

Moteurs de Recherche

M2 IAD Pro



Nicolas Usunier

nicolas.usunier@lip6.fr

Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6

Transparents de cours: Massih-Réza Amini, Nicolas Usunier

Plan

- Description de l'UE
- Problématiques de Recherche d'Information,
- Modèle booléen non ordonné
 - Appariement requête-document,
 - Indexation.

Description de l'UE

- ❑ Présentation des modèles de recherche d'information sur des documents textuels,
 - Modèles d'appariement pour les documents plats,
 - Utilisation des liens dans une collection de documents hypertextes,
 - Utilisation des méta-informations, combinaison de modèles de recherche,
- ❑ Cours effectué par des intervenants d'Exalead début Mars.

Description de l'UE

□ Evaluation par un projet

- réalisé au fil des TMEs,


- objectif :

 - réaliser un moteur de recherche sur des petites collections de documents,

 - comparer les performances des modèles vus en cours.

Moteurs de recherche

Web [Images](#) [Maps](#) [Actualités](#) [Vidéo](#) [Gmail](#) [plus ▼](#)





Rechercher dans : ☒ Web ☐ Pages francophones ☐ Pages : Fr

Web

[Contactez un expert Google pour votre visibilité !](#)
La visibilité de votre site mérite un professionnalisme sans faille... nous mettrons en oeuvre les mesures nécessaires au positionnement de votre site.
[kalitic-referencement.fr/expert-google.php](#) - 13k - [En cache](#) - [Pages similaires](#)

[David DURAND PICHARD, Expert référencement Google](#)
DDP-FRANCE vous propose la création de sites internet et le référencement dans les moteurs de recherche. **Expert google**, l'optimisation de votre site sera ...
[www.ddp-france.com/expert-google.php](#) - 13k - [En cache](#) - [Pages similaires](#)

[expert google](#) : UCatchit : Tel  +33 1 53 62 00 60 

Pour devenir un **expert de google** notre CMS de référencement vous donne tous les conseils en ... Devenir un **expert google** est accessible à tous maintenant. ...
[www.ucatchit.com/ucatchit_expert_google.html](#) - 8k - [En cache](#) - [Pages similaires](#)

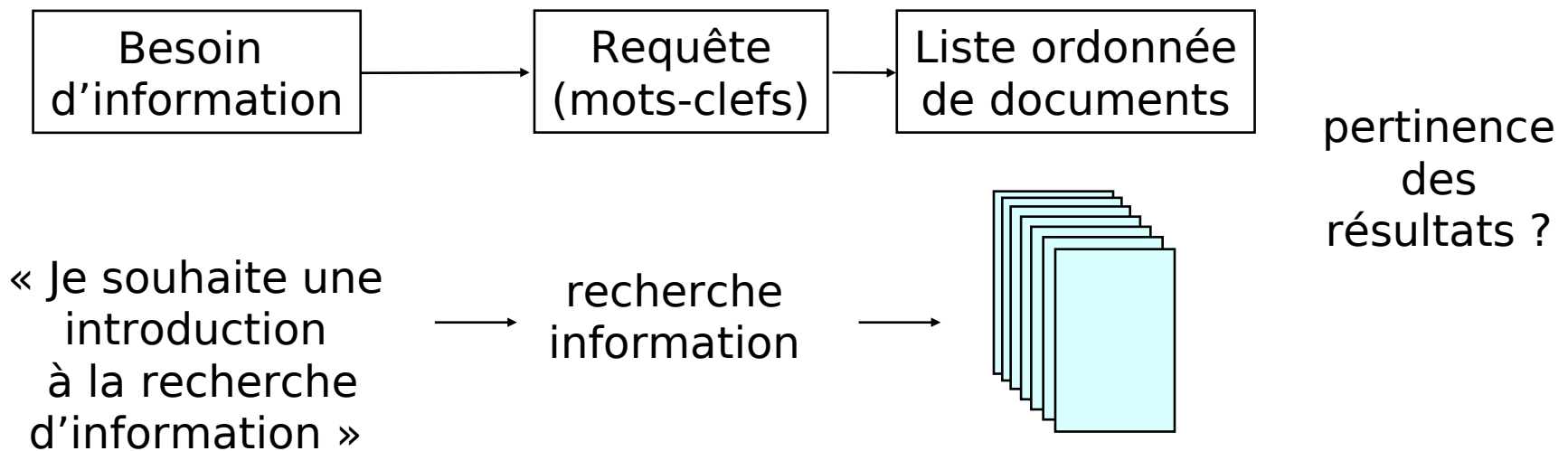
[Expert Google](#)
24 jul 2008 ... Suis-je devenu un **expert Google** ? Rien n'est moins sûr, l'algorithme le plus protégé de la planète ne peut être découvert si facilement. ...
[www.ericnguyen.net/Expert-Google.html](#) - 15k - Il y a 7 heures - [En cache](#) - [Pages similaires](#)

