

Methode du TALN, Traitement Automatisé du Langage Naturel, **notion de l'indexation automatique**

Plan :

Principaux modes de recherche de l'information

- Par navigation arborescente
- Par navigation hypertextuelle
- Par requête sur les metadonnées du document
- Par requête sur le texte intégral

Indexation :

- Documentaire : langage documentaire
- Linguistique : TALN

Pièges du Langage Naturel :

- L'implicite
- La redondance
- L'ambiguïté

Indexation automatisée

- Analyse linguistique
 - Indexation morphologique
 - Indexation lexicale
 - Indexation syntaxique
 - Indexation par analyse sémantique
- Analyse statistique
 - Indice de pondération

**NOTIONS SUR
L'INDEXATION AUTOMATISEE
DU LANGAGE NATUREL**

Rappel des quatre principaux modes de recherche de l'information

Retour sur l'indexation

Quelques pièges du langage naturel

Notions sur l'indexation automatisée

LES QUATRE PRINCIPAUX MODES DE RECHERCHE DE L'INFORMATION

<i>Modes de recherche</i>	<i>Principe, démarche intellectuelles</i>	<i>Type d'information concernée</i>	<i>Exemples d'outils</i>
Recherche par navigation arborescente	<p>Arbre</p> <p>Démarche systématique, du général au particulier</p> <p>Recherche par menus successifs</p>	<p>Information structurée, organisée en plan de classement</p> <p>Information secondaire</p>	<p>Tables des matières</p> <p>Classifications documentaires (CDU, Dewey)</p> <p>Annuaire thématiques (Yahoo, Nomade...)</p> <p>Page d'accueil d'un site web</p>
Recherche par navigation hypertextuelle	<p>Réseau</p> <p>Démarche associative, d'une notion à l'autre.</p> <p>Navigation dans un réseau de noeuds et de liens</p>	<p>Information non structurée</p> <p>Information primaire, brute</p>	<p>Renvois dans une encyclopédie</p> <p>Hypertextes sur CD-ROM</p> <p>Sites web</p>
Recherche par requête sur les "métadonnées"	<p>Index</p> <p>Démarche</p>	<p>Information structurée en champs.</p>	<p>Index des livres</p>

du document	d'indexation de l'information Recherche par champs, logique booléenne	Information secondaire	Banques de données Catalogues de bibliothèques
Recherche par requête sur le texte intégral	Texte Démarche d'analyse linguistique Recherche contextuelle sur le contenu	Information non structurée ; Information brute, primaire	Outils de TALN (<i>Traitement Automatique du Langage Naturel</i>) Moteurs de recherche Outils linguistiques

Représentation très simplifiée et schématisée des méthodes de recherche d'information, regroupées ici autour de principes très généraux ;
Sur le web, à l'intérieur du quatrième mode de recherche (la recherche par requête sur le texte intégral), existe une très grande diversité de travaux, de nombreuses technologies et de nombreux outils.

Coexistence de deux grandes écoles :

- **les moteurs de recherche :**
focalisés sur deux problèmes :
 - la **quantité d'informations à traiter** (cf la taille du web et des index des moteurs)
 - la **capacité à traiter simultanément des milliers de requêtes**
- **les technologies du "text mining" :**
focalisées sur deux autres problèmes :
 - le **traitement d'une grande pluralité de sources d'informations**
 - les **méthodes automatiques de classification de l'information** : analyse linguistique, classement automatique...
 - outils concernés : agents intelligents, outils de cartographie de l'information, certains moteurs (Exalead) et métamoteurs (MapStan, Kartoo...)

Retour sur l'indexation

Opération d'indexation au cœur des activités documentaires, du traitement des documents et au fondement de la recherche de l'information : recherche de documents fondée sur l'appariement entre :

- les termes d'une requête
- les termes décrivant le contenu du document

Deux approches de l'indexation : documentaire et linguistique

- Selon la définition classique, documentaire :
Indexation : Représentation par les éléments d'un langage documentaire des notions résultant de l'analyse d'un document ou d'une question en vue d'en faciliter la recherche.

