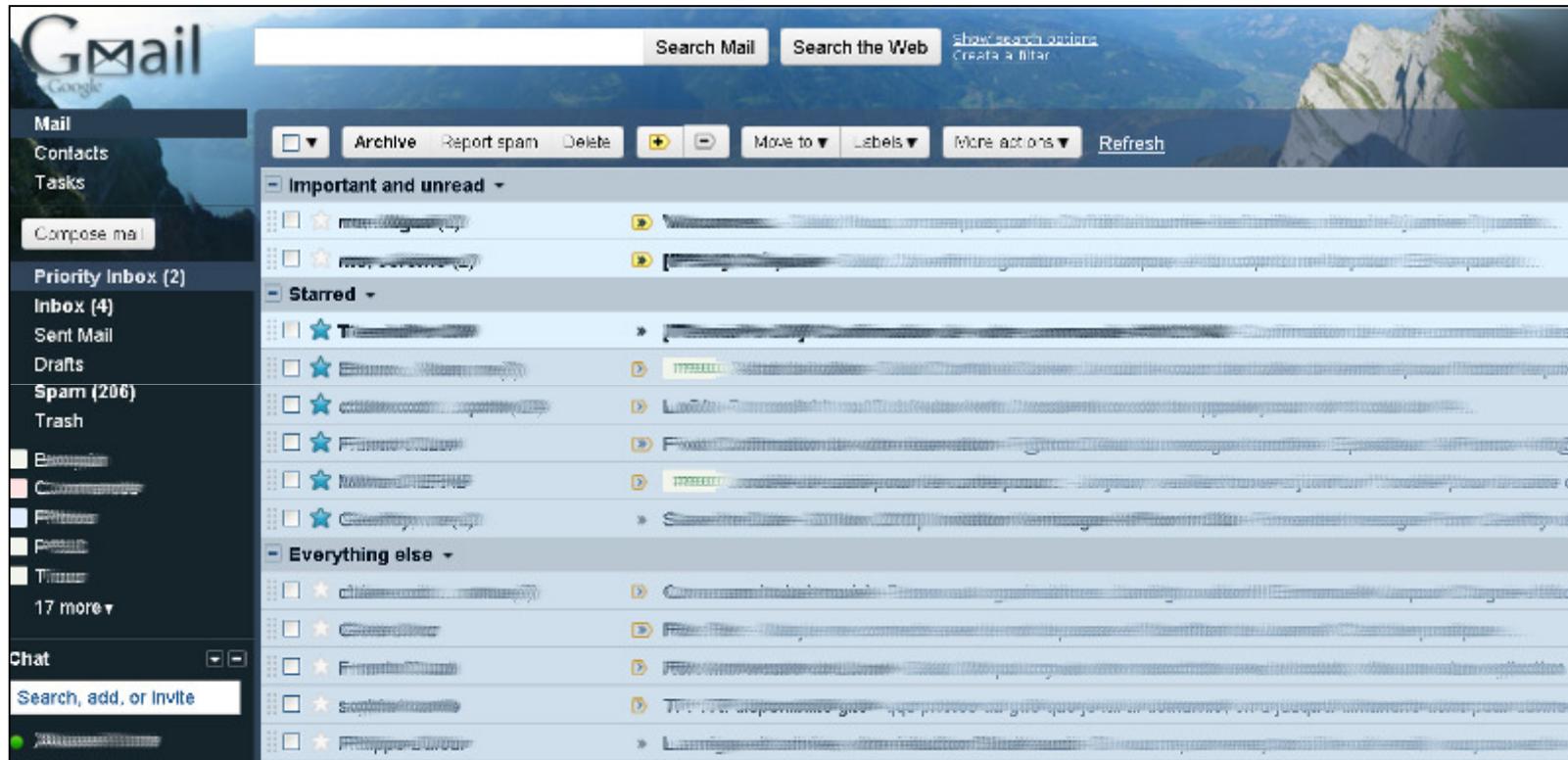
# Un contenu dynamique

The screenshot shows a Google search results page for the query "liliane bettencourt". The search bar at the top contains the text "liliane bettencourt". Below it, a link to "Actualités correspondant à liliane bettencourt" is shown. The results include several news items and tweets. Three specific items are circled in red:

- A news item from "nouvelobs.com" about an agreement between Liliane Bettencourt's daughter and François-Marie Banier, with a timestamp "Il y a 11 minutes".
- A tweet from "LeVolontaire" (@ToutelaPresse) about the end of the legal battle, with a timestamp "Publié il y a 1 minute".
- A tweet from "Twitter" (@liberation\_fr) about the end of the legal battle, with a timestamp "Il y a 2 minutes".

The left sidebar of the Google interface is also visible, showing options like "Tout", "Images", "Vidéos", "Actualités", "En temps réel", "Plus", "Paris", and "Date indifférente" (which is circled in green). The "Actualités" section is currently selected.

# *Un contenu dynamique*



# Un contenu dynamique

[Modifier la recherche](#)

Paris (FR) » Buenos Aires (AR) ven. 14 janvier 2011 - ven. 21 janvier 2011

[Ajouter un hôtel](#) [Ajouter une voiture](#)

**268 résultats** Taxes et frais inclus pour tous les passagers en EUR

Trier par : Prix total   Tableau escales / compagnies Page : < >

Prix	Airline	Type d'escale	Départ	Arrivée
848 €	TAM Lin Aer	1 escale (20h 10m)	20:00 CDG	12:10 EZE
854 €	TAM Lin Aer	1 escale (24h 05m)	05:55 AEP	10:00 CDG
859 €	TAM Lin Aer	2 escales (19h 50m)	20:20 CDG	12:10 EZE
933 €	Delta Air Lines	1 escale (22h 00m)	13:50 CDG	07:50 EZE

**Vols Paris - Buenos Aires à partir de 289 €**

[www.edreams.fr/buenos-aires](http://www.edreams.fr/buenos-aires) Réservez vite sur Edreams

**Publicité**

**Choisir mes critères**

Mon budget

Avec ou sans escale ?  
 vols directs  
 1 escale  
 2 escales ou plus

Durée  
13h 30m - 43h 40m

Horaires du vol aller  
 Décollage  Atterrissage  
Décollage: 07:10 - 23:20

Horaires du vol retour

# Un contenu redondant

Google françois hollande

**Recherche** Environ 143 000 résultats (0,17 secondes)

Web [François Hollande pose les bases de son mandat sur un premier ...](#) Les Échos - il y a 2 heures

Images [Jusqu'au bout, François Hollande aura pris le contre-pied de Nicolas Sarkozy. ...](#)

Maps [François Hollande n'hésite pas à revendiquer l'effort de redressement « le plus ...](#)

Vidéos [Les principaux points du Budget 2013](#) L'Expansion

Actualités [Budget 2013. Enseignement : 10.011 créations de postes en 2013](#) Le Télégramme

[Challenges.fr](#)

[Autres articles \(113\) »](#)

Shopping [Copé sur Hollande et Ayraut : "on cherche en vain les hommes d'Etat"](#)

Plus [Le Point - il y a 4 heures](#)

Sur France 2, le député-maire de Meaux a commenté : "les prestations télévisées de François Hollande et de Jean-Marc Ayrault se succèdent ...

À la une [Copé et l'UMP dénoncent le renvoi par Hollande à l'héritage ...](#) Le Parisien

[- Copé à propos de Hollande sur l'héritage : "pas une réflexion d ...](#) 20minutes.fr

[Le Figaro - Europe1](#)

[Autres articles \(606\) »](#)

Recherche sur le Web

Rechercher les pages en Français

Toute l'actualité

[Mondial de l'auto: François Hollande prend le pouls d'un secteur en ...](#)

Le Point - il y a 1 heure

Le Mondial de l'automobile de Paris accueille vendredi le président François Hollande, qui pourra prendre le pouls d'un secteur en crise dont ...

[Mondial de l'automobile : Hollande très attendu](#) Le Nouvel Observateur

[Autres articles \(234\) »](#)

 DES PAROLE  
DES ACTES  
  
L'Express

 Copé sur Hollande et Ayraut : "on cherche en vain les hommes d'Etat"  
Le Point

 Mondial de l'auto: François Hollande prend le pouls d'un secteur en ...  
Le Nouvel Observateur

# *Un contenu souvent non informatif*

Menus  
Liens  
Index  
  
Publicités

**Le Point.fr**

**EN CONTINU**  
→ 24h d'info  
→ Flux RSS  
→ Mobile  
→ Newsletters

**LE MAGAZINE**  
→ Sommaire  
→ Abonnement  
→ Édition digitale  
→ Nos hors-séries

**LES SERVICES**  
→ Météo  
→ Bourse  
→ Jeux-Concours  
→ Programme télé

Rechercher sur le site

**ACTUALITÉ** POLITIQUE ÉCONOMIE TECH & NET SANTÉ SPORTS CULTURE ART DE VIVRE TUTÉRIAPOLIS Le Point TV Diaporamas

Actualité | Monde | **Société** | Médias | Sciences | Villes | Insolites | Confidentiels | Ces gens-là | Chroniques | Échos | GUIDE DU VIN

Australie - Nouvelle-Zélande à partir de **1162 € TTC**

**ACTUALITÉ** > Société

Le Point.fr - Publié le 09/12/2012 à 08:38 - Modifié le 09/12/2012 à 09:08

## Miss Bourgogne devient Miss France 2013

Marine Lorphelin, étudiante en médecine de 19 ans, rêve de devenir obstétricienne ou pédiatre.