Problématique

- ❑ Répondre à un *besoin d'information* de l'utilisateur :
 - « je recherche une introduction à la recherche d'information »
- ❑ recherche de l'information *pertinente*...
 - Par exemple la page « Information Retrieval » de Wikipedia
- ❑ ...dans une grande collection de données
 - Texte, image, ...
 - Ex: 1000 milliards de pages Web indexées par Google en 2008

Recherche *ad-hoc*: principe

- ❑ Moteurs de recherche *ad-hoc*:
 - une liste de résultats par requête
 - *ad-hoc*: « *formed or used for specific or immediate problems or needs* »



Moteurs de recherche vs Bases de données (1)

□ Bases de données:

■ Données structurées:

la sémantique des champs est connue

■ Langages de requêtes très expressif

`SELECT * FROM web WHERE text LIKE "%recherche
%information%"`

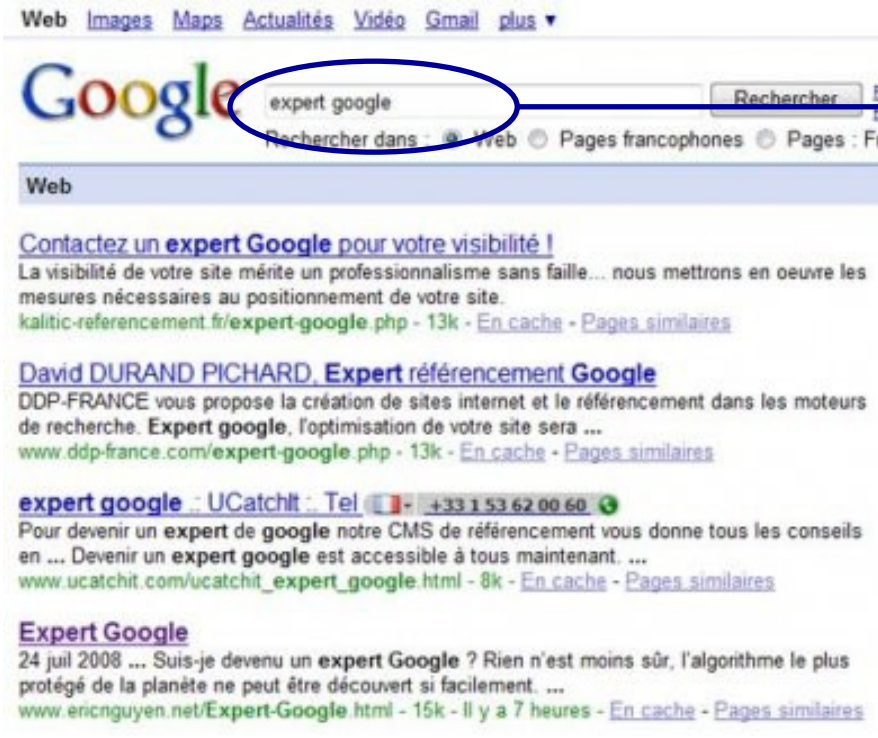
■ besoin d'information exprimé de façon exacte

donc pas de notion de pertinence du résultat

Moteurs de recherche vs Bases de données (2)

- Moteurs de recherche
 - Information peu structurée
 - HTML, XML, fichiers textes,
 - quelques méta-informations (titre, auteur, ...),
 - Langage de requêtes simple,
 - Besoin d'information exprimé partiellement.

