Module RI - Chapitre 2 : Représentation et Indexation de l'information

1

1. Introduction

La représentation ou l'indexation de l'information est un processus permettant de construire un ensemble d'éléments « clés » permettant de :

- ✓ Bien caractériser le contenu d'un document
- ✓ Réduire la taille d'un document
- ✓ Faciliter le processus de recherche
- ✓ Réduire le temps de recherche.

1.1. Les éléments clés

Pour l'information textuelle les éléments clés peuvent être :

- ➢ mots simples : exemple : pomme
- ➢ groupe de mots : exemple : pomme de terre

Pour une image les éléments clés peuvent être :

- ➢ Couleurs, formes

2

2. L'indexation

L'indexation peut se faire en utilisant :

- **Un vocabulaire non contrôlé :**
  - Indexation plein-texte : des mots clés sont extraits du contenu
- **Un vocabulaire contrôlé :**
  - Utilisation d'un thésaurus
  - Utilisation d'une ontologies

3

2.1. Thésaurus

Liste de mots clés + relation sémantiques entre les mots clés.

Pour élaborer un thésaurus il faut :

- Déterminer les termes qui peuvent être pris en compte
- Associer des relations sémantiques entre ces termes :
  - ✓ Hyperonymie/hyponymie(généralisation/spécialisation) (is-a),
  - ✓ antonymie (opposé à)
  - ✓ … etc

2.2. Ontologie

- ✓ Liste de concepts (exemple : regroupement des termes pour certaine sémantique) + relations entre les concepts.

4

2.3. Avantage du vocabulaire contrôlé

- ✓ Permet la recherche par concepts (par sujets, par thèmes), plus intéressante que la recherche par mots simples
- ✓ Permet la classification (regroupement) de documents (par sujets, par thème)
- ✓ Fournit une terminologie standard pour indexer et rechercher les documents

2.4. Inconvénients du vocabulaire contrôlé

- ✓ Indexation très coûteuse (pour construire le vocabulaire, pour affecter les concepts (termes) aux documents.
- ✓ Difficile à maintenir (la terminologie évolue, plusieurs termes sont rajoutés tous les jours)
- ✓ Les utilisateurs ne connaissent pas forcément le vocabulaire utilisé.

5

2.5. Avantage du vocabulaire non contrôlé

- ✓ Indexation plus rapide
- ✓ Facile à maintenir sa mise à jour
- ✓ Les utilisateurs connaissent facilement le vocabulaire

2.6. Inconvénients du vocabulaire non contrôlé

- ✓ Indexation basée sur des statistiques, pas de sens
- ✓ Ne permet pas la recherche par concept, par sujet, par thème

6

Remarque :

Dans notre cours, on s'intéresse à **l'indexation plein-texte (vocabulaire non contrôlé)**.

Les autres formes d'indexations (utilisation d'un thésaurus ou d'une ontologie) sont utilisées pour la recherche d'information sémantique.

7

3. Le processus d'indexations

L'indexation peut être :

- Manuelle (par des experts en indexation)
- Automatique (par programmes sur un ordinateur)
- Semi-automatique (combinaison des deux)

3.1. Indexation manuelle

- ✓ Choix des mots effectué par les indexeurs (experts)
- ✓ Basée sur un vocabulaire contrôlé
- ✓ Approche utilisée souvent dans les bibliothèques, les centres de documentation
- ✓ Dépend du savoir faire de l'indexeur (expert)

8

3.2. Indexation Automatique

- Approche statistique par distribution des mots
- Approche statistique compréhension du texte (TALN)
- Approche courante, plutôt statistique avec des hypothèses simples
  - ✓ Redondance (fréquence) d'un mot marque son importance
  - ✓ Cooccurrence des mots marque des concepts sur le sujet d'un document

3.2. Indexation Manuelle VS Automatique

L'indexation manuelle nécessite des experts des domaines de des documents, elle est couteuse en temps et en main d'oeuvre. L'indexation automatique est plus rapide en temps, mais nécessite des approches efficaces. Dans ce cours, on s'intéresse à l'indexation automatique.