**Société**

- 09h23 Un homme blessé dans le bois de Vincennes : un suspect interpellé
- 09h01 Négociation emploi: le recul de la précarité n'est pas abordé, regrette...
- 08h32 Le déficit commercial recule à 4,3 milliards d'euros en novembre
- 08h21 Le décret sur les rythmes scolaires examiné mardi au CSE
- 08h15 Affaire du viol collectif en Inde: cinq suspects comparaissent à huis clos
- 08h06 Maisons de retraite : qui doit payer la facture ?

**Société : l'actualité en direct** 55

185

# *Un contenu non contrôlé*

- Tout le monde peut à la fois être **lecteur** et **producteur** d'info
  - Ajouter son propre contenu au Web est devenu simple et gratuit
  - Pages perso, blogs, wikis, forums, listes de diffusion...
  - Les institutions et les particuliers sont *a priori* sur le même pied
- Le Web fourmille d'**informations fausses**
- Les **métadonnées** sont peu utilisées
- Les créateurs de pages peuvent modifier le contenu pour améliorer leur classement sur les moteurs de recherche
  - **Répétition de mots-clés** dans des couleurs non visibles ou dans les métadonnées
  - **Spamming** : pas de contenu mais une énumération de mots-clés destinés uniquement à être visible sur les moteurs de recherche

# *Un contenu non contrôlé*

- De la **structure**
  - Formulaires, menu : [\[HTML\]](#), [\[TEXTE\]](#)
  - Tableaux : [\[HTML\]](#), [\[TEXTE\]](#)
- Du **contenu sans information**
  - [\[HTML\]](#)
- Du **spam**
  - [\[HTML\]](#)

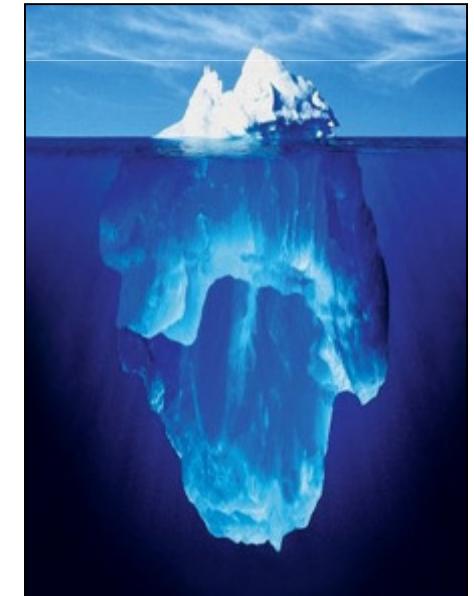
# *Un contenu « caché »*

- « Web profond »
  - Les sites "non- liés" (Web opaque)
  - Le contenu à accès limité
  - La plupart des pages dynamiques

↳ La majeure partie du Web n'est pas indexée

*(Web profond = 500 fois le Web de surface selon BrightPlanet 2001)*

↳ Protocole *sitemap* de Google



# *Des besoins d'utilisateurs différents*

- **Besoins d'information**

- Accès à la **connaissance** sur un sujet
- « *Chevaliers paysans de l'an mil au lac de Paladru* »
- Entre 40 et 65 % des requêtes

- **Besoins de navigation**

- Recherche d'une **page précise**
- « *Facebook* », « *Le Monde* »
- Entre 15 et 25 %

- **Besoins de transaction**

- Recherche d'un **service** opéré sur le Web
- « *billets SNCF* », « *Canon EOS 550 D* », « *météo Argentine* »
- Entre 20 et 35 %

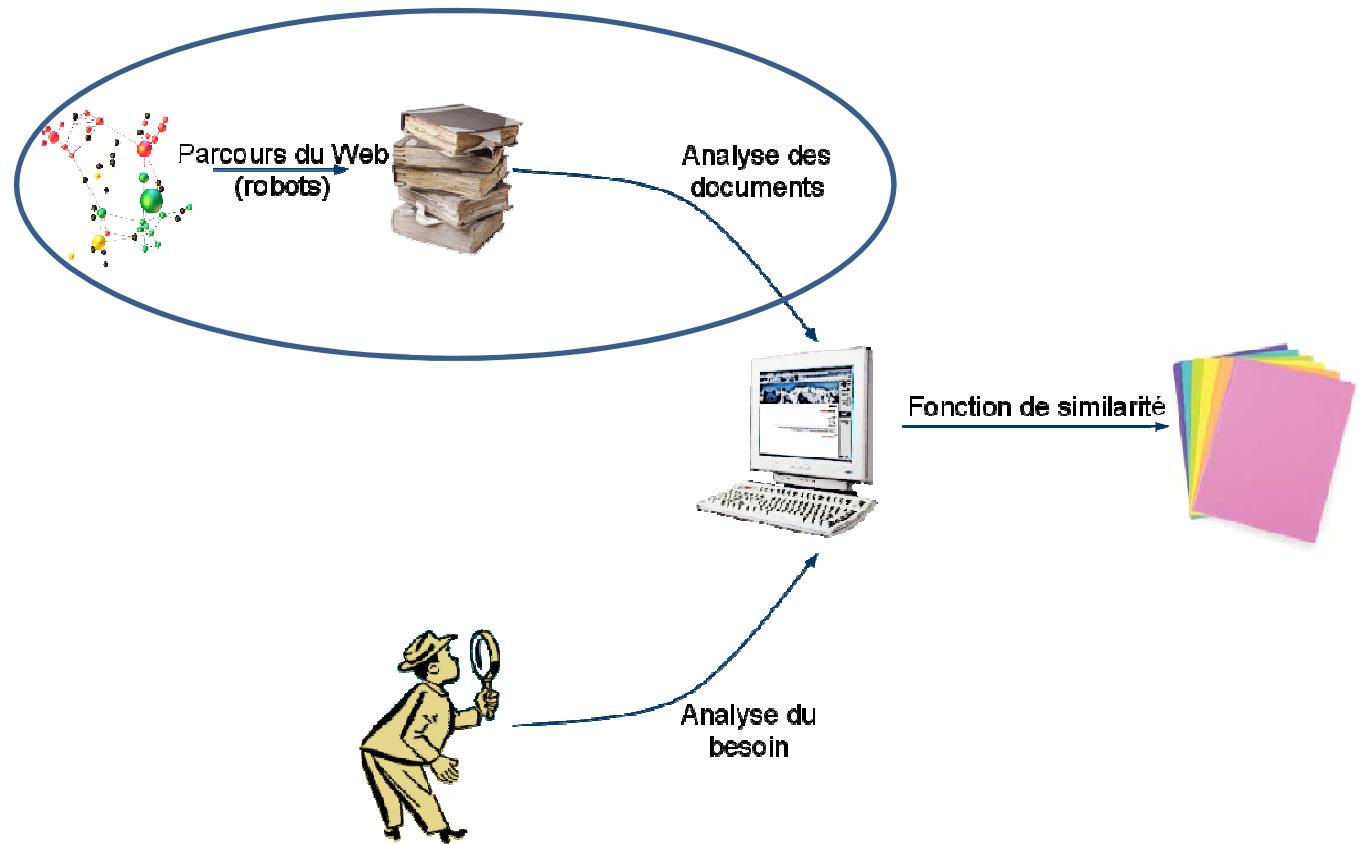
- Les frontières ne sont pas toujours claires...



# *Des enjeux importants*

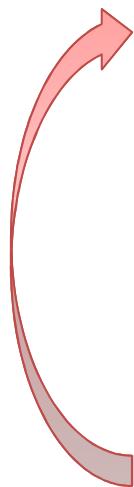
- Des enjeux **économiques**
  - 115 milliards d'euros générés par le Web en 2009 au Royaume-Uni
  - 75 % des sites sont trouvés par l'intermédiaire d'un moteur de recherche  
(source : WebPosition)
- Des enjeux **culturels et politiques**
  - Une prédominance américaine sur la gouvernance et les outils
  - Des algorithmes de recherche et des moyens mis en œuvre secrets
  - Une volonté de contrôle par les pouvoirs en place
- Des enjeux **éthiques**
  - Un réseau international, mais des droits nationaux
  - Le plagiat est devenu monnaie courante
  - Attention au respect de la vie privée
  - L'information est très difficile à effacer !

