

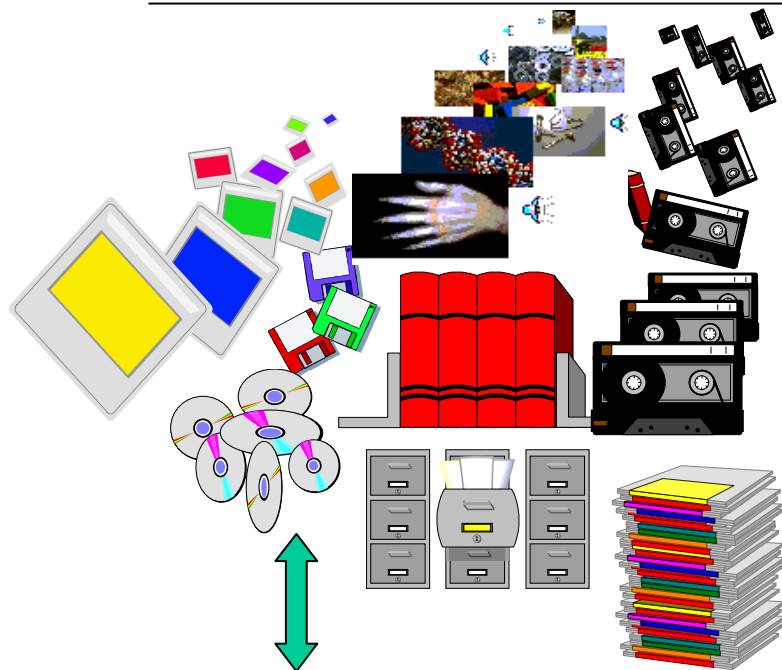


Systemes de recherche d'informations

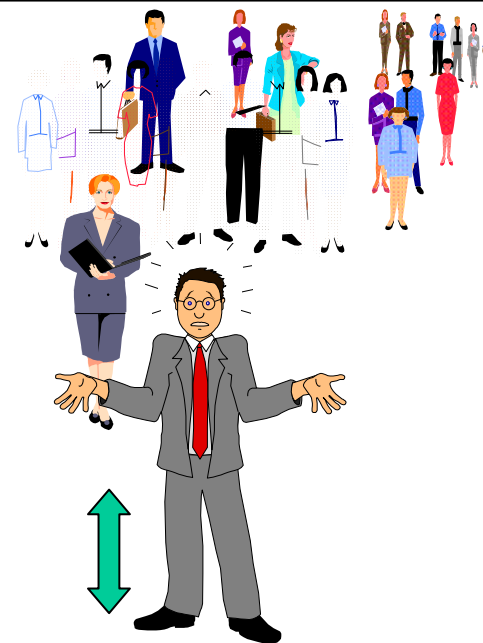


Plan

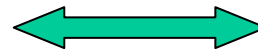
- Problématique et objectifs
- Processus d'indexation et de recherche
- Modèles et approches pour la recherche d'information
- Évaluation d'un système d'indexation
- Côté web : Principe de google



Documents
indexés



Requêtes



indexation

langage d'indexation

Processus d'indexation

interrogation

interface

fonction de correspondance

Méthodes d'indexation manuelle (1/4)

La forme la plus répandue d'indexation : le rôle de l'indexeur est d'attribuer au document à archiver un certain nombre de descripteurs :

- **Mots-clés unitermes**

Descripteurs formés d'un seul mot. Par conjonction de plusieurs unitermes, on obtient des expressions composées.



Méthodes d'indexation manuelle (2/4)

- **Descripteurs composés :**

Constitués d'expressions de deux ou trois termes. On peut utiliser des expressions de différents types :

- nom-adjectif (ex: Droit social)
- nom-complément du nom (ex: Histoire de la musique)
- les termes possédant un trait d'union (ex: Libre-échange)
- des termes avec rejet (ex Boole, algèbre de) bien adaptées pour les catalogues manuels
- des termes avec un qualificatif (ex: Mercure (métal), Mercure (planète))



Méthodes d'indexation manuelle (3/4)

- **Descripteurs structurés**

Un descripteur structuré contient plusieurs informations sous une même entrée dite "vedette".

