



# RECHERCHE D'INFORMATION

Master1 Miage Nanterre  
Sonia GUEHIS



# La Recherche d'Information (RI) - Définition

En anglais : Information Retrieval (IR)

- Étant donnée **une collection de documents** constitués essentiellement de texte, comment trouver les plus **pertinents** en fonction d'un **besoin** exprimé par quelques mots-clés?
- Un système de recherche d'information (SRI) : un système permettant de retrouver **une information pertinente** par rapport à une **requête** dans **une grande collection de documents**.
- Une branche de l'informatique qui étudie la construction des systèmes ayant pour objectif principal de permettre de retrouver une information spécifique, correspondant au besoin de l'utilisateur, dans un ensemble de documents.
- Trouver des documents **peu ou faiblement structurés**, dans une grande *collection*, en fonction d'un *besoin d'information*.
- Le domaine d'application le plus connu est celui de la recherche « plein texte ».

## RI - Mission

- La RI développe :
  - *Des modèles pour :*
    - Interpréter les documents d'une part,
    - Interpréter le besoin d'information d'autre part,
    - en vue de faire correspondre les deux,
  - *Des techniques pour calculer des réponses rapidement même en présence de collections très volumineuses.*
  - *Des systèmes (moteurs de recherches ) fournissant des solutions avancées prêtes à l'emploi.*

## RI – Contexte

- La RI est utilisée dans plusieurs contextes informatiques:
  - La recherche sur le Web, utilisée quotidiennement par des milliards d'utilisateurs,
  - La recherche de messages dans votre boîte mail,
  - La recherche de fichiers sur votre ordinateur,
  - La recherche de documents dans une base documentaire, publique ou privée,
  - ...

## RI – Stratégies de recherche

- La RI peut se faire en se basant sur des différentes stratégies :
  - La recherche **par mots clés**: Expression du besoin en information via un ensemble de mots clés. Les résultats retournés sont cherchés à travers un appariement requête-documents.
  - La recherche **par navigation**: s'opère le plus souvent lorsque l'utilisateur n'a pas une connaissance préalable sur les éléments informationnels du système, il s'agit d'une recherche par exploration. L'utilisateur fait des choix entre plusieurs alternatives pour affiner et exprimer son besoin.
  - La recherche **par facettes**: le concept est basé sur une navigation multidimensionnelle: La recherche est basée sur un ensemble de métadonnées définissant des catégories dans le contenu informationnel et permettant de regrouper les éléments de ce contenu. Exemple de facette de recherche: la langue si on a des documents en plusieurs langues, les marques des produits, les catégories de produits....

## RI- Enjeux

- Avec une base de données classique:
  - Un schéma de données connu, une organisation générale, avec des contraintes qui garantissent une certaine régularité.
- En RI:
  - les données ou documents , sont souvent **hétérogènes**, de **diverses provenances**, présentent des **irrégularités** et des **variations** dues à l'absence de contrainte et de validation au moment de leur création.
  - Utilisation par des **utilisateurs non-experts**.
  - L'enjeu: interpréter et décrire le contenu des documents + comprendre le besoin utilisateur, exprimé souvent de manière très partielle.

## RI–Documents et requêtes

- Un **document** peut être
  - un texte
  - un morceau de texte
  - une page Web
  - une image
  - une vidéo
- Une **requête** exprime le besoin d'information de l'utilisateur
  - Besoin en information est une expression mentale d'un utilisateur
  - Requête
    - Ensemble de mots-clés
    - Une représentation possible du besoin en information

## RI-Principe

- L'information recherchée se trouve dans des documents numériques. Elle est « cachée », inaccessible.
- L'utilisateur recherche parmi les documents ceux qui contiennent la réponse qui l'intéresse.
- Pour utiliser un moteur de recherche il est obligé d'exprimer son besoin par une requête.
- Les outils de recherche ne comprennent pas la langue naturelle. La requête doit être conforme au langage de requête qui est employé par le moteur de recherche :
  - *mots clés*
  - *opérateurs booléens, guillemets, ...*