Indexation vise à représenter le contenu des documents : thèmes, sujets, aspects, etc., selon un langage

artificiel partagé ; indexation comme opération de traduction des concepts d'un document, dans une sorte d'interlangue documentaire.

Selon l'approche documentaire : **indexation = représentation extérieure, forcément réductrice du contenu :**

- Approche linguistique de l'indexation :
représentation du document par le document lui-même, qui est la représentation la plus fidèle.
Evolution actuelle : essor des outils de TALN (Traitement Automatique du Langage Naturel)

A quoi sert l'indexation ?

- A la recherche documentaire :
Selon l'approche documentaire : « L'indexation a pour but de faciliter l'accès au contenu d'un document ou d'un ensemble de documents à partir d'un sujet ou d'une combinaison de sujets » (Pomart et Sutter, 1997)
- Objectif de l'indexation documentaire :
permettre la recherche du document dans une collection organisée, une banque de données...
- A la recherche d'informations :
- Objectif de l'indexation automatisée :
permettre la recherche de l'information dans le texte intégral des documents

- **A l'analyse et à la cartographie de l'information :**
Avec les logiciels de TALN et les nouveaux outils linguistiques, l'indexation sert à :
 - analyser des corpus de textes
 - dresser des cartographies informationnelles : réseaux sémantiques, analyses de sites web
 - la veille stratégique et l'intelligence économique
 - la scientométrie : analyse des traces et productions de la science

Quelques pièges du langage naturel

Principal défi de la recherche d'information : les pièges et difficultés du langage naturel

Tableau récapitulatif
(d'après P. Lefèvre)

Caractéristiques du langage naturel	Les difficultés dans la recherche d'informations	Définitions	Exemples
1/ L'implicite	<ul style="list-style-type: none"> • La pragmatique : Impossible à prendre en compte par des logiciels ou des langages documentaires	Liée au contexte du message, aux connaissances sur le monde, à l'usage... <ul style="list-style-type: none"> • domaine de la pragmatique : étude du " langage en action " 	" Paul donna le billet à la jeune femme " : <ul style="list-style-type: none"> • transaction commerciale : billet de banque ? • spectacle : billet d'entrée ? • relation amoureuse : billet doux ? • espionnage : message chiffré ?
2/ La redondance	<ul style="list-style-type: none"> • La synonymie : 	Mots ou expressions différents ayant le même sens, ou des sens voisins.	voiture et automobile ; tremblement de terre et séisme ; train et chemin de fer...

	<ul style="list-style-type: none"> • La paraphrase : 	Expressions équivalentes mais de structure ou de termes différents	<i>Mon fils a cessé de fumer</i> <i>Jean a renoncé au tabac</i>
	<ul style="list-style-type: none"> • Le glissement de sens : 	<p>La dénotation : sens propre d'un mot</p> <p>La connotation : sens d'un mot dans un contexte particulier</p>	<i>Il prend un bain</i> <i>Il est dans le bain</i>
3/ L'ambiguïté	<ul style="list-style-type: none"> • L'homonymie : 	Mots ayant la même forme, la même graphie mais des sens différents.	<i>Je porte la porte</i> <i>Les poules du couvent couvent</i>
	<ul style="list-style-type: none"> • La polysémie : 	Mots ou expressions ayant plusieurs sens	Mémoire humaine, mémoire d'ordinateur, le mémoire de maîtrise...
	<ul style="list-style-type: none"> • L'homotaxie : <p>> problèmes pour les logiciels de TALN</p>	Une même syntaxe recouvrant des réalités différentes	<i>Jean est facile à convaincre</i> <i>Jean est habile à convaincre</i>

Notions sur l'indexation automatisée

1/ L'analyse linguistique

2/ L'analyse statistique

L'indexation automatisée (dans de nombreux moteurs de recherche) repose sur les techniques de **TALN** :

Traitement Automatique du Langage Naturel.