On fait succéder les descripteurs dans l'ordre suivant :

- tête de vedette, significative du sujet
- sous-vedette de point de vue
- sous-vedette de localisation géographique
- sous-vedette de localisation chronologique
- sous-vedette de forme (dictionnaire, bibliographie, congrès)



Méthodes d'indexation manuelle (4/4)

- **Indices de classification**

La classification permet de situer un document dans un système de domaine de connaissances :

Les classifications hiérarchiques :

5 Science

51 Mathématique

512 Algèbre

513 Arithmétique

Méthodes d'indexation automatique

- **Indexation par des méthodes sémantiques**
 - relations sémantiques entre termes
 - représenter le document dans un langage de description tenant compte des relations sémantiques
 - extension de la notion de thesaurus
- **Méthodes linguistiques**
 - analyse plus ou moins profonde : plusieurs types de traitements
 - Avantage
 - analyse fine des unités de sens du texte
 - Inconvénients
 - nécessité d'un dictionnaire complet
 - fonctions d'analyse linguistique lourdes et coûteuses
- **Méthodes statistiques**
 - basées sur le calcul de fréquences de termes

Types d'indexation automatique

- **Indexation orientée document**
 - L'objectif est de résumer ou de présenter le contenu de chaque document.
- **Indexation orientée requête**
 - Pour chaque document, refléter les requêtes pour lesquelles il est pertinent : l'indexation d'un document doit alors représenter les raisons pour lesquelles un utilisateur consulte ce document (i.e : confronter chaque document de la base à une liste de requêtes prédéfinies)

Quelle indexation choisir ?

Les indexations proposées dans les systèmes de recherche d'informations doivent être mixtes. En effet, les besoins des systèmes sont doubles :

- Afin de servir le spectre **le plus large d'utilisateurs**, le système doit disposer du maximum d'informations sur les documents. De ce point de vue, l'indexation doit donc être orientée document.
- Afin de servir au mieux **un utilisateur**, de ce point de vue, l'indexation doit être orientée requête.

Paramètre du langage d'indexation

➤ Vocabulaire d'indexation

- **Les *langages libres*, ou *non contrôlés*** : langages évolutifs car ils sont constitués « **a posteriori** » sur la base de l'indexation en langage naturel des documents déjà enregistrés dans une collection.
- **Les *langages contrôlés*** : langages figés car construits « **a priori** » avant de commencer à indexer des documents d'une collection. Ces langages éliminent tout nouveau mot ou concept.

Paramètres du langage d'indexation

- Pré-coordination ou post-coordination du langage d'indexation
 - Coordination « a priori » des termes qui sont reliés entre eux (**avant l'interrogation**).
 - Coordination « a posteriori » des termes qui sont reliés entre eux (**après l'interrogation**).

Les modèles d'indexation automatique

- Le modèle booléen (exple : Medline)
 - Ce modèle représente les documents et les requêtes sous forme d'une **proposition de termes** et **d'opérateurs logiques**, et fournit une fonction de correspondance basée sur l'implication $D=Q$
 - Une requête est une expression logique composée de termes connectés par les **opérateurs logiques** \wedge, \vee et \neg
 - Simplicité, bonnes performances quantitatives, même sur de très grandes collections de documents. Mais limitations sur le plan qualitatif (évaluation binaire)

Les modèles d'indexation automatique

➤ Le modèle Probabiliste

- Mesure la probabilité de pertinence d'un document par rapport à une requête
- Utilise 2 probabilités pour chaque document :
 - la probabilité de pertinence du document D , $P(per/D)$,
 - la probabilité de non pertinence du document D , $P(nonper/D)$
- La fonction de recherche $g(D)$ donne un ordonnancement des documents: $g(D) = (P(per/D)/P(nonper/D))$
Probabilités calculées en fonction des termes d'indexation de la base.

Les modèles d'indexation automatique

- Le modèle vectoriel (SMART[Salton 83])
 - Un document D_i est représenté par un vecteur de *termes d'indexation*. $D_i = (W_{i1}, \dots, W_{ij}, \dots, W_{in})$ où n représente le nombre de termes d'indexation et W_{ij} la pondération du terme t_j dans le document D_i .
 - $W_{ij} = FLOC_{ij} / FDOC_j$, où
 $FLOC_{ij}$: Fréquence locale de t_j dans le document D_i ,
 $FDOC_j$: Fréquence documentaire de t_j dans tout le corpus .
 - Vecteur requête $Q = (k_1, \dots, k_j, \dots, k_n)$
 - Similarité $(D_i, Q) = (\sum_j W_{ij} k_j / (\sum_j W_{ij}^2 \sum_j k_j^2))^{1/2}$