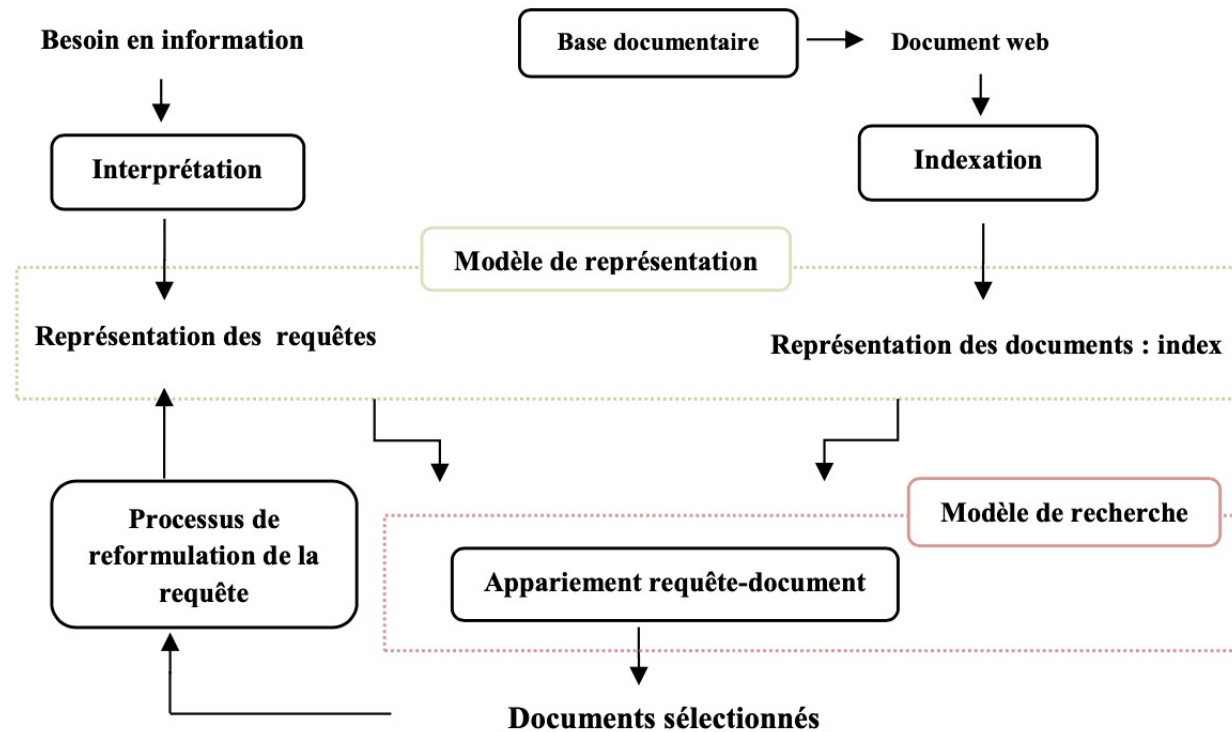
## RI-Démarche

- Collecter les documents (sources documentaires, Web, ...)
- Nettoyer et analyser l'ensemble des documents
- Créer un index inversé de l'ensemble des termes jugés représentatifs des documents
- Traiter la requête de recherche :
  - *mots clés*
  - *opérateurs booléens*
  - *métadonnées documentaires (auteur, titre, date d'édition, collection, ISBN, ...)*
- Classer les documents résultats selon leurs pertinences

## RI-Index

- Un index dans un livre:
  - Liste des termes ( mots, expressions) avec les emplacements où sont situés ces termes au sein du livre.
  - L'objectif: avoir un accès direct aux termes sans devoir lire l'ensemble du livre.
  
- Un index dans un moteur de recherche:
  - Liste des termes ( mots, expressions) avec pour chaque terme sont référencés les emplacements (sites Web) où ce terme est employé.
  - L'objectif: trouver un site Web à partir des termes qui y sont employés.

# SRI – Architecture



Architecture d'un SRI selon Salton et McGill 1986

## Exemple de recherche par mots clé: une souris



# Un site e-commerce connu

Toutes nos catégories

souris

ine et Maison

Chez sonia

Melieuses Ventes

AmazonBasics

Acheter à nouveau

Service Client

Ventes Flash

Iddes cadeaux

Livres

Dernières Nouveautés

High-Tech

Chèques-cadeaux

Ebooks Kindle

Informatique

Prévoyez et Economisez

Vendre sur Amazon

Jeux et Jouets

ins

Melieuses ventes

Offres reconditionnées

Nos idées cadeaux

Services Amazon

Amazon Assistant

1-48 sur plus de 100 000 résultats pour "souris"

Trier par: Amazon présente

Amazon Prime

☐ prime

Affiner la catégorie

Souris

Jeux vidéo

Souris gaming pour PC

Accessoires pour Nintendo Switch

Accessoires pour PlayStation 4

Xbox One: Consoles, jeux et accessoires

Accessoires pour PlayStation

Accessoires pour Xbox

Accessoires pour PlayStation 2

Voir les 19 catégories

Moyenne des commentaires client

★★★★★ & plus

★★★★★ & plus

★★★★★ & plus

★★★★★ & plus

Marque

☐ Logitech

☐ ViTsing

☐ Logitech G

☐ inphic

☐ PICTEK

☐ AmazonBasics

☐ TECKNET

Voir plus

Prix

0 à 20 EUR

20 à 50 EUR

50 à 100 EUR

100 à 200 EUR

200 à 500 EUR

Plus de 500 EUR

EUR Min

EUR Max

Aller

Promotions

☐ Ventes Flash

Nos marques

☐ Nos marques

Condition

☐ D'occasion

☐ Neuf

Nouveautés

Depuis 1 mois

Depuis 3 mois

Livraison internationale


☐ Livraison internationale disponible

Offre du moment


KLIM

Clavier Gaming KLIM - La qualité garantie 5 ans


En savoir plus sur KLIM >




KLIM Chroma Clavier sans Fil Gamer AZERTY - FRANÇAIS + Fil, Durable, Ergonomique, Discrét...  
★★★★★ 7 285  
prime