Moteurs de recherche : principe



requête

Documents triés
par scores
décroissants

Échelle des
scores

Recherche par mots-clefs

- ❑ La requête décrit le *contenu* du résultat souhaité le plus souvent sous forme de mots-clefs
- ❑ Le moteur de recherche renvoie les documents dont le contenu correspond *le plus* à la description
 - nécessité d'établir des indicateurs de pertinence d'un document par rapport à une requête,
 - nécessité de les combiner.
- ❑ Ex : les mots-clefs apparaissent dans le titre du document, les mots-clefs apparaissent fréquemment, indicateurs relatifs à la collection de document (PageRank), ...

Notions fondamentales

- ❑ Besoin utilisateur: exprimé par la requête
- ❑ Unité d'information: ex: document, paragraphe, image
- ❑ Fonction d'appariement:
 - comment calculer le score d'un document par rapport à une requête ?
- ❑ Notion de pertinence:
 - un document répond-il à la requête ?
- ❑ Le rang d'un document dans la liste triée:
 - Défini par l'ordre de présentation des documents à l'utilisateur
 - Le premier document a le rang 1
- ❑ Le rappel:
 - Quelle proportion des documents pertinents a été trouvée ?
- ❑ La précision:
 - Parmi les documents renvoyés, quelle est la proportion de documents pertinents ?

Rappel et Précision

□ Rappel:

- Le pourcentage de tous les documents pertinents trouvés par le système de recherche.

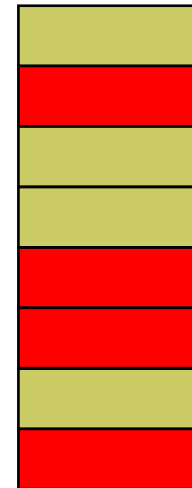
$$Rap = \frac{nb\ de\ docs\ pertinents\ retournés}{nb\ de\ docs\ pertinents}$$

□ Précision:

- Le pourcentage de documents retournés par le système qui sont pertinents.

$$Pre = \frac{nb\ de\ docs\ pertinents\ retournés}{nb\ de\ docs\ retournés}$$

Docs retournés, Docs per=10



Rap=4/10, Pre=4/8



Pertinent



Non-pertinent

Recherche Ad-hoc: Web vs. Autres collections

□ Web

- Moyen d'accès le plus utilisé,
- Source d'information dynamique

□ Autres bases de données

- Bibliothèques électroniques, Wikipedia
- Journaux (Le Monde, Wall Street Journal, ...)
- Intranets d'entreprise, bases d'articles scientifiques, bases de brevets...

Recherche Ad-hoc:

Web vs. Autres collections (2)

- Le Web est un environnement à forte précision et faible rappel
 - Beaucoup de demandes d'information sont :
 - « trouver une réponse / un site / un service » ,
 - Les utilisateurs ne vont pas chercher plus loin que le premier écran pour leur réponse.

- Le Web est une source énorme et variée de données
 - Un appariement simple et exact d'une requête à un document pourrait marcher
 - Plus complexe dans une collection restreinte de documents,
 - Nécessité de prendre en compte la structure du Web (lien hypertextes) et des documents (balises « titre », ...)
 - Pas forcément nécessaire dans des collections plus petites

Appariement requête-documents

- Appariement **Exact** (AE) avec la requête
 - Chaque document renvoyé doit correspondre exactement à la requête
 - Chaque mot de la requête doit se trouver dans les documents renvoyés
 - Rapide, bon pour de gros volumes de données
 - Ex: le Web

- **Meilleurs** Appariements (MA) avec la requête
 - Les bons documents apparaissent en tête de tri,
 - Usuellement, les bons documents ne s'apparient pas exactement avec la requête,
 - La requête est trop pauvre pour exprimer l'attente de l'utilisateur ou bien elle ne reflète pas toute(s) l(es) idée(s) présente(s) dans les documents.