9

4. Processus d'indexation automatique

Pour indexer les documents de façon automatique, nous utilisons l'approche courante, qui est une approche basée sur des statistiques sur l'apparition des mots dans les documents.

Généralement un mot est une suite de caractères séparés par blanc, signe de ponctuation, caractères spéciaux,…). Donc, il faut d'abord définir la façon de localiser un mot. Puis, plusieurs questions se poses :

1. Est ce qu'on garde dans l'index tous les mots du document, ou bien seulement les mots représentatifs du document. ?
2. Les mots extraits seront utilisés comme ils sont, ou bien ils doivent subir certaines transformations ?
3. Les mots ont la même importance ou bien chaque mot a une certaine importance qu'il faut définir ?

10

Démarche à suivre dans l'indexation

Afin de faire une indexation qui permet une représentation réduite, représentative et facile à utiliser dans la phase de la recherche, il faut la faire en 3 étapes :

- Étape 1 : extraction de mots représentatifs
- Étape 2 : normalisation des mots extraits
- Étape 3: pondération des mots normalisés, pour mesurer leurs importances

11

4.1. Etape 1 : Extraction des mots à utiliser

Cette étape s'appelle aussi « tockenization ». Généralement un mot est une suite de caractères séparés par blanc, signe de ponctuation, caractères spéciaux,…). Donc :

1. Définir la façon de localiser un mot.
2. Extraire les mots
3. Définir les mots non utiles (liste de mots vides / stoplist / common words). *Exemple :*
   - *Anglais : the, or, a, you, I, us, …*
   - *Français : a, le , la, de , des, je, tu, …*

   ➢ *Attention à :*
   - *US : «USA » ; « give us information »*
   - *a de (vitamine a)*
   - *….. etc*

   Donc, il faut bien définir cette liste de mots vides !
4. Supprimer (ignorer) les mots vides, et garder que les mots représentatifs.

12

4.2. Etape 2 : Normalisation

Il existe plusieurs façons pour la normalisation :

1. Normalisation par Lemmatisation : (radicalisation) / (stemming)

C'est un processus morphologique permettant de regrouper les variantes d'un mot :

- Exemple : économie, économiquement, économiste, ⇨ économ
- Exemple en anglais : retrieve, retrieving, retrieval, retrieved, retrieves ⇨ retriev

Pré **traite** **ment**

Forme morphologique d'un mot

13

2. Normalisation par l'utilisation de règles de transformations

- ❑ Règles de type : condition action

  L'algorithme le plus connu est : Porter (utilisé pour l'anglais)

  Plusieurs implantations sont accessibles

  http://www.tartarus.org/~martin/PorterStemmer/

- ❑ Analyse grammaticale
  - Utilisation de lexique (dictionnaire)
  - Tree-tagger (gratuit sur le net)

- ❑ Troncature

14

Détail de l'algorithme de porter

Notations :

v : une voyelle.

y : une voyelle si elle est précédée par une consonne.

c : une consonne.

m : mesure combien de fois il y a de « vc » avant les lettres à transformer

*e : le préfixe se termine par la lettre e

v : le préfixe contient une voyelle

*d : le préfixe se termine par une consonne doublée

*o : le préfixe se termine par « cvc » où le second « c » n'est ni « w » ni « x » ni « y »

15

Etapes de l'algorithme de porter :