L'indexation automatisée repose sur la notion de fichier inverse :

Fichier inverse :

fichier organisé par ordre alphabétique de **descripteurs**, **de mots-clés ou de mots**, derrière lesquels figurent les numéros des notices possédant ces termes. Ce fichier est " inversé " par rapport au " fichier direct " (ou principal.) Il est lu en accès direct sur les mots-clés de la question.

Dans les systèmes d'index en texte intégral, les fichiers inverses sont des fichiers contenant les mots du texte, classés alphabétiquement, avec l'adresse précise de leur occurrence dans le texte.

Dans les fichiers inverses des moteurs de recherche, chaque terme pointe vers les URL des pages qui contiennent le terme.

Différence essentielle entre **l'indexation automatisée** et **l'indexation documentaire manuelle** :

- l'indexation documentaire manuelle (avec un langage documentaire) porte sur les **concepts**, représentés par des mots-clés ou des descripteurs

- l'indexation automatisée porte sur les **mots** des documents

- Exemple : pour une requête portant sur la notion suivante : "cours d'informatique", deux possibilités, selon le mode d'indexation utilisé par le système d'information :
 - **en indexation manuelle** (avec langage documentaire) : si l'indexation propose "*enseignement de l'informatique*", nécessité d'utiliser ce terme dans la requête
 - **en indexation automatisée** : l'indexation plein texte devra prendre en compte tous les termes et expressions suivants : *cours d'informatique, enseignement de l'informatique, TP d'informatique, on enseigne l'informatique dans cette université*, etc., en réponse à une requête *cours d'informatique*

Deux grandes **méthodes d'analyse** dans **l'indexation automatisée** :

- **analyse linguistique** : fondée sur la **reconnaissance des mots**

- **analyse statistique** : fondée sur **la fréquence des mots**

Ces deux méthodes sont généralement utilisées conjointement dans la plupart des logiciels, mais reposent sur des principes différents.

1/ L'analyse linguistique et les différents niveaux d'indexation

A/ Indexation morphologique : indexation en texte intégral ou plein texte (*full text*)

B/ Indexation lexicale : la lemmatisation ou la normalisation

C/ Indexation syntaxique

D/ Indexation par analyse sémantique

Au moins **quatre grands niveaux d'analyse linguistique** du texte intégral :

- **niveau morphologique** : reconnaissance du mot
- **niveau lexical** : réduction du mot à sa forme canonique > lemmatisation
- **niveau syntaxique** : niveau d'utilisation de la grammaire
- **niveau sémantique** : niveau de la reconnaissance des concepts

Cinquième niveau d'analyse du langage naturel : **niveau pragmatique** :

- niveau de la connaissance du monde réel :

> **impossible à automatiser et hors de portée des outils informatiques**

- compréhension d'un mot ou d'une phrase dans son contexte d'énonciation

A/ Indexation morphologique : indexation en texte intégral ou plein texte (*full text*)

- Première étape de traitement du langage naturel : l'analyse morphologique est le préalable à toute indexation automatisée (linguistique et statistique)
- Peut parfois constituer le seul niveau d'indexation

Fondée sur l'analyse morphologique des mots : leur **forme**

Méthode constituée de deux niveaux d'indexation :

- **Niveau 0 : Indexation libre par fichier inverse brut.**

- **Niveau 1 : Indexation libre par fichier inverse de mots significatifs**

- **Niveau 0 : Indexation libre par fichier inverse brut.**

Procédé : constitution d'un lexique de tous les mots du texte, par découpage du texte. Classement de tous les mots dans un fichier inverse, avec l'adresse de chaque mot dans le document. Les mots-vides ne sont pas éliminés.