Appariement exact : Modèle booléen

- ❑ recherche de documents s'appariant de façon exacte avec la requête.
- ❑ requête = expression logique ET..OU..NON.
- ❑ Exemples:
 - « information » ET « retrieval »
signifie : trouver tous les documents contenant les mots « information » et « retrieval »
 - « information » ET « retrieval » ET NON « wikipedia »
signifie : trouver tous les documents contenant les mots « information » et « retrieval », mais pas « wikipedia »

Modèle booléen non ordonné: recherche et indexation (1)

- ❑ Il n'est pas nécessaire de stocker le texte exact des documents:
 - uniquement l'information présence/absence des mots est nécessaire
- ❑ Il est nécessaire de définir la notion de *token* (\approx mots):
 - Ex: « Information » et « information » sont-ils le même « mot » ?
- ❑ Il est nécessaire d'accéder rapidement aux documents contenant les mots de la requête:
 - Indexation, et création de l'index inversé

Modèle booléen : indexation

□ Etape 1:

Création du vocabulaire:

Ensemble des *tokens* considérés pour la recherche

- Ex: considérer toutes les séquences de lettres [a-zA-Z]+, puis tout mettre en minuscule

□ Etape 2:

création de la matrice documents x tokens

1 ligne par document, 1 colonne par token

à la case (i,j):

1 si le i-ième document contient le j-ième token, sinon 0

La matrice documents x tokens s'appelle l'**index**

La matrice transposée tokens x documents est l'**index inversé**

Modèle booléen :

Exemple d'index inversé (1)

- 18 documents sur un vocabulaire de 32 mots:

A1	business, depression, tax, unemployment
A2	business, economy, market
A3	business, market, production
A4	depression, liquidation, prosperity, recovery
A5	compensation, unemployment, welfare
A6	business, market, price
A7	benefit, compensation, unemployment
B1	basin, fault, submergence
B2	depression, fault
B3	basin, drainage, valley
B4	depression, drainage, erosion
B5	basin, drainage, volcano
C1	counseling, illness, mental
C2	counseling, depression, emotion
C3	depression, rehabilitation, treatment
C4	disturbance, drug, illness
C5	distractibility, illness, talkativeness
C6	counseling, drug, psychotherapy

Modèle booléen :

Exemple d'index inversé (2)

Terms	Documents																	
	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5	C6
basin	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0
benefit	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
business	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
compensation	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
counseling	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
depression	1	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0
distractibility	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
disturbance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
drainage	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
drug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
economy	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
emotion	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
erosion	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
fault	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
illness	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0
liquidation	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
market	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
mental	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
price	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
production	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
prosperity	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
psychotherapy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
recovery	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rehabilitation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
submergence	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
talkativeness	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
tax	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
treatment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
unemployment	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
valley	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
volcano	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
welfare	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Stockage: matrice creuse. Exemple: basin -> {B1, B3, B5}

Modèle booléen non ordonné: recherche avec l'index inversé

- Sur l'exemple précédent, considérons la requête:

- « business » ET « depression »

business -> {A1, A2, A3, A6}

depression -> {A1, A4, B2, B4, C2, C3}

- Réponse: {A1}

- En général:

- trouver des documents est équivalent à des unions/intersections d'ensembles

- Le temps de traitement d'une requête ne dépend pas du nombre de documents (ou de termes) dans la collection