Étape 1	**a**	• *SSES* → *SS* • *IES* → *I* • *SS* → *SS* • *S* →	caresses → caress ponies → poni caress → caress cats → cat
	b	• (m>0) *EED* → *EE* • (*v*) *ED* → • (*v*) *ING* →	feed → feed, agreed → agree plastered → plaster, bled → bled motoring → motor, sing → sing
	c	• (*v*) *Y* → I	happy → happi, sky → sky
Étape 2		• (m>0) *ATIONAL* → *ATE* • (m>0) *TIONAL* → *TION* • (m>0) *ENCI* → *ENCE* • (m>0) *ANCI* → *ANCE* • …	relational → relate conditional → condition, rational → rational valenci → valence hesitansi → hesitance …
Étape 3		• (m>0) *ICATE* → *IC* • (m>0) *ATIVE* → • (m>0) *ALIZE* → *AL* • (m>0) *ICITI* → *IC* • …	triplicate → triplic formative → form formalize → formal electriciti → electric …
Étape 4		• (m>1) *AL* → • (m>1) *ANCE* → • (m>1) *ENCE* → • (m>1) *ER* → • …	revival → reviv allowance → allow inference → infer airliner → airlin …
Étape 5		• (m>1) *E* → • (m=1 and not *o) *E* → • (m>1 and *d and *L) → lettre non doublée	probate → probat, rate → rate cease → ceas controll → control, roll → roll

16

Exemple de normalisation par l' algorithme de Porter

- **Texte original :**

  marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

- **Texte : après porter + suppression mots vides + fréquences :**

  Market 4, strategi 1, carri 1, US 1, compani 1, agricultur 1, chemic 2, report 2, prediction 2, share 1, statistic 1, agrochem 1, pesticid 1, herbicid 1, fungicid 1, insecticid 1, fertiliz 1, sale 2, stimulat 1, demand 1, price 1, cut 1, volum 1

17

3. Normalisation par troncature

Consiste à Tronquer les mots à X caractères

- Tronquer plutôt les suffixes
- Exemple troncature à 7 caractères
  - économiquement : économi

Quelle est la valeur optimale de X ? : 7 caractères pour le Français

4.3. Etape 3 : Pondération des mots (mesurer leurs importance)

Pour mesurer l'importance d'un mot dans un document, il faut lui attribuer une valeur. La 1ère valeur qu'on peut exploiter pour le moment c'est la fréquence du mot dans le document.

18

Exemple par troncature :

- **Texte :** un système de recherche d 'informations (document) (SRI, base de données documentaires, recherche documentaire) permet d 'analyser, d 'indexer et de retrouver les documents pertinents répondant à un besoin d'un utilisateur en information.

- **Extraction des mots et suppression des mots vides :** système, recherche, informations, document, SRI, base, données, documentaires, recherche, documentaire, analyser, indexer, retrouver, documents, pertinents, répondant, besoin , utilisateur, information

- **Normalisation par trocature à 7 caractères et mettre tout en miniscule :** système, recherc, informa, documen, sri, base, données, documen, recherc, documen, analyse, indexer, retrouv, documen, pertine, reponda, besoin, utilisa, informa

- **Pondération des termes par fréquences :**
  système 1, **recherc 2**, **informa 2**, **documen 4**, sri 1, base 1, données 1, analyse 1, indexer 1, retrouv 1, pertine 1, reponda 1, besoin 1, utilisa 1

19

5. Inconvénients de la normalisation

- Les algorithmes de "Stemers" sont souvent difficiles à comprendre et à modifier

- Peut conduire à une normalisation "agressive" (perdre le sens du mot) :
  - **Exemple de normalisation par Porter :** « general » devient « gener »
  - **Exemple de normalisation par troncature** : « Internet » devient « Interne »

Note : Il existe des techniques (analyse de corpus) pour réduire ces effets négatifs.

20

6. La méthode des n-grammes

- Définition : un n-gram est une succession de n lettres.
- Généralement n = 1,2,3
- Utilisée pour le chinois
- Intéressant pour la radicalisation
- **Exemple** : retrieval
  - 1-gram : r, e, t, r, i, e, v, a, l
  - 2-gram : re, et, tr, ri, ie, ev, va, al
  - 3- gram : ret, etr, tri, rie, iev, eva, val

6.1. Comparer deux mots par n-grammes

Exemple : retrieve et retrieval par 3-gram

- A=retrieve :
  - ret, etr, tri, rie, iev, eve
- B=retrieval
  - ret, etr, tri, rie, iev, eva, val

Sim (A,B)=2*nb_comm/(nb_A +nb_B)