KLIM Light V2 Clavier sans Fil AZERTY FR + Fil, Ergonomique, Discrét, Waterproof, Silencieux + ...  
★★★★★ 222  
prime




Suggestions parmi nos marques  
Amazon Basics Souris sans fil ergonomique - DPI réglables - Bleu  
★★★★★ ~ 6 502  
14,61€  
prime GRATUIT Livraison en 1 jour.  
Recevez-le demain le 1 mars




Sponsorisé @  
inphic Souris sans Fil, Souris Optique Ergonomique sans Fil Rechargeable 2,4 G avec Nano-récepteur USB pour Ordinateur Portable PC MacBook...  
★★★★★ ~ 3 230  
12,99€  
Economisez 6 % avec coupon  
prime Livraison GRATUITE d'ici mercredi 3 mars




Sponsorisé @  
inphic Souris sans Fil Rechargeable, Mini Souris Optique sans Silence Click, Ultra Mince 1600 DPI pour Ordinateur Portable, PC, Ordinateur...  
★★★★★ ~ 36 278  
11,59€  
Economisez 8 % avec coupon  
prime GRATUIT Livraison en 1 jour.  
Recevez-le demain le 1 mars




Sponsorisé @  
inphic Souris Filaire pour Filles, Clic Silencieux, 4800DPI Réglable et 7 Boutons Programmables, Beau Design Rose Girly avec Ergonomie...  
★★★★★ ~ 402  
14,98€  
Economisez 7 % avec coupon  
prime GRATUIT Livraison en 1 jour.  
Recevez-le demain le 1 mars




N°1 des ventes  
Souris Sans Fil 2.4G, ViTsing 2400 CPI Souris Optique Mobile avec Récepteur Nano USB 6 Boutons 2400 DPI (5 Niveaux Réglables) pour...  
★★★★★ ~ 14 701



HP X1500 - Souris Filaire Noire, USB, 1000 DPI, Ambidextre)  
★★★★★ ~ 1 165  
6,59€  
prime GRATUIT Livraison en 1 jour.



Amazon's Choice  
Logitech M90 Souris Filaire USB, Suivi Optique 1000 PPP, Ambidextre, Compatible avec PC/Mac/Portable - Noire  
★★★★★ ~ 8 075



PICTEK Souris Gamer RGB Filaire/Souris de Jeu 8 Boutons Programmables, 7200 dpi réglable/Souris Gaming...  
★★★★★ ~ 2 806

# Recherche par facettes

« Souris »

Affiner

189 produit(s) trié(s) par

Pertinence

Catégories

Décoration et éclairage

- ☐ Coussin et housse (13)
- ☐ Galette de chaise (10)
- ☐ Papier peint (7)
- ☐ Lustre et suspension (6)
- ☐ Rideau (6)
- ☐ Coussin de sol (3)
- ☐ Store enrouleur (1)
- ☐ Stickers (1)

Terrasse et jardin

- ☐ Anti nuisibles (32)
- ☐ Mangeoire, nichoir et hôtel à insectes (3)
- ☐ Pouf de jardin (1)
- ☐ Lame bois pour terrasse et jardin (1)
- ☐ Pose et entretien de sol composite (1)

Peinture et droguerie

Peinture murale couleur / Te...

Salle de bains

Panneaux muraux décoratifs ...

Menuiserie

Peinture et traitement d'étan...



CAUSSADE

Pâte antirats et souris CAUSSADE, 150g  
★★★★★ 4 avis

8.25 €



CAUSSADE

Tapette antisouris CAUSSADE  
★★★★★ 26 avis

0.95 €



RETO

Répulsif ultrasons antisouris RETRO,  
lot de 3  
★★★★★ 15 avis

46.90 € / Lot



Pâte antisouris PROTECT EXPERT, 120g

★★★★★ 2 avis

5.95 €



CAUSSADE

Nasse antisouris CAUSSADE  
★★★★★ 48 avis

5.10 €



Céréales antirats, mulots et souris  
PROTECT EXPERT, 150g

★★★★★ 2 avis

6.50 €



CAUSSADE

Blocs antirats et souris CAUSSADE,  
300g  
★★★★★ 5 avis

9.80 €



CAUSSADE

Céréales antisouris CAUSSADE, 100g  
★★★★★ 2 avis

5.30 €

# Exemple: un moteur de recherche populaire

Google


souris

Tous Images Shopping Vidéos Actualités Plus Paramètres Outils


Environ 82 600 000 résultats (0,74 secondes)