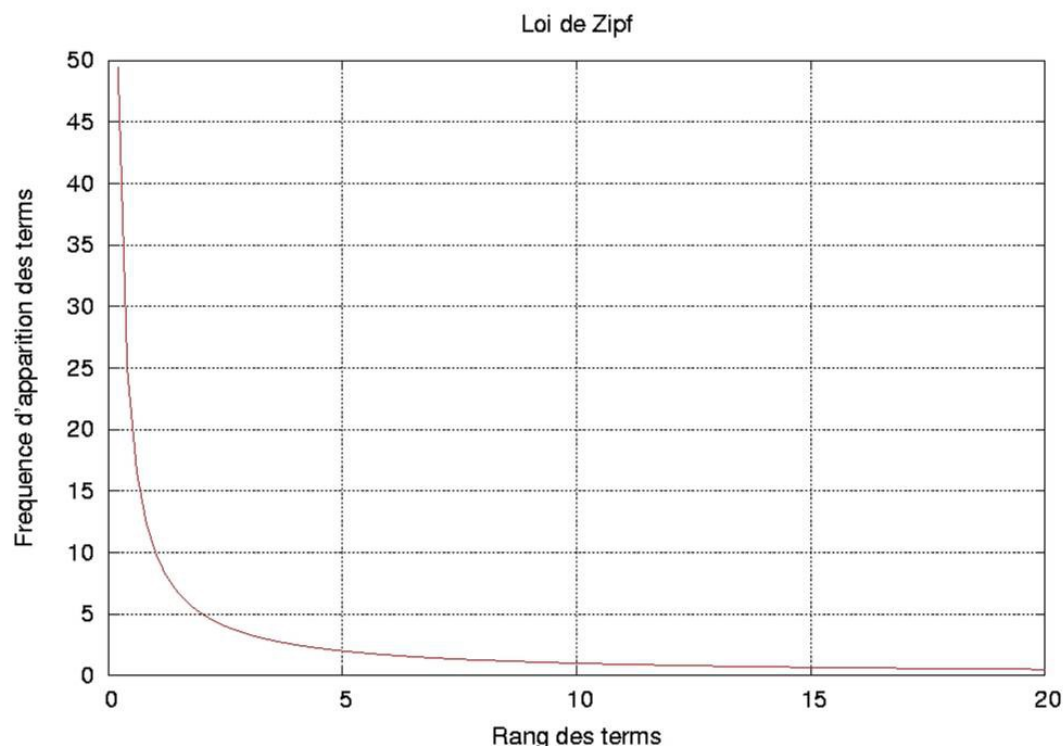
Retour sur l'indexation : détermination du vocabulaire

- Quelles règles pour déterminer les *tokens* ?
- Quels sont les mots importants ?
 - Quelles sont leurs caractéristiques ?
 - Comment déterminer le dictionnaire ?