# Crawling

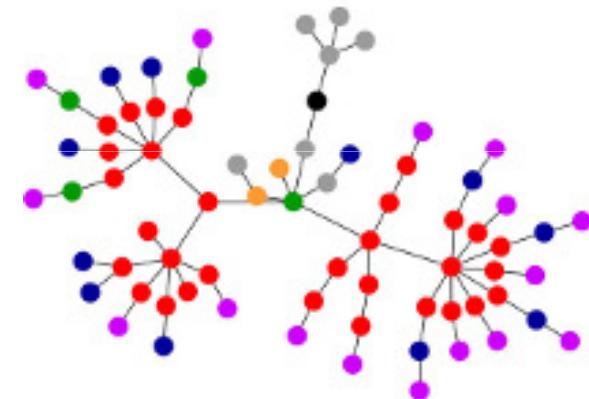


# *Le crawling dans le principe*

Des **robots** (ou **crawlers**, **spiders**, ...) partent d'une liste d'URL connues (*amorces*)



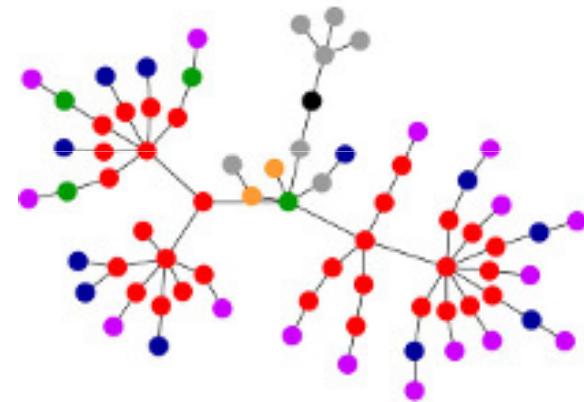
- Ils "parcourent" les sites et indexent leur contenu
  - Extraction des URL présentes dans les nouveaux sites
  - Placement de ces URL dans la file d'attente



Ils explorent ensuite la **file d'attente**

## *Oui, mais...*

- Une seule machine ne peut pas gérer le parcours du Web entier
  - Le crawling doit être **distribué**
- Certaines pages sont « malicieuses »
  - **Spam**
  - Des « **pièges à robots** » (intentionnels ou non)
- Et il faut gérer :
  - La **latence** et la **bande passante** des serveurs
  - Les **stipulations** des webmasters
  - Les **sites miroirs** et les **pages dupliquées**
  - Les règles de **politesse**



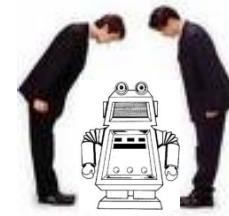
# *Le crawling dans la pratique*

Un crawler doit :

- Être **poli** :
  - Ne pas solliciter un serveur trop souvent
  - Ne parcourir que les **pages autorisées**
  - Respecter les stipulations du **robots.txt**
- Être **robuste** :
  - Éviter les **pièges** et les comportements malicieux des serveurs
  - Passer à l'échelle convenablement
    - Tâches distribuées
    - Conception propre à augmenter régulièrement le nombre de sites visités
  - Savoir **s'adapter** à de nouveaux formats ou protocoles
- Être **pertinent** :
  - Parcourir les pages de « **qualité** » en priorité
  - Repasser souvent dans les **endroits animés**

# *La politesse du robot*

- Politesse **implicite** : ne pas solliciter un serveur trop souvent
- Politesse **explicite** : le *robots.txt*
  - Un protocole permettant de choisir l'accès donné au crawler sur un site
  - Le site annonce les restrictions dans un fichier *robots.txt* à la racine. Exemple :
    - Aucun robot ne doit visiter les URL du répertoire /fichiers/perso,  
sauf le robot nommé « lebonmoteur »
    - User-agent: \*
    - Disallow: /fichiers/perso
    - User-agent: lebonmoteur
    - Disallow:
  - [www.robotstxt.org/wc/norobots.html](http://www.robotstxt.org/wc/norobots.html)



# *Les étapes du crawling*

- 
1. Choisir une URL dans la file d'attente
  2. Récupérer le document correspondant
  3. Analyser le document
    - Extraire les liens externes
    - Pour chaque lien externe :
      - Vérifier qu'il vérifie les contraintes (robots.txt, contraintes spécifiques au moteur)
      - Vérifier qu'il n'est pas déjà dans la file d'attente
      - Si c'est bon, l'ajouter dans la file d'attente
  4. Vérifier que le contenu du document est nouveau
    - Si oui, ajouter à l'index

Priorités des URL

DNS

Normalisation  
des URL

Mise en cache des  
robots.txt

Détection de  
doublons

# Détection de doublons

- Le Web est rempli de doublons (plagiat, éditions multiples, etc.)
- Mais ce ne sont pas des doublons parfaits
  - Les procédés de *fingerprinting* ne suffisent donc pas
  - Il faut mettre en œuvre des mesures de similarité

The image shows two side-by-side screenshots of web pages. The left screenshot is from a blog post titled "Comment sauver le soldat Wikipedia" by Cédric Le Merrer on August 19, 2011. It discusses the decline of Wikipedia's activity due to SEO pressure and the Panda update. The right screenshot is from a news article titled "IL FAUT SAUVER WIKIPEDIA" on August 24, 2011, from the website owni.fr. Both pages show social sharing buttons for Twitter, Facebook, and email, and readability options for font size.

**Comment sauver le soldat Wikipedia**

Posté par Cédric Le Merrer le 19.08.11 à 15:01 | tags : wikipedia | 9

Malgré des positions SEO toujours dominantes – et enco Panda - Wikipedia semble avoir perdu son Mojo et n' contenus bénévoles que le web se dispute. Wikipedi comment ? Flu analyse les possibles.

D'abord les chiffres : jusqu'à 90 000 en 2010, les contributeur Beaucoup sont persuadés que la chute du nombre de co l'encyclopédie serait complète et surtout Wikipedia reflétant le moyen, "un geek masculin de 26 ans" selon son fondateur Jimmy Wales , l'encyclopédie mar fatalement de points de vue féminins et non occidentaux.

Mais cette baisse somme toute assez limitée cache un p correspond plus aux usages en vogue aujourd'hui sur le web évident du web 2.0. Reposant sur des outils de programmation, Wikipedia était avant tout ce qu'en font ses usage Symbole des mutations de l'époque, Wikipedia reflète aussi

**IL FAUT SAUVER WIKIPEDIA**

Tweeter 194 Recommander 148 email A- A+

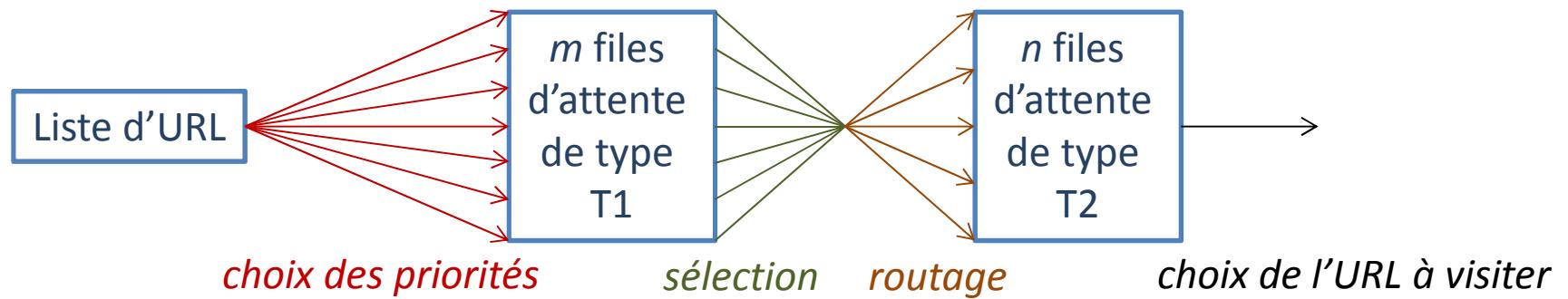
D'abord les chiffres : jusqu'à 90 000 en 2010, les contributeurs actifs n'étaient que 82 000 en Beaucoup sont persuadés que la chute du nombre de contributeurs n'est qu'un phénomène r l'encyclopédie serait complète et surtout Wikipedia reflétant les centres d'intérêt de son cont moyen, "un geek masculin de 26 ans" selon son fondateur Jimmy Wales , l'encyclopédie mar fatalement de points de vue féminins et non occidentaux.

Mais cette baisse somme toute assez limitée cache un phénomène plus alarmiste : Wikipedi correspond plus aux usages en vogue aujourd'hui sur le web, après avoir été pourtant le sym évident du web 2.0. Reposant sur des outils de programmation dynamique, facile à modifier s programmer, Wikipedia était avant tout ce qu'en font ses usagers. Que s'est-il passé ?

