Mutation des métiers de l'information

Eric Debonne

https://www.linkedin.com/in/solaci/

Eric.Debonne@solaci.com

Usage en copie privée uniquement, diffusion et reproduction interdite

SOLACI Créée en 2003

- Expert
  - Veille / Intelligence Economique,
  - Moteur de Recherche d'Entreprise,
  - Data Intelligence,
  - Intelligence Collective
- www.solaci.com
- www.taligentia.com

Facilitateur, j'accompagne mes client en tant que responsable de projet, AMOA

Usage en copie privée uniquement, diffusion et reproduction interdite

Programme

- Le métier de Documentaliste 25 ans auparavant
- Les premières évolutions
- La recherche et l'accès à l'information
- La veille, le collaboratifs, l'intelligence des données
- Le métier actuel de professionnel de l'information
- Archiviste : archives départementales

Usage en copie privée uniquement, diffusion et reproduction interdite

Métiers de l'information

- Documentaliste
- Professionnel de l'information

Usage en copie privée uniquement, diffusion et reproduction interdite

Il y a 25 ans et plus - Missions

- Gérer l'information
- Documents papiers : livre, publications, presse, documents d'entreprise
- Gérer l'information électronique
- Classification, thésaurus
- Relation éditeurs, gestion des abonnements
- Gestion d'une bibliothèque – CDI

Usage en copie privée uniquement, diffusion et reproduction interdite

Il y a 25 ans et plus – Missions (2)

- Recherche de documents
- Production de synthèses
- Relations Interlocuteurs – secrétaire – documentaliste
- Pas d'intranet

Usage en copie privée uniquement, diffusion et reproduction interdite

Expertise

- Plan de classement, catégorisation, ajout de méta données
- Thésaurus
- Lire, comprendre l'information
  - Compétences métier, sans être expert
- Résumer, rédiger des synthèses
- Administrateur, Utilisateur de logiciels
  - Gestion électronique de Documents
  - Gestion de Bibliothèque

Usage en copie privée uniquement, diffusion et reproduction interdite

Expertise (2)

- Recherche d'information
  - Booléen AND / OR/ NOT / Proximité
  - Associé, générique, spécifique
  - Connaissance des revues et autres abonnements papier
- Recherche sur Internet (peu utilisé)
- Langues ?