Annonces · Acheter : souris


**PROMOTION**




Souris filaire HP X500  
6,29 € 8,99-€  
HP Store  
★★★★★ (126)  
Par Google



Souris Sans Fil 2.4G, VicTsing 2400 CPI...  
11,99 €  
Amazon.fr  
Par Google



HandShoeMouse filaire - Souris...  
131,99 €  
Ergo Accessoires  
Par Google



Logitech M185 Swift Grey Souris sans fil  
11,99 €  
Boulangier  
★★★★★ (9k+)  
Par Keyade

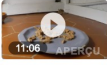
fr.wikipedia.org › wiki › Souris

**Souris — Wikipédia**


Souris » est un nom du vocabulaire courant qui peut désigner toutes sortes de mammifères rongeurs ayant généralement une petite taille, un museau pointu, ...

Mus musculus · Souris de Californie · Souris de coton · Souris kangourou

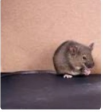

Vidéos



La souris voleuse Le retour !  
YouTube · MIAOU Jeux pour chats  
25 sept. 2018



Les Souris  
YouTube · Documentaire Animalier  
11 juil. 2020



Plus d'images

**Souris**

Animal

« Souris » est un nom du vocabulaire courant qui peut désigner toutes sortes de mammifères rongeurs ayant généralement une petite taille, un museau pointu, des oreilles rondes, un pelage gris-brun et une queue relativement longue. [Wikipédia](#)

**Espérance de vie :** Mus minotoides: 2 ans, Pachyuromys duprasi: 5 – 7 ans


**Période de gestation :** Souris grise: 20 jours, [PLUS Encyclopédie de la Vie](#)

**Poids :** Souris grise: 19 g, Apodemus sylvaticus: 23 g, [PLUS Encyclopédie de la Vie](#)


**Longueur :** Apodemus sylvaticus: 8,8 cm, [PLUS Encyclopédie de la Vie](#)

**Espèces représentatives**


Voir d'autres éléments (plus de 35)




Souris grise




Peromy... manicul...



Mus minotoides



Pachyur... duprasi



Souris à pattes blanches

Recherche par mot clé: souris

15

## Autres questions posées

Est-ce que les souris sont dangereuses ?	▼
Comment chasser les souris dans la maison ?	▼
Où il vit la souris ?	▼
C'est quoi la souris de l'ordinateur ?	▼
Commentaires	

lemagdesanimaux.ouest-france.fr › dossier-84-souris-pe... ▼

### La souris, petit rongeur que l'on déteste à la maison

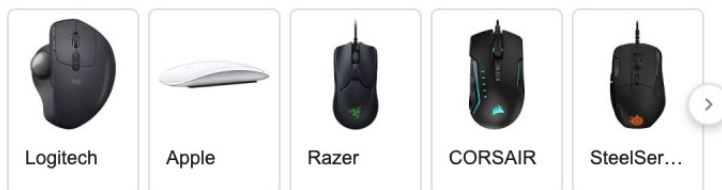
15 sept. 2019 — Parfois, on appelle communément **souris** d'autres rongeurs comme le mulot ou le campagnol, mais tous ces animaux possèdent leurs propres ...

theconversation.com › mieux-connaître-les-souris-et-arr... ▼

### Mieux connaître les souris (et arrêter de les confondre avec ...

24 nov. 2019 — S'agit-il d'un petit rat ou d'une **souris** domestique ? Pour identifier ce rongeur, il nous faut en connaître les caractéristiques morphologiques.

## Affiner par marque



www.ldlc.com › ... › Clavier, souris, saisie ▼

### Souris PC - Achat Souris PC - LDLC.com

Souris PC Logitech, Microsoft, ASUS... 570 références et 63 marques à partir de 4€ sur LDLC.com, n°1 du high-tech, élu Service Client de l'Année.

www.produit-antinuisible.com › contenu › 34-Fiche-sur... ▼

### Souris : Souris domestique ou souris grise - produit-antinuisible

Apprenez-en plus sur les **souris** : Conseils, infos pratiques, traitement et produits anti-souris. ☎ 01.40.38.38.38 ✓ Produits professionnels.

www.materiel.net › Périphériques PC › Clavier et souris ▼

### Souris PC - Achat Souris ordinateur au meilleur prix | Materiel ...

Souris PC - gamer, sans-fil, bluetooth, ergonomique. Les spécialistes de Materiel .net ont sélectionné pour vous un vaste choix de **souris** informatique pour PC ou ...