Symbole des mutations de l'époque, Wikipedia reflète aussi les aspirations libertaires de son Jimmy Wales est un libertarien un individualiste qui ne croit ni en la "société" ni en la légitimiti

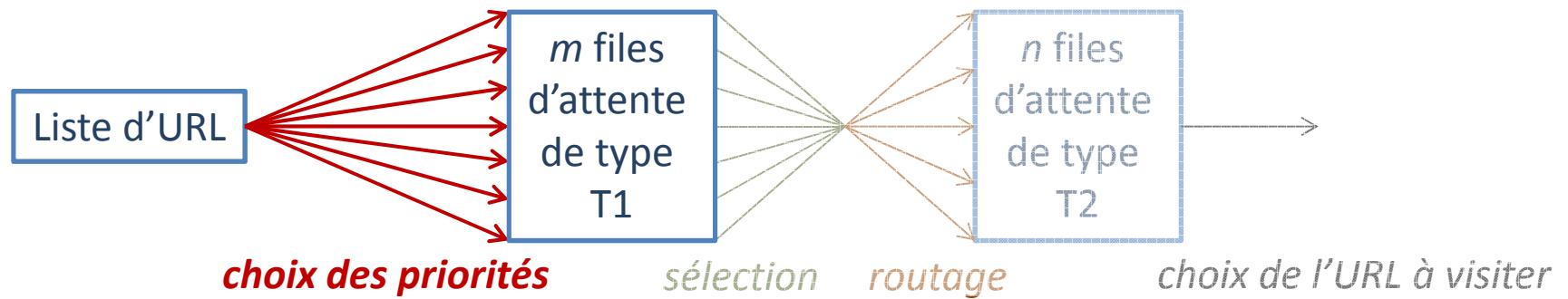
# Priorités des URL : l'exemple de Mercator

- Deux contraintes en conflit
  - Être **poli**
  - Visiter plus souvent les **pages fréquemment modifiées**
- Une file avec priorité simple (FIFO) est malpolie : les pages proposent souvent des liens vers leur propre site
- **Deux types de files d'attente :**
  - T1 pour gérer les priorités
  - T2 pour gérer la politesse



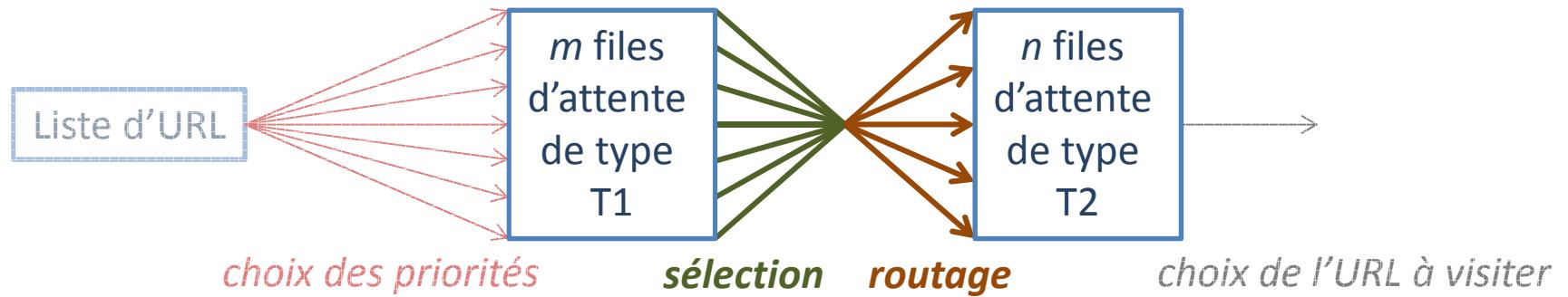
## Files de type T1

- Chaque URL reçoit une **priorité** comprise entre **1 et  $m$** .
  - Fréquence de visites en fonction des précédentes visites et des changements de la page lors de ces visites
  - Choix spécifiques (par exemple, *visiter les sites de news plus souvent*)
- L'URL est ajoutée à la file d'attente correspondante
  - Chaque file d'attente est de type **FIFO**



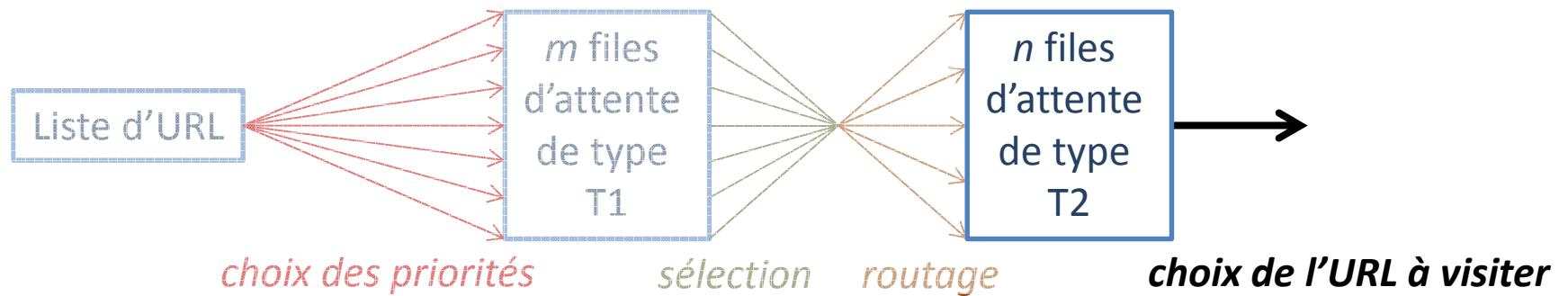
# Passage des URL des files T1 aux files T2

- Quand une file T2 cherche une **nouvelle URL**, on en sélectionne une dans une file T1
  - En fonction des priorités ou au hasard
- Chaque file T2 contient les URL **d'un seul et même hôte**
  - On maintient un **lexique** reliant un hôte à la file correspondante
  - Chaque file doit rester **non-vide** !
  - Pour chaque hôte, on conserve également le moment  $t_e$  auquel il pourra être sollicité de nouveau (heuristique fonction du dernier accès)



# Choix de l'URL à visiter

- Un **thread** du crawler qui cherche une nouvelle URL à visiter :
  - Choisit un hôte  $h$  ayant un  $t_e < \text{maintenant}$
  - Prend le premier élément de la file d'attente  $T_{2,h}$  correspondant à cet hôte
    - Si  $T_{2,h}$  est maintenant vide, sélectionner une URL  $u$  dans les files T1
      - Si l'hôte de  $u$  possède déjà une file T2, l'ajouter à cette file et recommencer
      - Sinon utiliser  $T_{2,h}$  pour cette URL et créer une nouvelle entrée pour cet hôte
    - Si  $T_{2,h}$  est non-vide, mettre à jour  $t_e$
- Recommandation : utiliser 3 fois plus de files T2 que de threads
  - Pour garder tous les threads occupés tout en restant poli



# Le spam

# *Qu'est-ce que le spam pour un moteur de RI ?*

- Un **spam** est une page Web qui remonte dans les résultats d'un moteur de recherche :
  - Pour des **mots-clés précis**
  - De façon **artificielle**
  - (Indépendamment de son **contenu réel**)
- Pourquoi faire du spam ?
  - Parce qu'être dans les premiers résultats pour certaines requêtes peut rapporter gros
  - L'optimisation pour moteurs de recherche (*SEO*) est une activité économique à part entière
- Pourquoi de contrer le spam ?
  - Parce que le spam est souvent à l'opposé de la notion de pertinence

# Comment spammer ?

- Les premiers moteurs de recherche s'appuyaient essentiellement sur le ***tf.idf***
  - Comment faire du spam avec du *tf.idf* ?
  - Comment conserver une page lisible tout en pratiquant ce type de spam ?
- Le ***cloaking*** : présenter un contenu différent à un moteur de recherche qu'à un humain
  - Avec l'adresse IP ou le *user-agent*
- La ***redirection*** : optimiser une page pour un seul mot-clé, puis rediriger l'internaute vers la « vraie » page

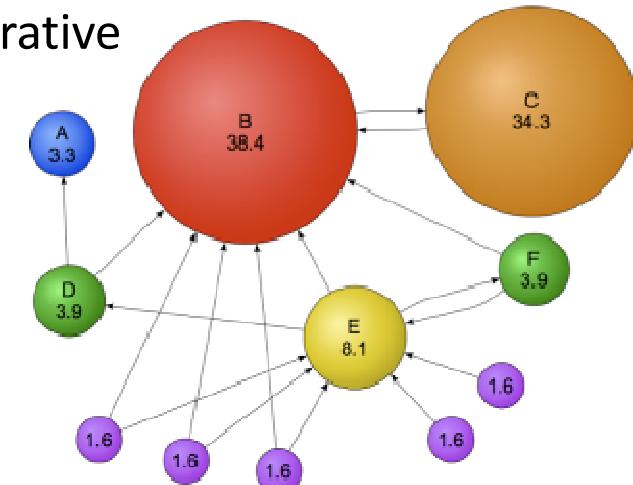
# *La lutte contre le spam*

- Repérer les pages qui font **autorité**
  - Votes des utilisateurs
  - Votes des autres auteurs de pages (voir plus loin)
- Des limites sur les **mots-clés**
- **Analyse de liens**
  - Repérer les chaînes de liens suspectes
  - Les spameurs font des liens vers d'autres spameurs
- **Apprentissage**
  - Trouver des « traits » importants pour différencier un spam d'une page honnête    **Lesquels ?**
  - Faire (un peu) d'analyse linguistique
  - Utiliser des ensembles d'apprentissage sur des listes connues de spams
- **Etc.**

# Le PageRank

# PageRank

- Mesure de l'**importance relative objective** d'une page Web :
  - Indice de **popularité** ; notion de **confiance** collaborative
  - Utilisation de la structure des liens du Web :
    - Les **liens sortants** (forward links)
    - Les **liens entrants** (backlinks)
- Justification intuitive :
  - Le nombre de liens entrants d'une page est révélateur d'une certaine importance  
(analogie : spéculation des futurs Prix Nobel par des comptages de citations)
  - Une page ayant un lien entrant provenant d'un site lui-même important (journal en ligne, grand site, portail, etc.) est plus importante qu'une page ayant des liens entrant provenant de sites peu importants : notion récursive de l'importance d'une page
- On voit le Web comme un **graphe**



# PageRank

- La **probabilité** pour qu'un utilisateur cliquant au hasard arrive sur une page.
- Obtenir un fort PageRank pour une page qui a de nombreux liens entrants et/ou des liens entrants provenant de pages elles-mêmes importantes :

$$PR(u) = d \sum_{v \in Bu} \frac{PR(v)}{C(v)} + (1 - d) \left( \times \frac{1}{N} \right)$$