Usage en copie privée uniquement, diffusion et reproduction interdite

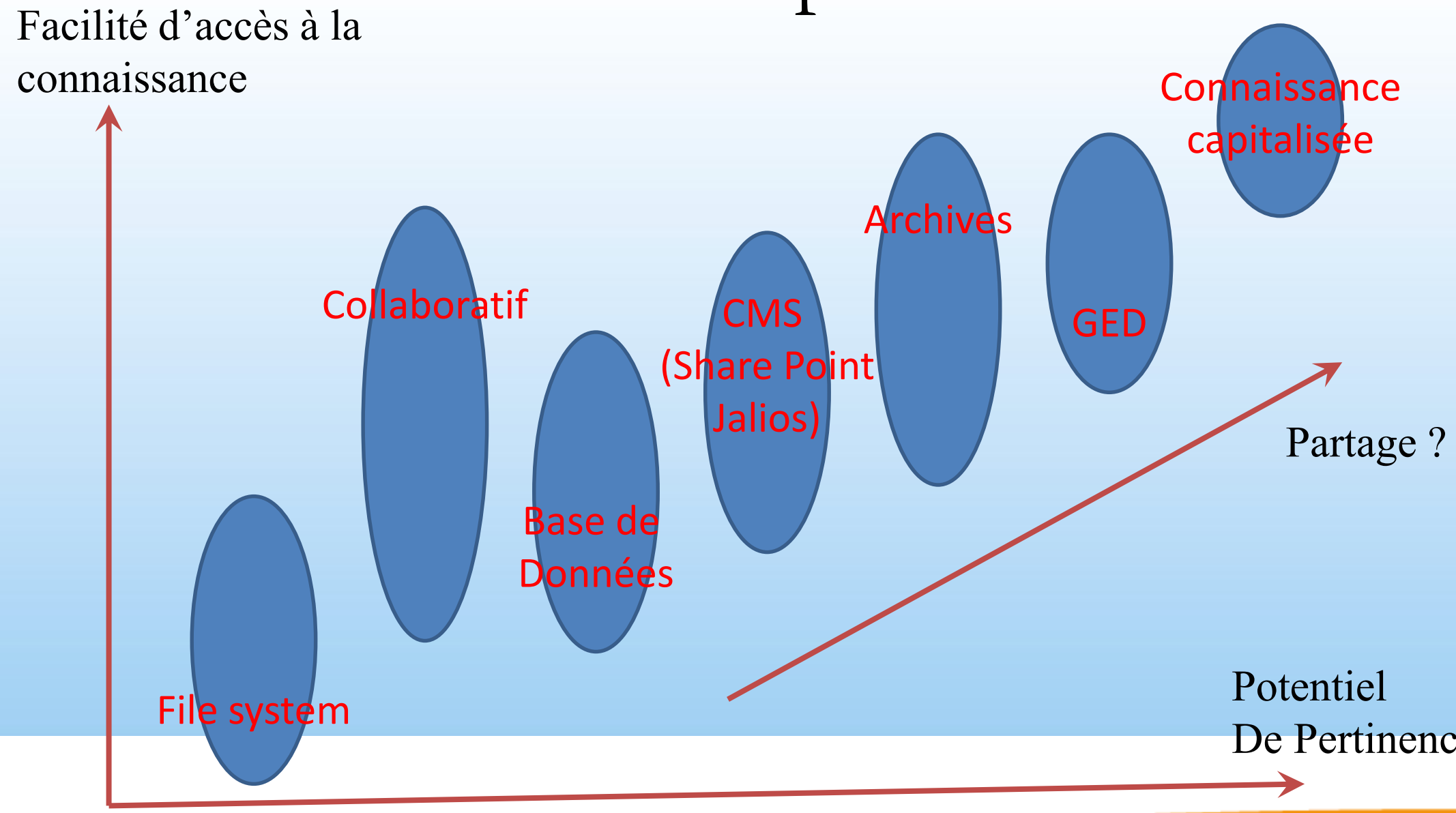
Premières évolutions

- Intranet
- Développement d'internet

Usage en copie privée uniquement, diffusion et reproduction interdite

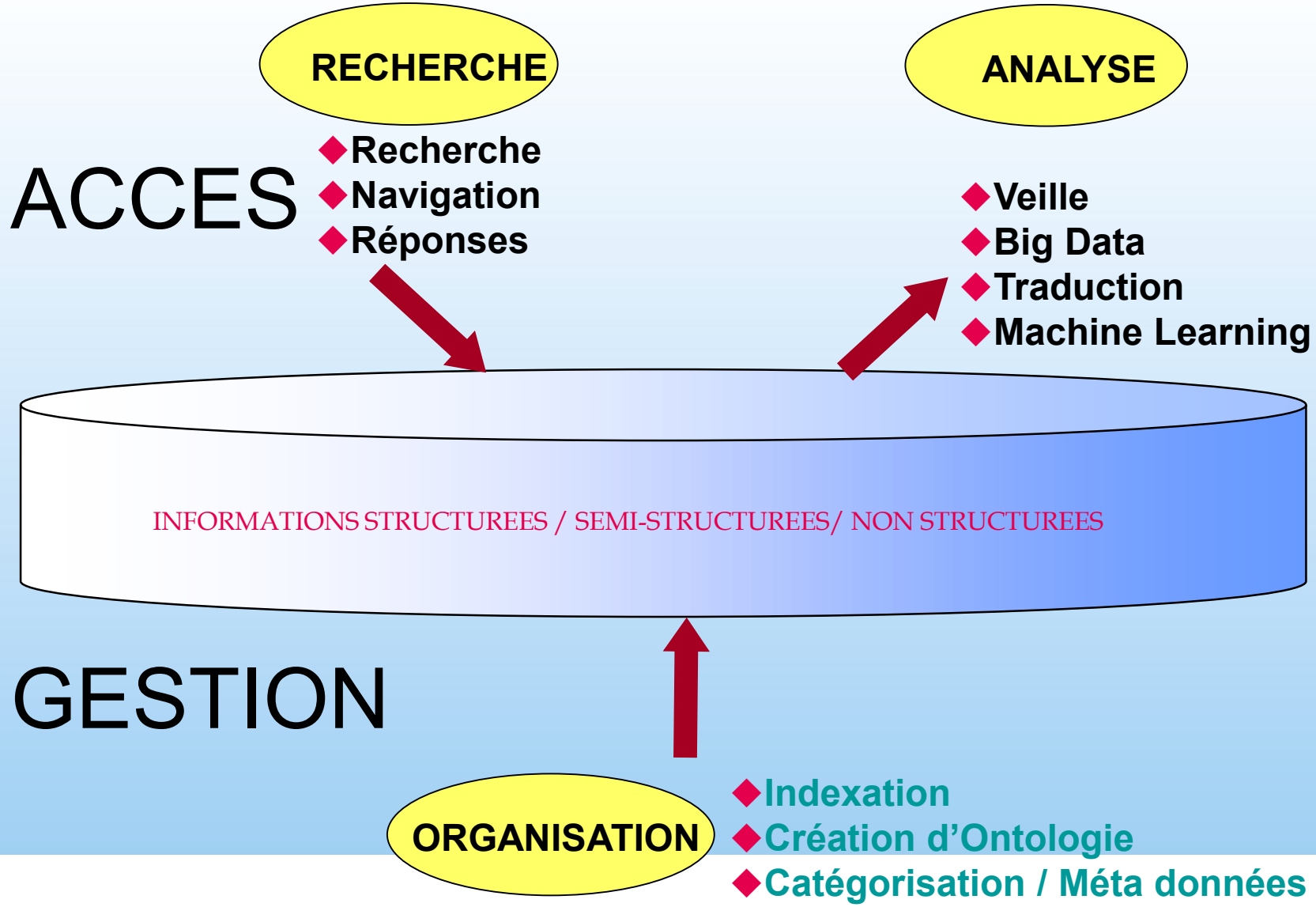
La recherche et l'accès à l'information

Trop d'information dans l'entreprise



Usage en copie privée uniquement, diffusion et reproduction interdite

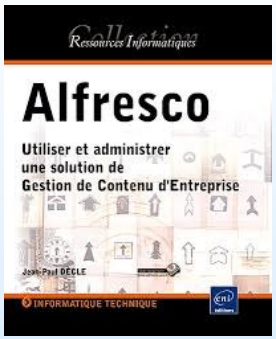
La fonctions d'accès à l'information



Usage en copie privée uniquement, diffusion et reproduction interdite

Documents

SharePoint



documentum

SAP

OPEN TEXT
The Content Experts™

LOCARCHIVES

Usage en copie privée uniquement, diffusion et reproduction interdite

Etapes

- Accéder aux documents
  - Localisation
  - Logiciels
- Lire les différents formats
  - Office, pdf, etc…
  - Application de filtres de conversion
- Structure de l'information
- Méta données
  - Titre, Auteur, Organisation
- Gérer la sécurité

Usage en copie privée uniquement, diffusion et reproduction interdite

Contenu, Texte, Ambiguïté