www.boulangier.com › souris ▼

### Souris - Livraison Offerte\* | Boulanger

Souris au meilleur prix ! Livraison Offerte\* - Garantie 2 ans\* - SAV 7j/7.

www.logitech.fr › fr-fr › mice ▼

### Souris Logitech pour ordinateurs PC et mac, souris filaires ou ...

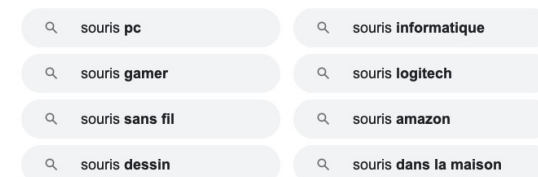
Rendez-vous sur le site Logitech pour trouver la parfaite **souris** filaire ou sans fil et améliorer votre productivité ou libérer votre créativité.

www.zooplus.fr › Rongeur & Co ▼

### Boutique pour souris, Nourriture et Accessoires souris - zooplus

Aliments & accessoires pour **souris** sur votre animalerie en ligne zooplus. Livraison gratuite dès 49 €!

## Recherches associées



Recherche par mot clé: souris



## RI – Pertinence

- La RI vise à proposer des réponses les **plus pertinentes possibles**.
- La notion de pertinence est centrale.
- En RI, un résultat constitué d'un ensemble de documents, n'est jamais exact, on mesure son degré de pertinence.
- On distingue:
  - **Les faux positifs** (false positive): les documents non pertinents inclus dans le résultat, sélectionnés à tort.
  - **les vrais positifs** ( true positive): les documents pertinents inclus dans le résultat
  - **les faux négatifs** (false negative): les documents pertinents qui ne sont pas inclus dans le résultat.
  - **les vrais négatifs** ( true negative): les documents non pertinents non inclus dans le résultat

	Pertinent	Non pertinent
Ramené (positif)	vrais positifs	faux positifs
Non ramené (négatif)	faux négatifs	vrai négatifs

## RI – Pertinence

- Plusieurs pertinences:
  - Thématique (topical): relation entre le sujet exprimé dans la requête et le sujet couvert dans le document.
  - Contextuelle (situation) : relation entre la tâche, le problème posé par l'utilisateur, la situation de l'utilisateur et l'information retrouvée.
  - Cognitive : relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée. elle est liée aux connaissances et à la perception de l'utilisateur envers un thème de sa requête. Elle est dite cognitive car elle permet d'améliorer la connaissance de l'utilisateur via le contenu renvoyé au cours de sa recherche.
- Processus subjectif (humain), dépend de plusieurs facteurs
  - difficile à automatiser

## RI – Mesure de qualité

- Afin de mesurer la qualité d'un SRI, deux indicateurs formels sont définis.
- On note:  $t_p(r)$  les vrais positifs et  $f_p(r)$  les faux positifs dans un résultat  $r$
- On note:  $f_n(r)$  : le nb de documents faux négatifs.

- **La précision:** La proportion des vrais positifs dans le résultat  $r$ .

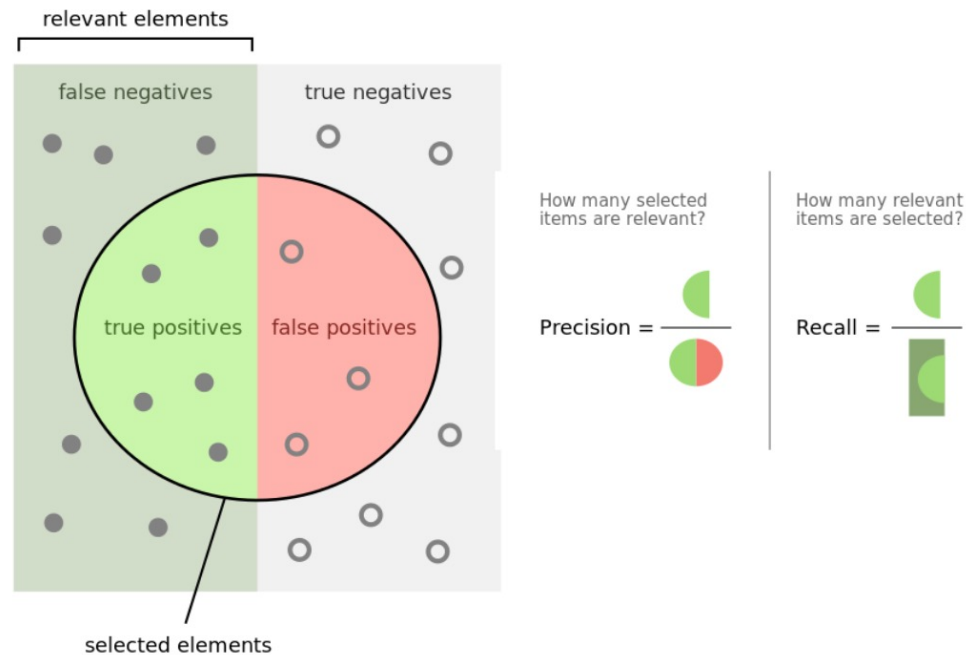
$$\text{Précision} = \frac{t_p(r)}{t_p(r) + f_p(r)} = \frac{t_p(r)}{|r|}$$