- **Bu** : ensemble des pages ayant un lien entrant sur la page  $u$
- **C(v)** : nombre de liens sortant de la page  $v$  (chaque page diffuse son vote de façon égale sur tous ses liens sortants)
- **d** : facteur d'amortissement ;  $d$  vaut 0.85, donc une page n'ayant aucun lien entrant aura un PageRank de 0.15

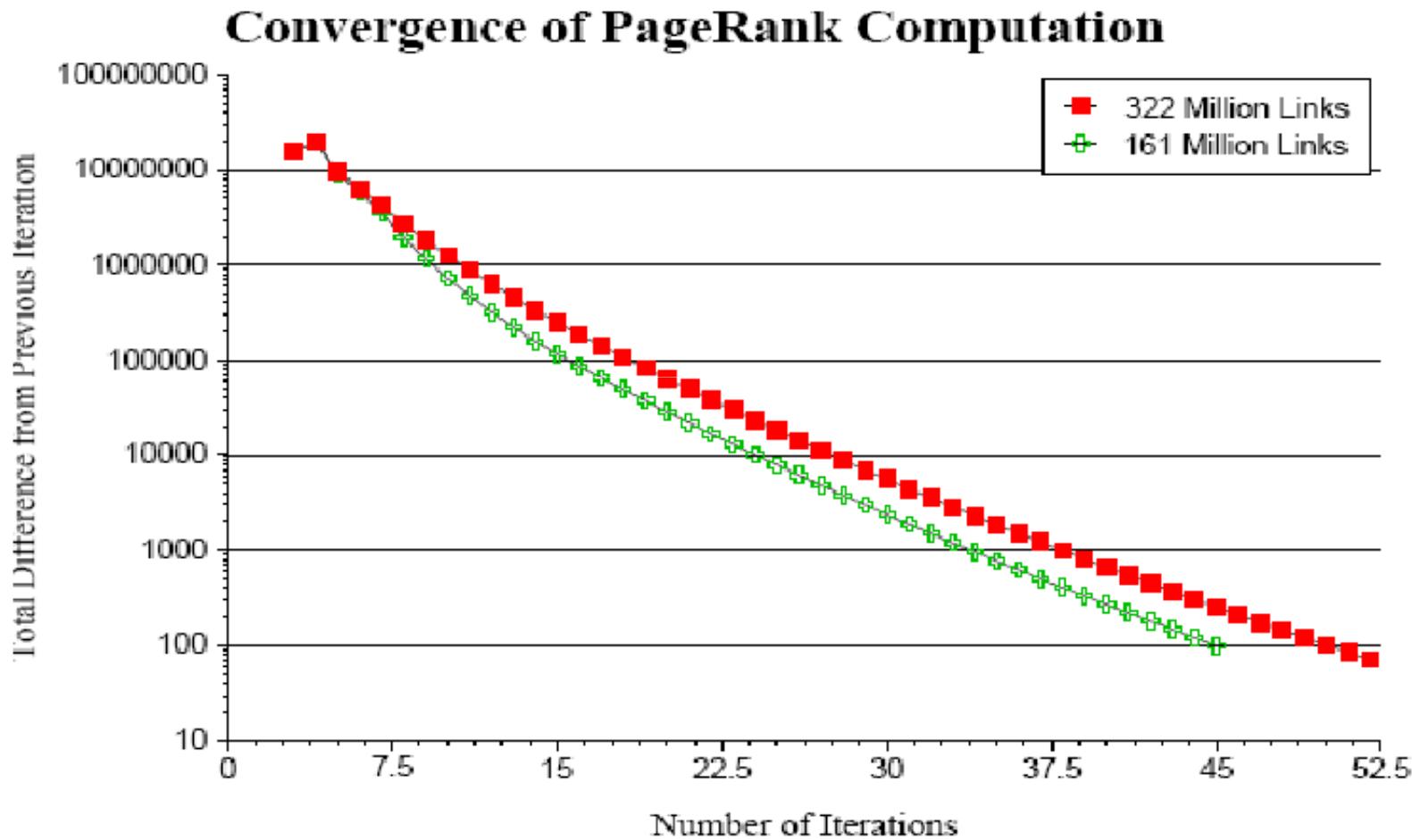
Quelle est la signification de  $d$  ?  
Son utilité ?

# *Calcul du PageRank*

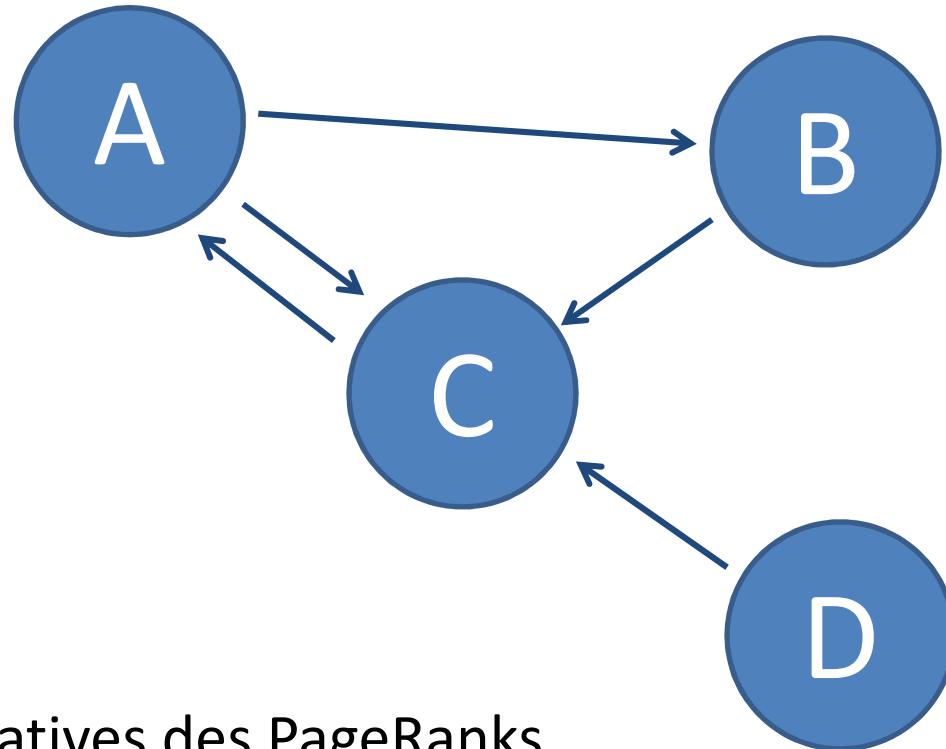
- Le PageRank moyen est 1 (avec  $1/N$ , la somme des  $PR$  est 1)
- Le PageRank d'une page dépend des PageRanks des pages qui pointent vers elle :
  - Calcul des PageRanks sans connaître la valeur finale de tous les PageRanks impliqués
  - **Itérations** qui approchent des valeurs finales jusqu'à convergence
  - La **valeur initiale** n'affecte pas les valeurs finales mais le nombre d'itérations pour atteindre la convergence (ex : prendre des valeurs initiales correspondant à la fréquentation des pages)
  - Le **coût** pour le calcul des PageRanks est très faible relativement au temps de construction d'un index complet