Pour Patrick Viveret, essayiste engagé, ancien conseiller référendaire à la Cour des comptes et signataire de l'Appel, *«nous avons besoin d'une politique fondée sur l'intelligence collaborative et la participation active des citoyens, bref ce que nous appelons "une mutation qualitative de la démocratie"»*.

Usage en copie privée uniquement, diffusion et reproduction interdite

De la clé d'accès au mot-clé

- Objectif visé : indexation par les mots du texte
- Difficultés
  - découpage des mots dans le texte
    - ambiguïté de la ponctuation
    - non-reconnaissance des mots composés
  - identification des différentes formes du même mot
    - problème de la casse (minuscules / majuscules)
    - formes au singulier et au pluriel, formes conjuguées

Usage en copie privée uniquement, diffusion et reproduction interdite

Racinisation ou Stemmatisation

- Réduction algorithmique d'un mot à sa "racine" par traitement des terminaisons
- Recherche sur tous les mots de même racine
- Pas toujours exacte, ni complète
  - absolut absolute absoluted absolutely absolutelys absoluteness absolutes absolution absolutions absolutism absolutization absolutize absolutized absolutizing absolutly absolutist absolutists
  - absolve absolved absolves absolving
- Fonctionne mieux pour certaines langues (anglais…) que pour d'autres (français …)