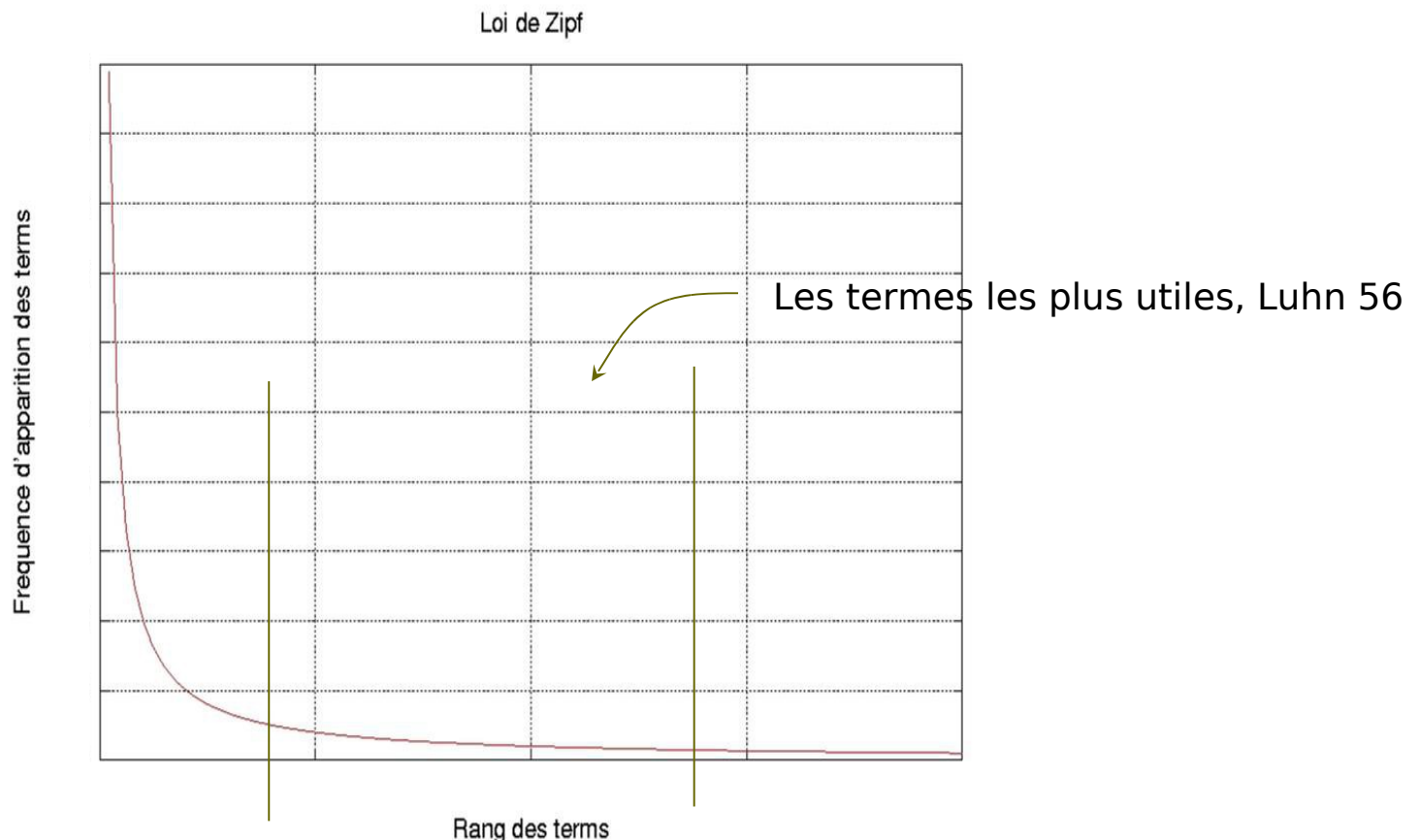
Loi de Zipf (1930)

- La probabilité d'apparition du $n^{\text{ième}}$ mot le plus fréquent dans une collection de n'importe quelle langue est approximativement inversement proportionnelle à n .

- $$P_n = \frac{C}{n}$$



Termes utiles: d'après Luhn (1956)



Détermination du vocabulaire (1)

□ Quelles unités conserver pour l'indexation ?

■ *Stop words - anti-dictionnaire*

- Les mots les plus fréquents de la langue "*stop words*" n'apportent pas d'information utile (ex: et, ou, le, ...)
- Les "*stop words*" sont les mots les plus fréquents d'un corpus + peuvent dépendre du domaine
- L'ensemble des mots éliminés est conservé dans un *anti-dictionnaire* (e.g. 500 mots).

■ Les mots les moins fréquents peuvent aussi être supprimés

- Ex: fautes d'orthographe, ...

Détermination du vocabulaire (2)

- Lexémisation (*stemming*) et lemmatisation
 - Utilisation d'une forme canonique pour représenter les variantes morphologiques d'un mot
 - e.g. dynamic, dynamics, dynamically, ...seront représentés par un même mot
 naviguer, naviguant, navireidem
 - Techniques :
 - systèmes itératifs à base de règles simples (e.g. pour l'anglais Porter stemming -largement employé) : on établit une liste de suffixes et de préfixes qui sont éliminés itérativement.
 - méthodes à base de dictionnaires mot - forme canonique. Intérêt : langue présentant une forte diversité lexicale (e.g. français)
- Regroupement de mots sémantiquement proches (ex: synonymes)