# *Calcul du PageRank*

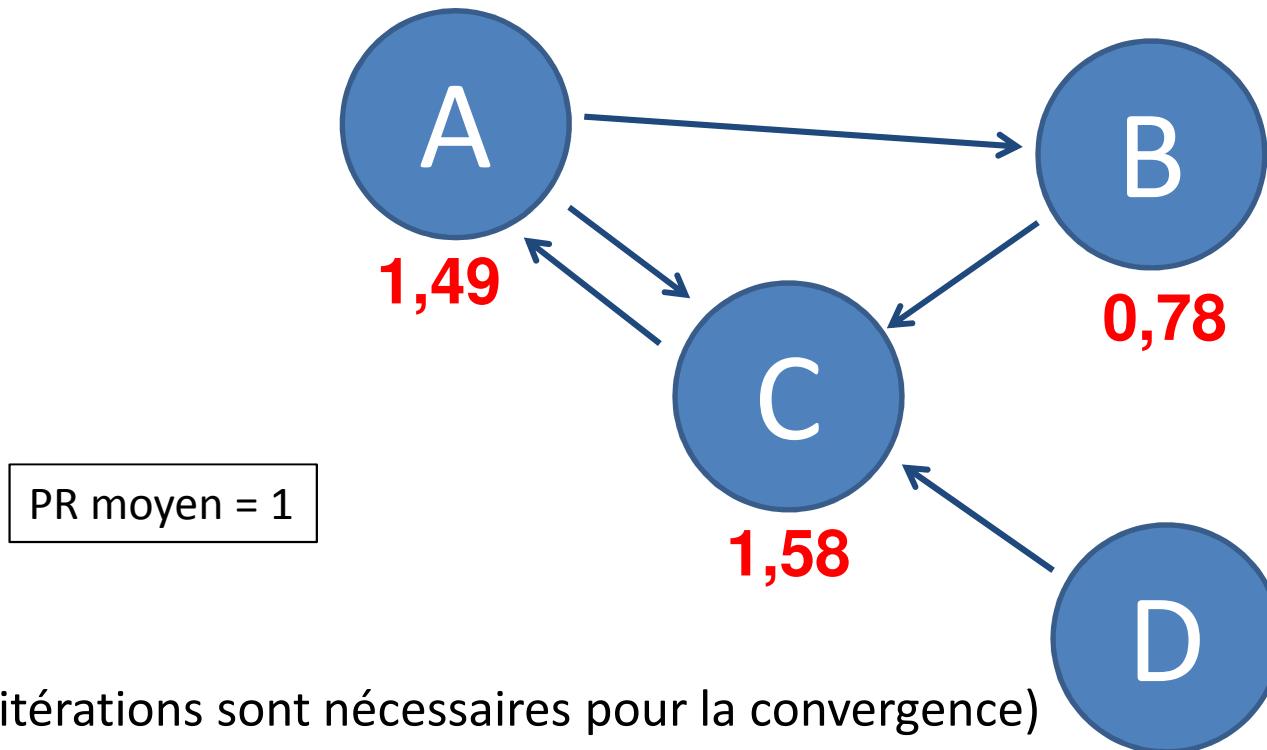


# *Calcul du PageRank*



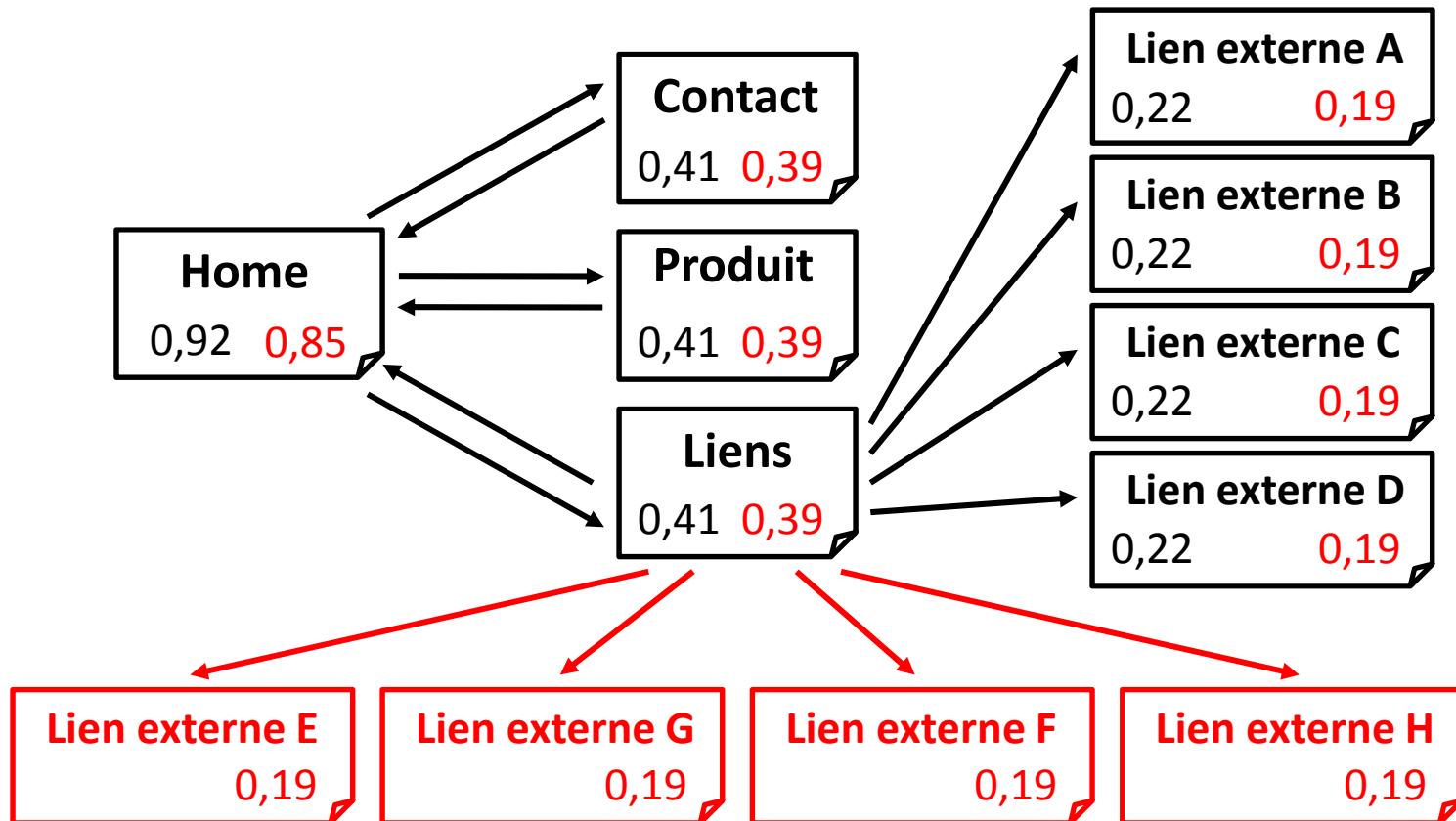
Valeurs relatives des PageRanks  
des pages ?

# *Calcul du PageRank*



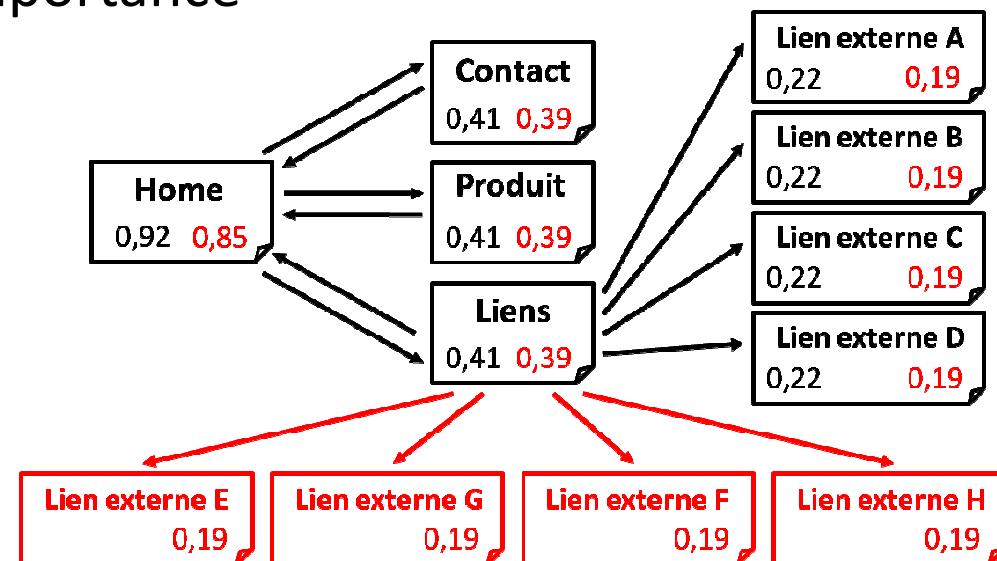
- (20 itérations sont nécessaires pour la convergence)
- La page **D** a une valeur minimale du PageRank (aucun lien entrant)
- La page **C** a de nombreux liens entrants
- La page **A** bénéficie du lien entrant provenant de la page C

# PageRank : cas simple

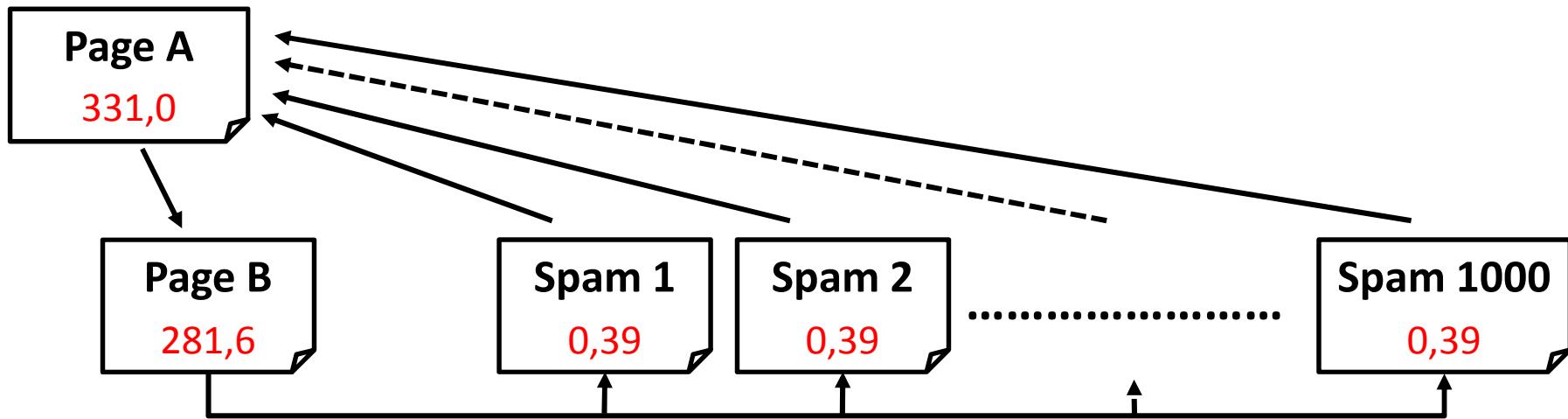


# *PageRank : cas simple*

- Rétroaction des valeurs des PageRanks pour la page Home
- Plus le nombre de liens sortant de la page Links est important, plus le partage du PageRank est diffus
- Plus le nombre de pages augmente, plus des pages sans nouveaux liens entrant perdent de l'importance
- Avoir un lien vers une page importante n'augmente pas le PR (!)