Usage en copie privée uniquement, diffusion et reproduction interdite

Limites

- Bruit
  - organisme / organiste
  - schisme / schiste
  - retraite / retraitement
  - marin / mariner
- Silence
  - école / scolaire
  - hanche / coxal
  - œil / yeux / oculaire / ophtalmique

Usage en copie privée uniquement, diffusion et reproduction interdite

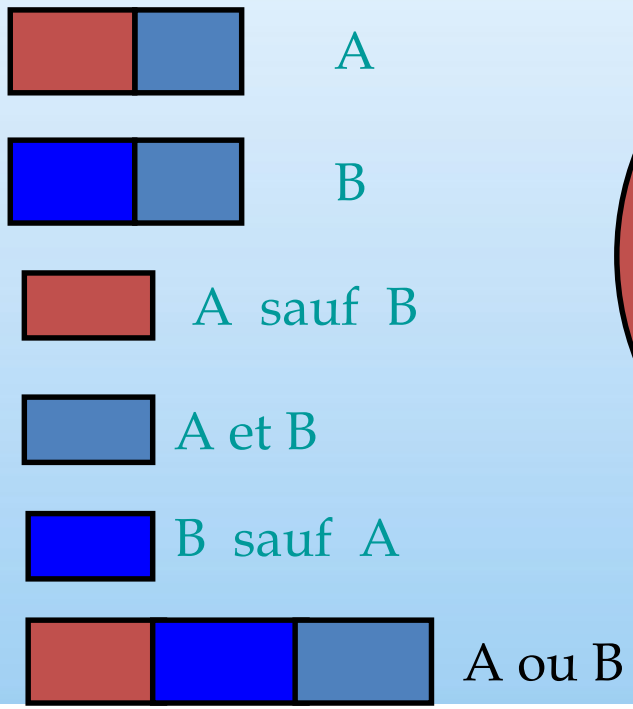
Anti-dictionnaires

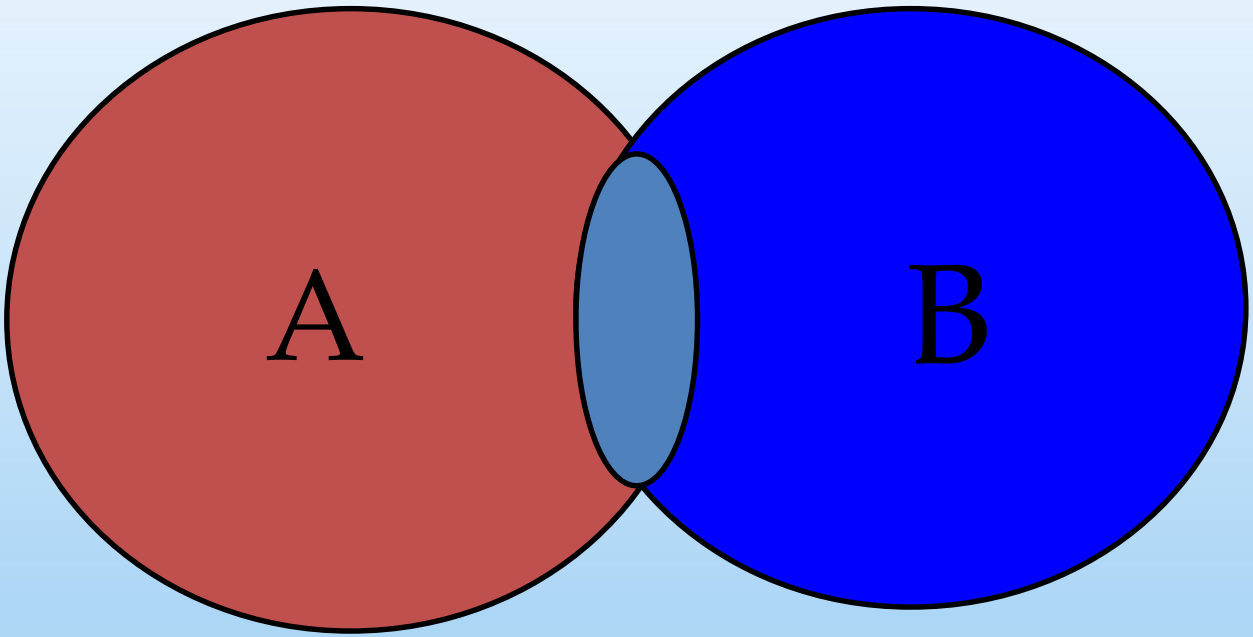
- Listes de mots non considérés comme clés d'accès (mots vides)
  - listes a priori ou mots trop fréquents dans une collection
  - exclusion radicale
    - vitamine A, sous marin, 100 % ne peuvent pas être recherchés
  - exclusion conditionnelle
    - A : ignored
    - vitamine A : OK