Exemple du texte suivant :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi

donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes. "

Indexation libre par fichier inverse brut :

a	donné	légué	paysages
à	elle	lentement	paysans
anonymes	façonnés	lui	ruraux,
aussi	générations	monuments	ses
Bretagne.	histoire	n'	seulement
de	l'	par	
des	la	pas	

Applications :

Au début, la plupart des moteurs de recherche fonctionnaient selon ce premier niveau d'indexation, en variant simplement la typographie majuscules-minuscules et en supprimant les accents (niveau 0+).

Inconvénients :

- taille volumineuse des fichiers inverses
- beaucoup de **bruit**

- **Niveau 1 : Indexation libre par fichier inverse de mots significatifs**

Méthode fondée sur :

- **l'élimination des mots-vides** (articles, prépositions, mots grammaticaux...), à partir d'un dictionnaire de termes (appelé "**stop list**")
- la **constitution d'un index des termes non éliminés**, considérés comme des chaînes de caractères.
Indexation libre, car les index retenus ne sont comparés à aucune liste préalable.

Exemple du texte :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes. "

anonymes	légué
Bretagne	lentement
donné	monuments
façonnés	paysages
générations	paysans
histoire	ruraux

Applications :

- logiciels de gestion documentaire
- logiciels de GEIDE
- moteurs de recherche sur le texte intégral

Exemples : Bull Searchway, OpenText Livelink, Basis, Oracle Intermedia Text...

- Recherche se fait selon logique booléenne ou avec des opérateurs de proximité ;
- méthode de recherche en full text quotidiennement utilisée dans les banques de données, pour les recherches portant sur les chaînes de caractères, dans les résumés par exemple.

Exemple : dans la phrase "*Prolétaires de tous les pays : unissez-vous*", l'indexation en full text éliminera les termes : de, tous, les, vous, et gardera "prolétaires", "pays" et "unissez"

A la recherche, il suffira de taper l'un de ces termes, ou une combinaison des termes, pour retrouver la phrase.

• **Quels problèmes et quels défauts de l'indexation morphologique (niveaux 0 et 1) ?**

Inconvénients évidents de cette méthode très fruste d'indexation :

- tous les mots non vides mis sur le même plan :
- pas de prise en compte de l'ordre des mots
- apparition des différentes formes d'un mot : par ex. un verbe va apparaître plusieurs fois sous des formes différentes
- l'analyse porte seulement sur des mots isolés (des unitermes), et délaisse toutes les expressions (les syntagmes), souvent porteurs de sens :

- ex. : pomme de terre donnera deux mots "pomme" et "terre", analysés séparément

- polysémie, synonymie du langage naturel pas prise en compte :

- *vol* = aussi bien *vol d'avion* que *vol à la tire*

- stricte équivalence entre termes de recherche et termes indexés : pas de faute de frappe...

>> limites très sérieuses de l'analyse morphologique, qui peut générer beaucoup de «bruit ou de silence documentaire »

> A noter : indexation morphologique est encore le seul niveau d'indexation utilisé par de nombreux moteurs de recherche.

B/ Indexation lexicale : la lemmatisation ou la normalisation

Réduction des mots à leur forme canonique, à leur racine : toutes les formes d'un verbe par exemple sont regroupées à l'infinitif ; tous les mots au pluriel sont ramenés au singulier...

- L'analyse lexicale consiste à ramener les mots à une forme de base, et à reconnaître toutes les variations liées à cette forme. Trois types de formes de base :
 - le radical : mang(e) pour manger, mangeoire, mangeables...
 - la racine : nation pour nationalité
 - le lemme : infinitif des verbes, masculin singulier...