# *PageRank : la tentation du spam*



- Le nombre de pages d'un site n'augmente pas le PR moyen
- Une certaine organisation hiérarchique d'un site peut fortement concentrer le PR sur la page principale
- Maintenant décelable par les robots (ex : Googlebot) qui pénalisent le site

# L'importance de l'interface

# *Les interfaces utilisateur (1/6)*

- Interfaces utilisateurs généralement dénudées
- Accès à une recherche avancée et à des paramètres
- Présentation d'extraits de résultats (*snippets*)
- Présentation de résultats commerciaux ("sponsorisés")

• Exemple:

The screenshot shows the Altavista search interface. The search query 'randonnée en montagne' is entered in the search bar. The 'TROUVER' button is highlighted in red. Below the search bar, there are filters for 'chercher' (Tous les pays, France) and 'resultats en' (Toutes les langues, Anglais, Français). The results section is titled 'Résultats sponsorisés' and contains four entries:

- Trekking Italy**  
Les Iles Eoliennes offrent occasions de **randonnées** à sept-oct.  
[www.carasco.com](http://www.carasco.com)
- Bienvenue aux chalet les Mélèzes à Lanslevillard - Val Cenis**  
Location d'appartement au pied du parc de la Vanoise qui vous propose de très nombreuses **randonnées en montagne** en Savoie.  
[www.chalet-les-mezezes.com](http://www.chalet-les-mezezes.com)
- Venez faire de la randonnée en montagne en Rhône-Alpes**  
Découvrez les nombreuses activités sportives proposées dans notre région. Laissez-vous guider sur le site officiel Rhône-Alpes Tourisme.com pour pratiquer sereinement vos loisirs préférés.  
[www.cnt-rhonealpes.fr](http://www.cnt-rhonealpes.fr)
- Camille Bonaventure: chalet hôtel et voyage d'aventure**  
Camille Bonaventure vous accueille dans leur Chalet-Hôt...  
[www.camillebonaventure.com](http://www.camillebonaventure.com)

# Les interfaces utilisateur (2/6)

- Présentation de résultats visuels
- Utilisation de termes associés et d'annuaires
- Exemple:

The screenshot shows the Exalead search engine interface. At the top, there is a search bar with the text "recherche d'information". Below the search bar are three radio buttons: "Tout le web" (selected), "Pages en français", and "Pages : France". To the right of the search bar are links for "S'identifier" and "Préférences". Below the search bar, the text "Résultats 1-10 sur environ 20 065 433 pour recherche d'information" is displayed. The main content area contains several search results:

- Toute l'actualité sur la recherche**  
www.atelier.fr - Découvrez l'Atelier, site référence de veille technologique : actualité, newsletters, conférences, ...
- Ecole Européenne d'Intelligence Economique**  
www.eeie.fr - Découvrez l'école Européenne d'intelligence Economique, sa démarche pédagogique unique, son module ...
- La recherche d'information sur Amazon.fr**  
www.amazon.fr - Amazon.fr met à votre disposition son stock de plusieurs milliers de livres constamment optimisé ...
- APPRI, Association Périnatalité Prévention Recherche Information**  
La vocation de l'association est de prévenir et lutter efficacement contre le tabagisme chez la femme. Actions de formation continue, ...  
www.appri.asso.fr/ - 1k - Ajouter aux raccourcis
- GDR TICS - Groupement de Recherche Technologies de l'Information ...**  
Groupement de Recherche Technologies de l'Information ... Dernières informations du GDR : La lettre du GDR & Société du 24 Juin 2006  
Abonnez-vous à la ...  
gdrtics.u-paris10.fr/ - 7k - Ajouter aux raccourcis

On the right side, there is a sidebar titled "Affiner la recherche" with the following sections:

- Termes associés**
  - Rechercher des informations
  - Information recherchée
  - Information scientifique
  - Recherche d'informations sur Internet
  - Stratégie de recherche
- Multimedia**
  - Audio
  - Vidéo
  - RSS
- Langues**
  - Français
  - Anglais
- Annuaire**
  - Commerce et économie
  - Sciences
  - Internet
- Type de fichiers**
  - Acrobat (.pdf)
  - Word (.doc)
  - Text (.txt)

A "Plus de choix..." button is located at the bottom right of the sidebar.

# *Les interfaces utilisateur (3/6)*

- Support pour des requêtes exprimées en langue
  - Focus sur l'utilisateur et non sur le fonctionnement interne du moteur de recherche
- Exemple :

The screenshot shows the Ask.com search interface. The search bar contains the query "Comment installer un programme sur ubuntu ?". Below the search bar are two buttons: "Web" and "Images", with "Web" being highlighted. To the right of the search bar is a blue "Rechercher" button. Further to the right is a link to "Recherche avancée". The main content area displays search results:

- Trouvez un logiciel** Résultat commerciaux  
www.telecharger-sans-risque.com Dernières versions disponibles Recherche, **installation** sans risque  
Télécharger OpenOffice
- Driver** Résultat commerciaux  
Vérifiez gratuitement que tous les Drivers de Votre PC sont à jour !  
drivers.avanquest.com
- Flash Player pour Linux** Résultat commerciaux  
Pour lire les contenus interactifs Téléchargez la dernière version ici  
www.eorezo.com/Logiciels\_Gratuits
- Installer un logiciel** Résultat commerciaux  
Tout savoir sur **Installer un logiciel** Retrouvez nos Experts !  
www.les-experts.com

On the right side of the results, there is a sidebar titled "Recherches Connexion" which lists various search terms:

- SUR X3 Art
- SUR 13
- South Side 13
- Southside X3
- Brown Pride
- Cholo Drawings
- Homies
- Surenos
- Cholos
- Tf1
- Latin Kings
- Poems of Surenos

# *Les interfaces utilisateur (4/6)*

- Aide à la formulation des requêtes durant la saisie
  - Par exemple, guidage par des suites de mots et le nombre de documents qui les contiennent

- Exemple:



# *Les interfaces utilisateur (5/6)*

- Présentation des résultats sous forme graphique (cartes...)
  - Aide à la navigation et à la reformulation de sa requête

- Exemple :



# *Les interfaces utilisateur (6/6)*

- Étiquetage "sémantique"

The screenshot shows a search interface for 'Paris' within a 'Wikipedia Articles' section. At the top, there's a search bar with the query 'What is the capital of France?' and a 'search' button. Below the search bar, the word 'Paris' is highlighted in a red box. To the right of 'Paris', there's a small thumbnail image of the Eiffel Tower. The main content area contains a brief summary of Paris as the capital of France, mentioning its administrative region and population. A sidebar on the right provides links to 'Hauts-de-Seine' and 'Seine-Saint-Denis'. Below the summary, a list of related Wikipedia articles is shown, each preceded by a blue square icon and a title. The titles include 'Paris', 'Capital of France', 'Capital punishment in France', 'Economy of France', 'Early Modern France', and 'Lyon'. The entire interface has a dark theme with light-colored text and buttons.

# *Le Web, un réseau public*

- Un réseau **international**, mais des droits **nationaux**
  - En Chine, des sites effacés des résultats de recherche
  - En France également
- Le **plagiat** est devenu monnaie courante
- Attention au **respect de la vie privée**
  - Documents mis à disposition sur la toile
  - Informations recueillies sur les internautes  
(formulaires ou informations sur les visiteurs)
- L'information est **très difficile à effacer** !
  - Google cache
  - Wayback machine
  - Listes de diffusion
  - Blogs, Facebook, ...

# *Des moteurs de recherche opaques (1/2)*

- Toutes les étapes de la recherche sont des **secrets industriels**
  - Comme pour toutes les activités à but commercial
  - Mais l'information et la connaissance sont des "marchandises" particulières
  - Le projet [Nutch](#) tente de promouvoir un moteur de recherche libre
  - Les réseaux tissés par les géants du Web peuvent poser question  
(sans tomber dans la paranoïa).

↳ Nécessité de varier ses sources d'information

## *Des moteurs de recherche opaques (2/2)*

- Rien ne permet de vérifier que les critères de classement des documents par rapport à la requête sont **équitables**
  - Séparation des liens commerciaux et des liens de recherche ?
  - Affinités particulières entre moteurs et sites commerciaux ; ex : Google et Yahoo avec Amazon, Voilà avec eBay et PriceMinister.
  - Google possède YouTube, cela influe-t-il sur le rang des liens YouTube ?
- Toutes les requêtes (**logs**) sont conservées
  - Tous dans l'illégalité vis-à-vis de la législation européenne (13 mois pour Yahoo, 9 mois pour Google, 18 mois pour Bing)
  - cf. demande des logs de Google et Yahoo par le gouvernement américain
  - Google, Yahoo et Microsoft veulent mettre les logs à disposition de la communauté scientifique (problèmes de droit)
  - Grâce aux comptes, les résultats (et les publicités) sont personnalisés

# *Quelques moteurs de recherche*

**YAHOO! SEARCH**



**le moteur**  
de la recherche

**exolead**  
France

**altavista**

**Google**  
France

et bien d'autres, sans compter les méta-moteurs...