- **Le rappel:** la proportion des documents pertinents qui sont inclus dans  $r$ .

$$\text{Rappel} = \frac{t_p(r)}{t_p(r) + f_n(r)}$$

## RI – Mesure de qualité

- Précision=1 : absence totale de faux positifs
- Précision=0 : aucun document pertinent
- Rappel=1 : tous les documents pertinents sont dans r
- Rappel=0: aucun document pertinent n'est dans r



## RI – Mesure de qualité

- Précision et rappel sont très difficiles à optimiser simultanément.
- Augmenter le rappel -> ajouter plus de document dans le résultat, diminuer la précision
- Un résultat avec toute la collection input a un rappel 1, mais une précision qui tend vers 0.
- On améliore la précision, si on ne garde que les documents dont la pertinence est sure, mais on augmente le risque de faux négatifs donc un rappel dégradé
- Evaluer un SRI est un tâche complexe et fragile car elle repose sur des enquêtes impliquant des utilisateurs et sur un échantillon.

## Exemple de Mesure de qualité d'un moteur de recherche

- Supposons que dans notre échantillon de données on a 90 pages pertinentes
- Un moteur de recherche retourne 60 pages web dont seulement 30 sont pertinentes,
  - $t_p(r) = 30$
  - $f_p(r) = 60-30=30$

Donc

$$\text{Précision} = \frac{t_p(r)}{t_p(r)+f_p(r)} = \frac{30}{60} = \frac{1}{2}$$

$$\text{Rappel} = \frac{t_p(r)}{t_p(r)+f_n} = \frac{30}{30+60} = \frac{1}{3}$$

# Mesure de qualité d'un moteur de recherche

- Dans certaines applications, le rappel est beaucoup plus important que la précision.
- Exemple : Trouver les courriels qui ne sont pas des pourriels:
  - ✓ il est très important de trouver tous les courriels qui ne sont pas des pourriels ;
  - ✓ il est moins grave que certains pourriels survivent au filtrage.

## Exemple de Recherche plein texte

- Considérons comme exemple l'ensemble de documents ci-dessous:

*d1: Le loup est dans la bergerie.*

*d2: Le loup et le trois petits cochons*

*d3: Les moutons sont dans la bergerie.*

*d4: Spider Cochon, Spider Cochon, il peut marcher au plafond.*

*d5: Un loup a mangé un mouton, les autres loups sont restés dans la bergerie.*

*d6: Il y a trois moutons dans le pré, et un mouton dans la gueule du loup.*

*d7: Le cochon est à 12 le Kg, le mouton à 10 E/Kg*

*d8: Les trois petits loups et le grand méchant cochon*



# Recherche plein texte

- Le besoin: Rechercher tous les documents parlant de loups, de moutons mais pas de bergerie
- Solutions:
  1. Parcourir tous les documents et tester la présence des mots-clés : Long face à un nombre conséquent de documents volumineux.
  2. Une autre solution possible:
    - ✓ Créer une structure où les données sont organisées en matrices ( d'incidence) représentant l'occurrence ( binaire à 1) de chaque mot dans chaque document.
    - ✓ Choix: mots en lignes et documents en colonnes ou l'inverse: deux techniques à étudier.
- A noter l'emploi de la notion **terme (token en anglais)** différente de celle de « mot ».  
Le **vocabulaire**, parfois appelé **dictionnaire**, est l'ensemble des termes sur lesquels on peut poser une requête.

## Exemple de Recherche plein texte

- **Choix1:** Une matrice d'incidence avec les documents en ligne. On se limite au vocabulaire suivant: {« loup », « mouton », « cochon », « bergerie », « pré », « gueule »}.

	<b>loup</b>	<b>mouton</b>	<b>cochon</b>	<b>bergerie</b>	<b>pré</b>	<b>gueule</b>
$d_1$	1	0	0	1	0	0
$d_2$	1	0	1	0	0	0
$d_3$	0	1	0	1	0	0
$d_4$	0	0	1	0	0	0
$d_5$	1	1	0	1	0	0
$d_6$	1	1	0	0	1	1
$d_7$	0	1	1	0	0	0
$d_8$	1	0	1	0	0	0

Matrice d'incidence

## Exemple de Recherche plein texte

- Considérons *les vecteurs d'incidence* de chaque terme contenu dans la requête, soit les colonnes dans notre représentation :
  - Loup: 11001101
  - Mouton: 00101110
  - Bergerie: 10101000
- On effectue un « ET logique » sur les vecteurs de *Loup* et *Mouton* et on obtient 00001100
- Ensuite, effectuons un « ET logique » du résultat avec le *complément* du vecteur de *Bergerie* (01010111):
  - 00001100 ET 01010111 on obtient 00000100
- Nous déduisons que la réponse est limitée au document d6, puisque la 6e position est la seule où il y a un “1”.