Usage en copie privée uniquement, diffusion et reproduction interdite

Approches booléennes

- 3 opérations ensemblistes de base
  - intersection (ET)
  - union (OU)
  - différence (SAUF)





Usage en copie privée uniquement, diffusion et reproduction interdite

Limites

- Pas de contrôle sur la taille des résultats obtenus (rien ou trop)
- Pas de classement des résultats
- Tous les termes sont d'une importance équivalente
- Résultats non intuitifs
  - en réponse à une requête en "OU"
    (A OU B OU C OU D …)
    - un document ne contenant qu'un seul des termes est estimé aussi pertinent qu'un document les contenant tous
  - en réponse à une requête en "ET"
    (A ET B ET C ET D…)
    - un document qui contient tous ces termes sauf un est écarté au même titre que ceux qui ne contiennent aucun de ces termes

Usage en copie privée uniquement, diffusion et reproduction interdite

Solution

- Utilisation d'un opérateur ET/OU
  - OU pondéré, interprétation non restrictive de l'opérateur ET
  - supposant aux documents une pertinence d’autant meilleure qu’ils répondent à davantage de critères de recherche

Usage en copie privée uniquement, diffusion et reproduction interdite

Pondération

- Possibilité d'affecter une pondération aux critères de recherche
- Prise en compte du nombre d'occurrences dans les documents
  - -> Possibilité d'ordonner les résultats

Usage en copie privée uniquement, diffusion et reproduction interdite

Tous les termes sont aussi importants ?

- **Requête utilisateur : Compte jeune**
- **Compte : présence à 40%**
- **Jeune : présence à 5%**

→ **Requête : Compte (poids 10) Jeune (poids 80)**

Usage en copie privée uniquement, diffusion et reproduction interdite

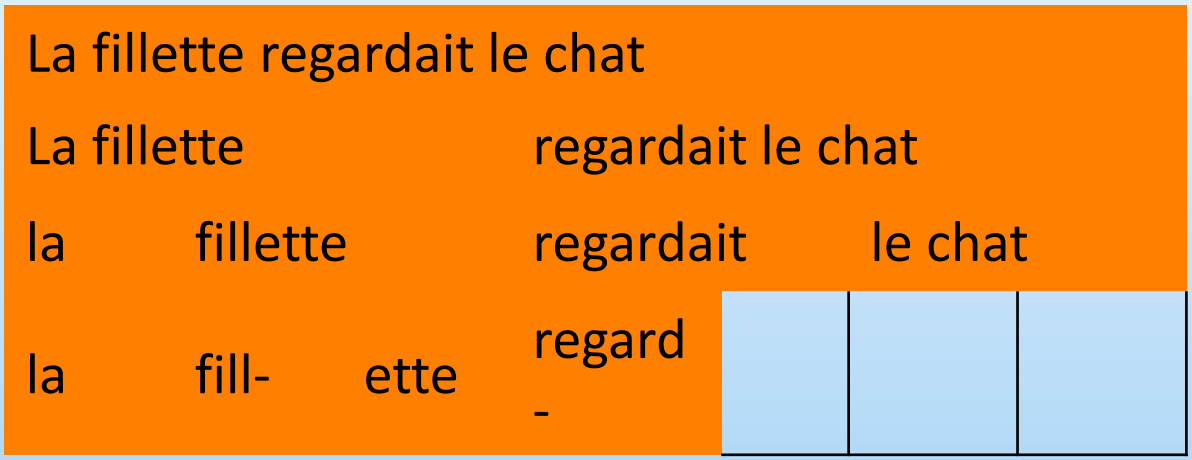
Adjacence

- Introduction d'opérateurs d'adjacence et de proximité
  - **mots contigus**
    - risque de silence
    - “validation rétroactive” ne permet pas de trouver
    - “la portée rétroactive de la validation […] ”
  - **mots voisins**
    - risque de bruit si le voisinage est trop “généreux”
    - permanents ~ syndicaux →
    - “des emplois permanents, ou à durée déterminée
    - […] représentants syndicaux au comité central d'entreprise”
  - **mots voisins à distance N**