Evaluation d'un SRI

- Utilisation de collections de tests
 - "étalon" de mesure de comparaison (relatif)
- Besoins exprimés en détails
- Sélection manuelle des meilleures réponses
- Confrontation automatique du SRI
- Mesures normalisées de comparaison

Evaluation

- **Précision relative**
 - nombre de documents trouvés jugés pertinents par l'utilisateur
- **Rappel relatif**
 - proportion de documents trouvés jugés pertinents par rapport au nombre global de documents pertinents
- **Digression**
 - nombre de documents trouvés et jugés non pertinents
- **Silence**
 - nombre de documents jugés pertinents et non retrouvés

Bouclage de pertinence

$$GainDeBouclage = \frac{|Pertinent_{u,i} \cap Trouvés_i|}{|\bigcup_{j=1}^i Pertinents_{u,j} \cap Trouvés_j|}$$

$$PerteDeBouclage = \frac{|NonPertinent_{u,i} \cap Trouvés_i|}{|\bigcup_{j=1}^i NonPertinents_{u,j} \cap Trouvés_j|}$$

Gain de bouclage désigne le nombre de nouveaux documents retrouvés à l'étape i et jugés pertinents par l'utilisateur.

Perte de bouclage désigne le nombre de nouveaux documents retrouvés à l'étape i et jugés non pertinents par l'utilisateur.



Côté Web : Google

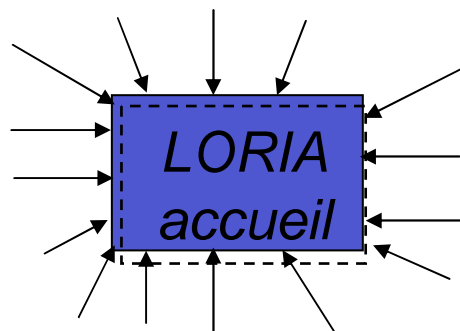
- Larry Page et Sergey Brin, deux étudiants provenant respectivement des facultés de Mathématique et d'Informatique de l'Université de Stamford.
- Ils conçoivent et gèrent un projet de recherche qui mène à la création de Google.
- Les années 1999-2000 marquent les premiers succès

Google : <http://www.google.com/>

- Principes de fonctionnement

- Idée : exploiter les liens hypertextes (à la manière de l'analyse des citations en science de l'information) avec l'hypothèse que les liens de citations entre pages WEB expriment une approbation

- Exemple :

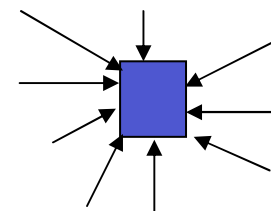


Google : <http://www.google.com/>

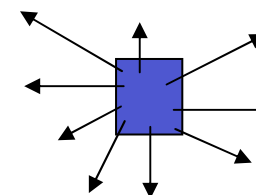
- **Algorithme de classement**

- 2 types de pages :

- les pages de références
(i.e. pages fréquemment citées)



- les pages pivots
(i.e. pages contenant un grand nombre de liens)



- Définition récursive de l'importance des pages

« plus une page de référence est pointée par de bonnes pages pivots, plus elle sera une bonne page de référence »

« plus une page pivot pointera de bonnes pages de références, plus, plus elle sera une bonne page pivot »

Dans la pratique, comment ça marche?

- La formule proposée [Brin, 1998] tient compte de la probabilité de suivre effectivement les liens
- La probabilité d'arriver à la page P sans suivre de lien est donc de $(1-d)$.
 - Avec d , facteur d'amortissement $0 < d < 1$
- $PR(P)$: PageRank de la page P
- $PR(P) = (1-d) + d \left[\frac{PR(P_1)}{C(P_1)} + \dots + \frac{PR(P_m)}{C(P_m)} \right]$
 - $C(P_j)$ = nombre de liens sortant de la page P

Exemple :

● $PR(P) = (1-d) + d [(PR(P1)/C(P1) + \dots + (PR(Pm)/C(Pm))]$

- L'algorithmme du PageRank expliqué avec quatre pages liées et un facteur d'amortissement $d = 0.85$
- Le PR de la page D sera donc de 0.15, grâce au premier terme de la formule du PageRank $(1 - d)$
- Bien qu'ayant un PR calculé, il est vraisemblable que cette page disparaîtra de l'index Google très vite, n'ayant aucun lien entrant.
- Au bout d'une vingtaine d'itérations, les valeurs de PR pour nos pages convergent vers les valeurs suivantes :

- Page A	1.49
- Page B	0.78
- Page C	1.58
- Page D	0.15