21

6. Résumer du processus d'indexation

Le processus d'indexation comporte trois étapes

1. Extraction des mots, élimination des mots vides et mettre les mots en minuscule **(étape obligatoire)**
2. Normalisation **(étape facultative)**
3. Pondération des mots **(étape obligatoire)**

22

7. Fichier inverse

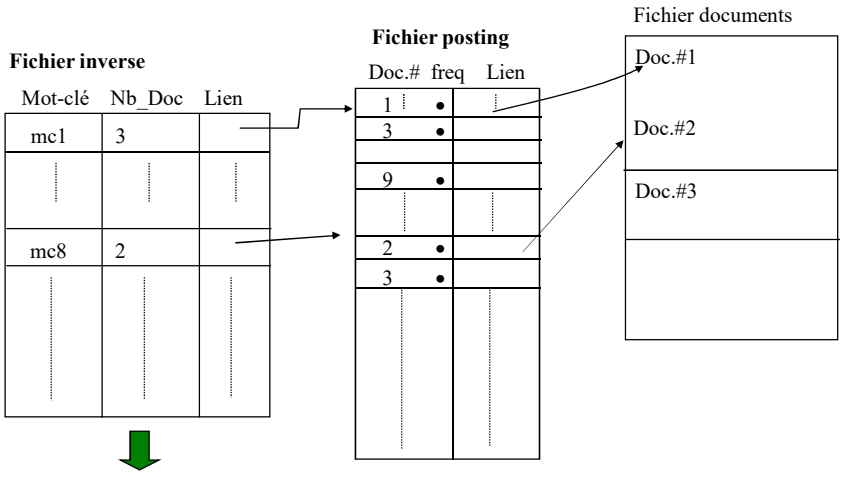
Une fois les documents indexés le résultat de chaque document aura donc un descripteur / une représentation

- Un descripteur :
  - ✓ Liste de mots
  - ✓ Fréquence de chaque mot
  - ✓ Exemple : système 1, recherc 1, informa 1, documen 3, sri 1, base 1, donnée 1, analyse1, indexer 1, retrouv 1, pertine 1, reponda 2, besoin 3, utilisa *1*

Ces mots sont ensuite stockés dans une structure appelée **fichier inverse,** pour pouvoir l'utiliser dans la phase de recherche (appariement document/requête)

23

Fichier inverse



- Liste triée
- B-Arbre
- Table de hashage (hash-code)
- ...

24

Exemple :

Pour une collection de 3 documents, on a les descripteurs suivants :

- D1 (sri 1 , recherche 1, information 2, document 3)
- D2 (document 1, information 4, vidéo 2)
- D3 (document 2, automatique 1)

La représentation théorique du fichier inverse de cette collection est :

	D1	D2	D3
sri	1	0	0
recherche	1	0	0
information	2	4	0
document	3	1	2
vidéo	0	2	0
automatique	0	0	1

25

Démarche de construction d'un fichier inverse

La construction d'un fichier inverse est une étape très importante en RI, elle peut prendre énormément de temps, mais ce temps n'a pas d'influence sur le processus de RI, car elle se fait avant la phase de recherche. Pour construire un fichier inverse il faut :

1. Faire toutes les étapes de l'indexation
2. Choisir la bonne structure à utiliser, qui permet le stockage de chaque mot, en le reliant avec son poids (fréquence pour l'instant) et son document.
3. Stocker chaque mot, avec son poids et son document.

26

Récapitulatif de la représentation de l'information en RI

- ✓ A l'issue de cette opération, chaque document sera représenté par une liste de termes pondérés. (sera détaillé dans le chapitre suivant)
- ✓ Le poids est fondamental et a une grande influence dans toutes les approches (modèles) de RI
- ✓ L'ensemble des termes extraits de tous les documents est stocké dans une structure spécifique appelée : fichier inverse
- ✓ Ce fichier permet de retrouver pour un terme donné tous les documents qui contiennent ce terme, avec son poids d'apparition.

27