Usage en copie privée uniquement, diffusion et reproduction interdite

Solutions Linguistiques

- Requête en langage naturel
- Comprendre le sens du document et le sens des requêtes
- Analyse linguistique
  - Morphologique
  - Syntaxique
  - Sémantique



Usage en copie privée uniquement, diffusion et reproduction interdite

Méta données - Extraction

- Utiliser, relier des méta données
  - Auteurs, départements
  - Concepts, etc…
- Extraire des entités nommées et autres informations
  - Auteur, organisation, lieu
  - Concepts, type d'information, actions (achats, hausse, baisse, innovations)
- Ajouter des méta données à chaque document
  - Enrichir, gagner en pertinence et en facilité d'accès

Usage en copie privée uniquement, diffusion et reproduction interdite

Co-occurence

- Statistique
  - Expressions les plus souvent présentes dans un corpus
- Relier des termes entre eux
  - Expressions souvent présentes
  - Expression souvent présentes ensembles
- Fréquence inverse : signaux faibles
- Apport d'algorithmes linguistiques

Usage en copie privée uniquement, diffusion et reproduction interdite

Catégorisation

- Prédéfinir des catégories
  - Un plan de classement
  - Une ontologie
  - La maintenir
- Catégorisation automatique
  - Définition automatique des expressions correspondantes à la catégorie
    - Extraction, statistiques, linguistique
  - Paramétrage de ces expressions et de leurs poids
    - Erreurs d'interprétation des statistiques pour la caractérisation d'une catégorie
  - Corpus d'apprentissage
  - Probabilités
  - Plusieurs catégories
  - Concept de non catégorisation

Usage en copie privée uniquement, diffusion et reproduction interdite

Clustérisation

- Découvrir la structuration d'un contenu
  - Peut aider à la constitution d'une ontologie
- Processus automatique
  - Extraction des expressions les plus courantes
  - Statistiques sur les expressions nommées ensemble, de manière proches
  - Paramétrage par l'ajout d'une liste de mots vides
    - Suppression d'expressions qui ne caractérise pas un contenu
  - Seuil de statistique
- Moteur de recherche : contenu pris en compte
  - La liste de résultat
  - Le corpus indexé
  - Influence des termes de la requête

Usage en copie privée uniquement, diffusion et reproduction interdite

Search & Big Data

- Volume de données, plusieurs Tera
- Architecture technique utilisant des serveurs « standard »
- Données versées par les outils de production
  - Problématique de la charge de sollicitation
- Recherche dans les données
  - Index du moteur de recherche, lui-même en architecture Big Data
- Machine Learning, Deep Learning

Usage en copie privée uniquement, diffusion et reproduction interdite

Un projet d'accès à l'information

- Qualité des données, des sources
- Questions de l'usage (expérience utilisateur)
  - Niveaux d'attentes, de pertinence, de transversabilité
- Niveau d'expertise des données et des utilisateurs
- Confidentialité
- Quelle restitution ?

Usage en copie privée uniquement, diffusion et reproduction interdite

Les interlocuteurs du projet

Le client	DSI	Professionnel de l'information
Interlocuteur Métier	Chef de projet MOE Mise en Œuvre	AMOA Chef de projet Consultant
Usager / représentant des utilisateurs	Spécifications techniques	
Exprime un besoin : «a Pain »	Développement / mise en œuvre de l'application Documentation	Propose des outils Accompagne le client pour défendre son projet
Participe aux réunions : description des besoins Cahier des charges	Support recette	Compréhension des besoins Formalisation des usages Traduction en CDC / Spécifications fonctionnelles
Recette : usages	Support / TMA	Gère la recette Cahier de recette Recette projet

Usage en copie privée uniquement, diffusion et reproduction interdite

L'expertise du professionnel de l'information

- Compréhension d'un usage en terme de
  - Flux d'information
  - Analyse d'information
  - Restitution pour l'utilisateur
- Connaissance des logiciels
- Connaissance du fonctionnements
  - Algorithmes
  - Fonctions
  - Limites

Usage en copie privée uniquement, diffusion et reproduction interdite