**Cuill**

**mylive search**

**swingly**

# **Web 2.0, réseaux sociaux et médias sociaux**

# *Web 2.0*

- Le Web 2.0 selon 3 dimensions :
- - **technique** : utilisation de technologies qui sont combinées (ergonomie des sites Web et interfaces utilisateurs, feuilles de style CSS, syndication de contenu, utilisation d'Ajax) ; transition vers des applications Web pour les utilisateurs ;
- - **sociale** : interactions entre les utilisateurs et le partage (blogs, wikis, réseaux sociaux) ;
- - relative aux **données collectées** : sont dépendantes de l'application Web 2.0 considérée et sont accessibles quel que soit le lieu de connexion au site 2.0.

# Le verrouillage des données par les applications Web 2.0

<b>Applications Web 2.0</b>	<b>Flickr</b>	<b>Del.icio.us</b>	<b>Linked-in</b>	<b>Gmail</b>
<b>Données collectées</b>	Photos	Favoris	Contacts	Méls et éléments joints

# Les réseaux sociaux

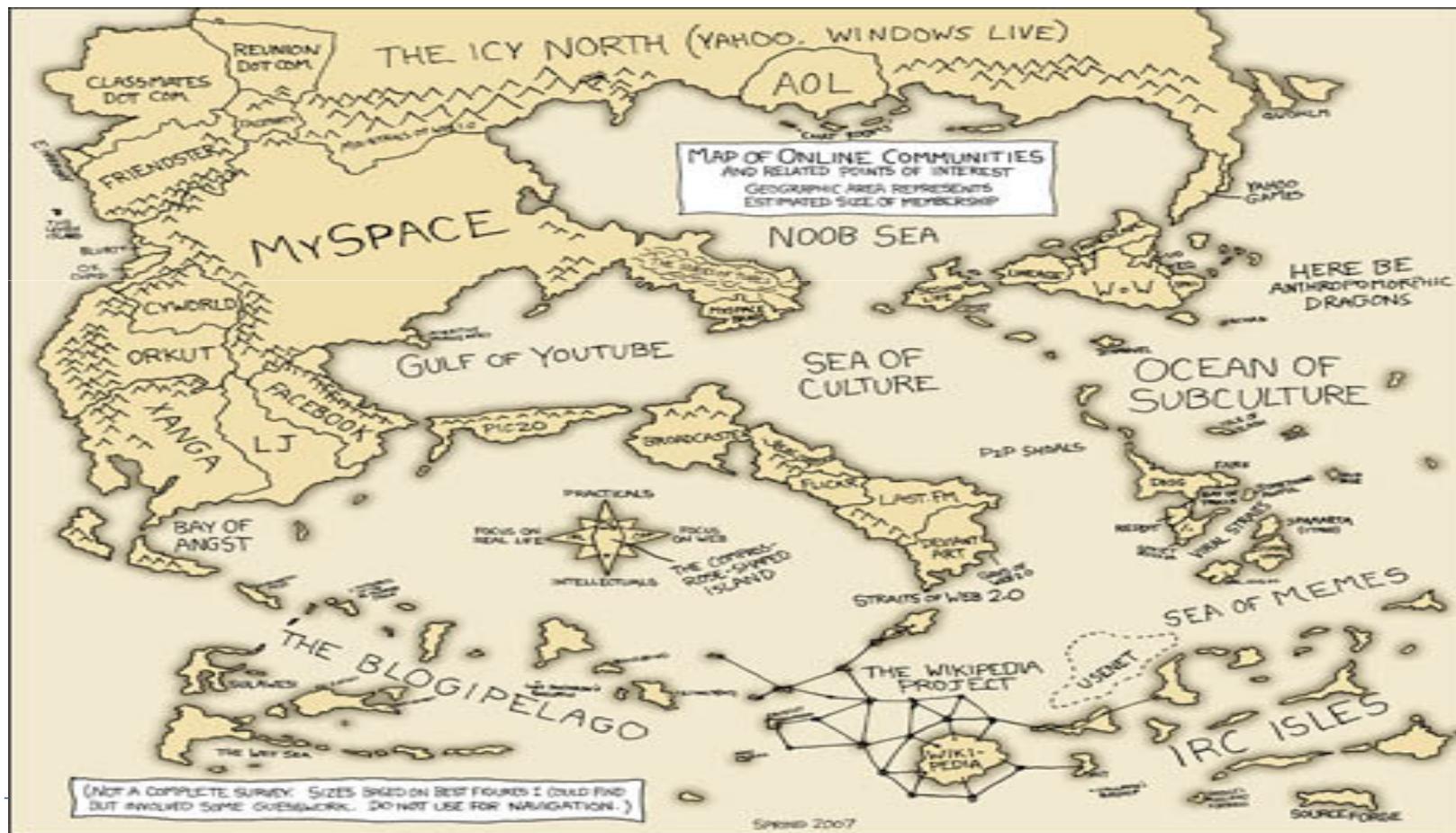
- Une mosaïque d'acteurs
- Apparition massive depuis 2003
- Applications phares du web 2.0



etc.

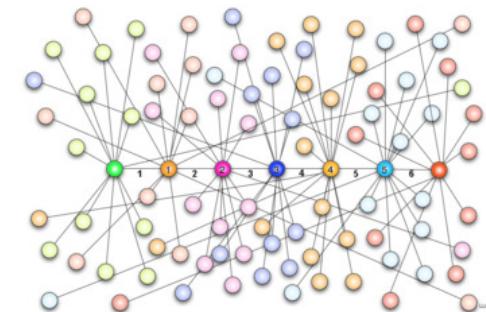
# Les réseaux sociaux

- Une cartographie amusante des réseaux sociaux et du Web 2.0



# Réseaux sociaux : concepts

- Organisations ou individus reliés entre eux par des liens qui sont créés à l'occasion d'interactions sociales
- Identité de l'internaute et son carnet d'adresses : pivot d'un réseau social + fonctions de messagerie et forums de discussion.
- Spécificité à chaque réseau social : ajout d'une série de fonctionnalités comme les « *news feed* » sur les actions des membres de son réseau
- Notion de distance entre individus
  - Phénomène du « petit monde » de Milgram



# Quelques réseaux sociaux

Réseau social\Critère	Nombre de membres (en millions)	Date de création	Utilisation	Fonctions spécifiques
LinkedIn	20	2003	Professionnelle	Importation possible de ses contacts depuis Yahoo, Gmail ou AOL.
Viadeo	2	2004	Professionnelle	Chemin existant entre soi et un membre si la distance est inférieur ou égale à 4. Après parrainage de 10 membres, 1 mois d'abonnement Premium offert.
MySpace	220	2003	Personnelle/ Professionnelle	Espace web personnalisé pour chaque membre qui permet d'héberger des créations, notamment musicales.
Facebook	70	2004	Personnelle/ Professionnelle	NOMBREUSES applications et API développées.

## *Des comportements qui évoluent*

- Web 2.0 : les internautes consommateurs et consommacteurs
- Loi du 1 % dans les médias participatifs (YouTube, Wikipédia, Agoravox, etc.)
- Nouveaux internautes, nouveaux usages : interactivité, commentaires et liens déposés sur les blogs. Récupération par des agrégateurs de contenu

# *Opportunités offertes par le Web 2.0*

- Possibilité de contacts gigantesques
  - Trouver des prospects, nouer des partenariats, recruter ou être recruté (LinkedIn ou Viadéo)
- Avec ses relations du RS : utilité des liens « faibles » ou « lâches » pour la mise en relation
- Blog : avoir une existence sur le Web, vitrine professionnelle,
- etc.

# *RI pour identifier de nouveaux risques*

- Risque en matière de sécurité (failles possibles dans les RS ou leurs API, vol, détournement ou utilisation frauduleuse des données personnelles)
- Spam, spyware, utilisation frauduleuse des données personnelles par des tiers malveillants
- Trop grande confiance accordée *a priori*, absence de vérification des données, des adresses IP, etc.
- Divulgation d'informations nominatives massives et sensibles

# *Questions soulevées par les réseaux sociaux / données*

- Persistance des données publiées (obsolètes, préjudiciables)
- Capacité à être retrouvé (obstacles : homonymes, faux profils)
- Reproductibilité des informations dans un contexte différent (message initial parfois dénaturé)

# *Données des RS & RI ??*

- Publicité peut être ciblée très finement
  - Pas seulement qui l'on est, mais qui l'on connaît (système Beacon de Facebook, Gmail par rapport aux méls et ses contenus)
- Mini-feed des RS : forme d'espionnage de la vie privée (en fonction de ses actions, de ses connaissances)

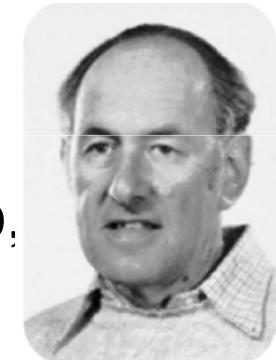
→ **Jusqu'où peut-on aller?**

# *Dates clés de l'historique 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, 1979)
- 1977 Modèle probabiliste (PRP) S. Robertson
- 1978 Première conférence SIGIR
- 1983 Début d' Okapi (modèle probabiliste)



Hans-Peter Luhn



Gerard Salton

## *Dates clés de l'historique 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)
- ERIA 2006