## Exemple de Recherche plein texte

### ■ Limites de la solution choix 1:

- Un million de documents, mille mots chacun en moyenne (ordre de grandeur d'une encyclopédie en ligne bien connue)
- Disons 6 octets par mot, soit 6 Go
- Disons 500 000 termes distincts (ordre de grandeur du nombre de mots dans une langue comme l'anglais)
- La matrice a 1 000 000 de lignes, 500 000 colonnes, soit  $500 * 10^9$ , soit 62 GO. Elle ne tient pas en mémoire de nombreuses machines.

# Recherche plein texte: index inversé

## ■ Choix2 : Les indexes inversés:

Une matrice d'incidence avec les termes en ligne.

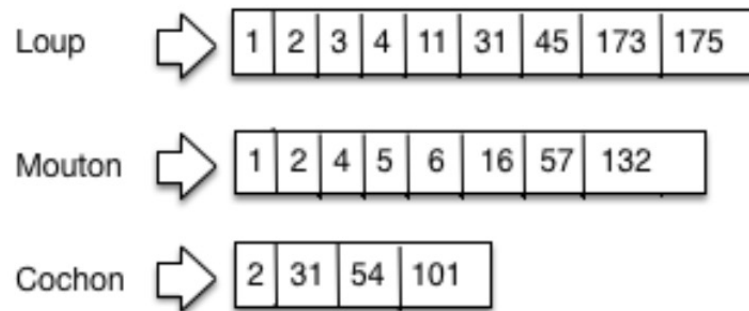
Loup	➡	<table><tr><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td><td>1</td></tr></table>	1	1	0	0	1	1	0	1
1	1	0	0	1	1	0	1			
Mouton	➡	<table><tr><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td></tr></table>	0	0	1	0	1	1	1	0
0	0	1	0	1	1	1	0			
Cochon	➡	<table><tr><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td><td>1</td></tr></table>	0	1	0	1	0	0	1	1
0	1	0	1	0	0	1	1			
Bergerie	➡	<table><tr><td>1</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr></table>	1	0	1	0	1	0	0	0
1	0	1	0	1	0	0	0			
Pré	➡	<table><tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr></table>	0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	0			
Gueule	➡	<table><tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr></table>	0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	0			

Inversion de la matrice d'incidence

# Recherche plein texte: index inversé

## ■ Choix2 : Les indexes inversés:

- Pour chaque ligne, seuls les identifiants du document où le terme existe sont mentionnés. Pour cela, on place dans les cellules l'*identifiant* du document (*docId*).
- Chaque liste est triée sur l'identifiant du document.



Index Inversé

- Tous les moteurs de recherche utilisent la structure d'index inversé.
- Excellentes propriétés pour une recherche efficace, avec en particulier des possibilités importantes de compression des listes associées à chaque terme.

# Recherche plein texte: index inversé

- Nous appellerons :
  - Le dictionnaire (dictionary): l'ensemble des termes de l'index inversé;
  - Le répertoire: la structure qui associe chaque terme à l'adresse de la liste inversée (posting list) associée au terme.
  - Un élément de la liste inversée est appelé entrée
- Le répertoire est toujours en mémoire, ce qui permet de trouver très rapidement les listes impliquées dans la recherche.
- Les listes inversées sont, autant que possible, en mémoire, sinon elles sont compressées et stockées dans des fichiers (contigus) sur le disque.