Lemmatisation permet de diminuer fortement le nombre de mots analysés, en **éliminant toutes les flexions et les dérivations grammaticales.**

Objectif : **ramener chaque terme à une forme unique**
- allègement des index, diminution du bruit à la recherche

Pas ou peu d'indexation lexicale sur Internet

Exemple du texte :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes. "

<i>anonyme</i>	<i>léguer</i>
Bretagne	lentement
<i>donner</i>	<i>monument</i>
<i>façonner</i>	<i>paysage</i>
<i>génération</i>	<i>paysan</i>
histoire	<i>rural</i>

C/ Indexation syntaxique

Passage **de la forme à la syntaxe.**

Analyse syntaxique d'un texte, par un logiciel d'indexation automatique, va permettre **plusieurs choses** :

- **identification des groupes nominaux, des expressions** : "accident du travail", pomme de terre", seront indexées comme expressions, et non mot par mot
- analyse syntaxique concerne la **place des mots dans une phrase**
- **reconnaissance des expressions contiguës** ou disjointes : par exemple, pouvoir reconnaître dans l'expression : Agence Française de presse l'expression Agence de presse
- l'élimination des problèmes d'homographie : termes ayant la même orthographe mais de sens différent : différence entre le substantif "porte" et la forme du verbe "porte"

Deux niveaux d'indexation syntaxique :

- **Indexation libre par fichier inverse de syntagmes ou mots composés**

- **Indexation libre par syntagmes nominaux étendus**

- **Indexation libre par fichier inverse de syntagmes ou mots composés**

Syntagme : unité syntaxique élémentaire (groupe nominal, groupe verbal)

Extraction de groupes de mots ou de mots composés (groupes nominaux) présents dans le texte. Possibilité d'ajouter d'autres mots significatifs (adjectifs, verbes), pour créer un syntagme.

Indexation libre.

Exemple du texte :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes. "

anonyme	léguer
Bretagne	monument
donner	paysage rural
façonner	paysan
génération	
histoire	

- niveau d'indexation présent dans certains logiciels de TALN
- sur Internet, le moteur **Exalead** dispose de fonctions d'analyse des groupes nominaux

- **Indexation libre par syntagmes nominaux étendus**

Traitement syntaxique plus poussé, fondé sur l'extrapolation, la transformation ou la dérivation d'expressions ou de mots composés. Création de syntagmes non présents dans le texte.

Exemple :

- légué ses monuments = legs de monuments
- paysages façonnés = façonnage (ou création) de paysage

Anonyme	<i>Legs de monument</i>
Bretagne	monument
Donner	paysan
<i>Façonnage de paysage</i>	
Génération	
histoire	

- niveau d'indexation assez rare dans les logiciels de TALN

D/ Indexation par analyse sémantique

Analyse sémantique, fondée sur le **sens des mots (les concepts)**.

Analyse sémantique va s'intéresser au regroupement de termes synonymes, aux familles de termes, pour **dresser un réseau des relations sémantiques dans un texte**.

Systemes relevant des systèmes experts, **qui intègrent un thesaurus dans l'indexation automatique** des textes.

2/ L'analyse statistique

- **Principes :**

Indexation fondée sur la fréquence des mots :

- Indexation déjà ancienne, fondée sur le **calcul statistique des occurrences**, cad de la fréquence d'apparition de mots dans un texte. Tous les mots significatifs d'un texte sont relevés (les occurrences) et leur fréquence est calculée, selon un indice moyen de fréquence (par exemple 1 / 1000).

- Méthode fondée sur le postulat que, **si l'indice de fréquence d'un mot est supérieur à son indice moyen, il doit s'agir d'un mot-clé pertinent, décrivant bien le sujet du texte**.

- Méthode permet **les calculs de pondération**, cad l'importance d'un mot dans un document déterminé et l'élimination de termes moins significatifs.

Applications :

- moteurs de recherche : classement des résultats selon l'indice de pertinence
- nombreux logiciels de TALN
- outils de scientométrie, fondés sur l'analyse des mots-associés :
Sampler, WordMapper...

Sources :

- Cours URFIST Bretagne-Pays de Loire. Alexandre Serres
- Philippe LEFEVRE. *La Recherche d'informations*, Hermès, 2000
- Pascale SEBILLOT, Traitement automatique des langues et recherche d'information. In *La Recherche d'information sur les réseaux, Cours INRIA, 2002.* , p. 137-168