# Index inversé: Opération de recherche

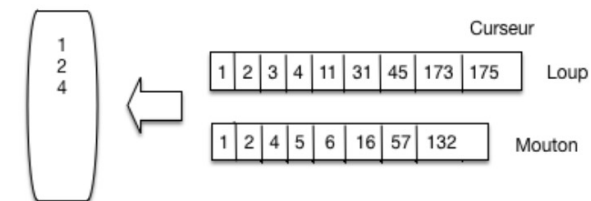
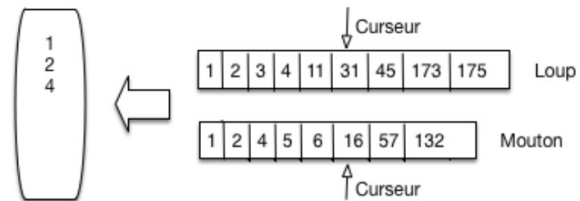
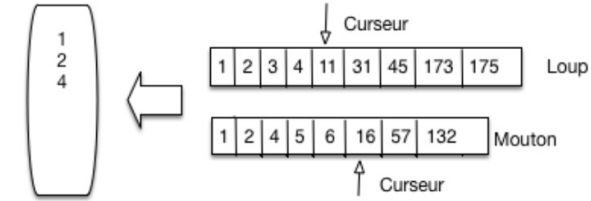
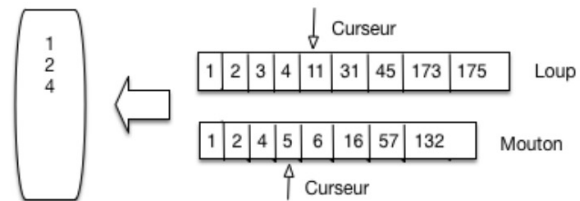
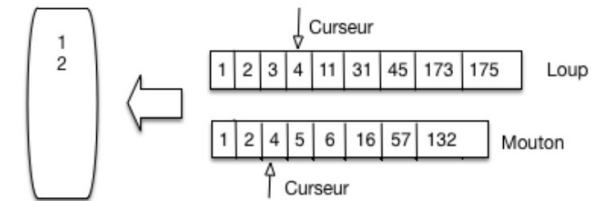
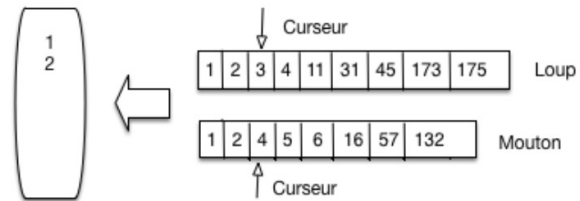
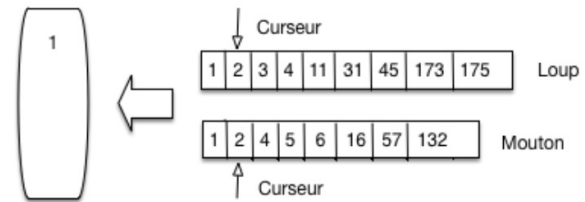
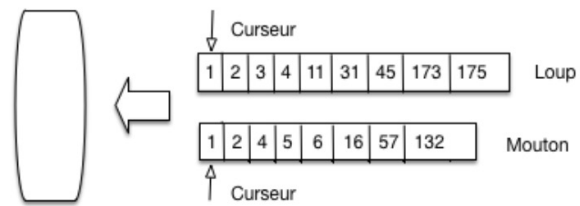
- L'algorithme employé: une fusion (« merge ») de liste triées.
- Un parcours parallèle et séquentiel des listes en une seule fois: une technique très efficace.
- Grâce au tri des listes par rapport à l'identifiant du document, le parcours est unique.

## Fonctionnement:

Exemple: rechercher les documents contenant les termes 'mouton' et 'loup':

- *Un curseur par terme, positionné au début de chaque liste.*
- *Une comparaison sur le docID est faite:*
  - ü Si égalité: le docID est placé dans les résultats
  - ü Sinon, le curseur qui pointe sur le docID le plus faible sera incrémenté.





# Index inversé: Algorithme

```
// Fusion de deux listes l1 et l2
function Intersect($l1, $l2)
{
    $résultat = [];
    // Début de la fusion des listes
    while ($l1 != null and $l2 != null) {
        if ($l1.docId == $l2.docId) {
            // On a trouvé un document contenant les deux termes
            $résultat += $l1.docId;
            // Avançons sur les deux listes
            $l1 = $l1.next; $l2 = $l2.next;
        }
        else if ($l1.docId < $l2.docId) {
            // Avançons sur l1
            $l1 = $l1.next;
        }
        else {
            // Avançons sur l2
            $l2 = $l2.next;
        }
    }
}
```

## Index inversé: Optimisation

- Si l'on souhaite effectuer une recherche entre 'loup' et 'mouton' et 'chèvre', comment optimiser la recherche?
- On analyse la taille des listes et l'on commence par faire l'intersection des deux plus petites listes pour constituer une liste résultat, puis effectuer l'intersection entre la liste résultat de l'étape précédente et la troisième liste.
- D'autres pistes d'optimisation existent :
  - *stocker l'une des listes en mémoire et calculer les intersections à la volée, en lisant depuis le disque*
  - ....

# Documents de référence

1. Cours de bases de données documentaires et distribuées: <http://b3d.bdpedia.fr>
2. Système de recherche d'information étendue basé sur une projection multi-espaces , thèse présentée par HANNECH AMEL -2018
3. Recherche d'information, applications, modèles et algorithmes , Massih-Reza Amini et Eric Gaussier, Eyrolles 2<sup>ème</sup> édition.
4. Introduction à la Recherche d'Information S1: les principes , Raphaël Fournier-S'niehotta, Philippe Rigaux