Cours 5 : Recherche d'information et Analyse de données textuelles

NGUYỄN Thị Minh Huyền
Email: huyenntm@vnu.edu.vn
VNU University of Science, Hanoi

Plan

Plan

Plan

Plan

1. Introduction

2. Recherche d'information - introduction

3. Système de recherche d'information (SRI)
   - Indexation et mécanismes fondamentaux de RI
   - Modèles de RI
   - Evaluation des SRI

4. Méthodes d'analyse de données textuelles (*Text Mining*)
   - Applications

Références

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- J. Han and M. Kamber. `https://hanj.cs.illinois.edu/bk3/`
- Charu Aggarwal, Data Mining, Springer 2015.
- Tan, Steinbach, Karpatne and Kumar, Introduction to Data Mining, 2nd ed. `https://www-users.cs.umn.edu/~kumar001/dmbook/index.php`
- Thèse de Fabienne Moreau, *Revisiter le couplage traitement automatique des langues et recherche d'information* `http://tel.archives-ouvertes.fr/docs/00/52/45/14/PDF/these_Fabienne_Moreau.pdf`

Plan

Fouille de textes

Fouille de textes = Lexicométrie + Fouille de données

- Dégager et structurer le contenu, les thèmes dans une perspective d'analyse rapide (non litéraire)
- Découverte d'information cachées $\Rightarrow$ surtout recherche d'information (RI)
- Prise de décision $\Rightarrow$ surtout extraction d'information (EI)

Différences entre RI et EI (1/2)

- La RI s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent, pour comparer les documents entre eux et détecter des typologies
- L'EI recherche des informations précises dans les documents, sans les comparer entre eux, en tenant compte de l'ordre et de la proximité des mots pour discriminer des énoncés différents ayant des mots clés similaires
  $\Rightarrow$ grande complexité de l'EI, qui doit effectuer une analyse lexicale et morpho-syntaxique pour reconnaître les constituants du texte (phrases, mots), leur nature et leurs relations.

Différences entre RI et EI (2/2)

- L'EI consiste en l'alimentation d'une base de données structurée à partir des données exprimées en langage naturel
- Il s'agit de détecter dans le texte en langage naturel les mots ou syntagmes correspondant à chaque champ de la base de données
- La RI cherche à détecter tous les thèmes présents
- L'EI ne s'intéresse qu'aux thèmes en rapport avec la base de données "cible"

Plan

Application des statistiques et de la FD

- Application des techniques de la FD :
  - individus = documents
  - caractères des individus = thèmes / termes des documents
- Remarque:
  - Les thèmes peuvent être très nombreux (plusieurs miliers) si le nombre de documents est important
  - On aboutit à des problèmes de FD avec un grand nombre de variables (plus de variables que d'individus)
  - Intérêt de :
    - techniques puissantes de FD
    - réduire le nombre de thèmes grâce à l'analyse linguistique

Techniques descriptives applicables

- Classification des documents
  - selon des thèmes non prédéfinis (découverts dans les documents)
  - suivie d'une extraction automatique des mots clés (thèmes / termes fréquents dans le segment et rares dans l'ensemble des documents)
- Analyse factorielle des correspondances

Techniques prédictives applicables

- Classement des documents
  - selon des thèmes prédéfinis (nomenclature)
  - utilisé pour du filtrage de documents
- Utilisation des chaînes de Markov pour les requêtes ouvertes (libres)

Représentation graphique

- On peut dresser une cartographie des documents et repérer :
  - les thèmes isolés
  - les thèmes formant des ensembles homogènes
  - l'intensité des liens entre thèmes d'un même ensemble (vocabulaire et problématique commune aux thèmes)
  - le nombre de documents pour chaque thème.

Plan

Problème général

- Un ensemble de documents
- Utilisateurs avec besoin d'information : requête en langage naturel
- SRI : établir un lien pertinent entre les documents et la requête

Composants d'un SRI

3 modules

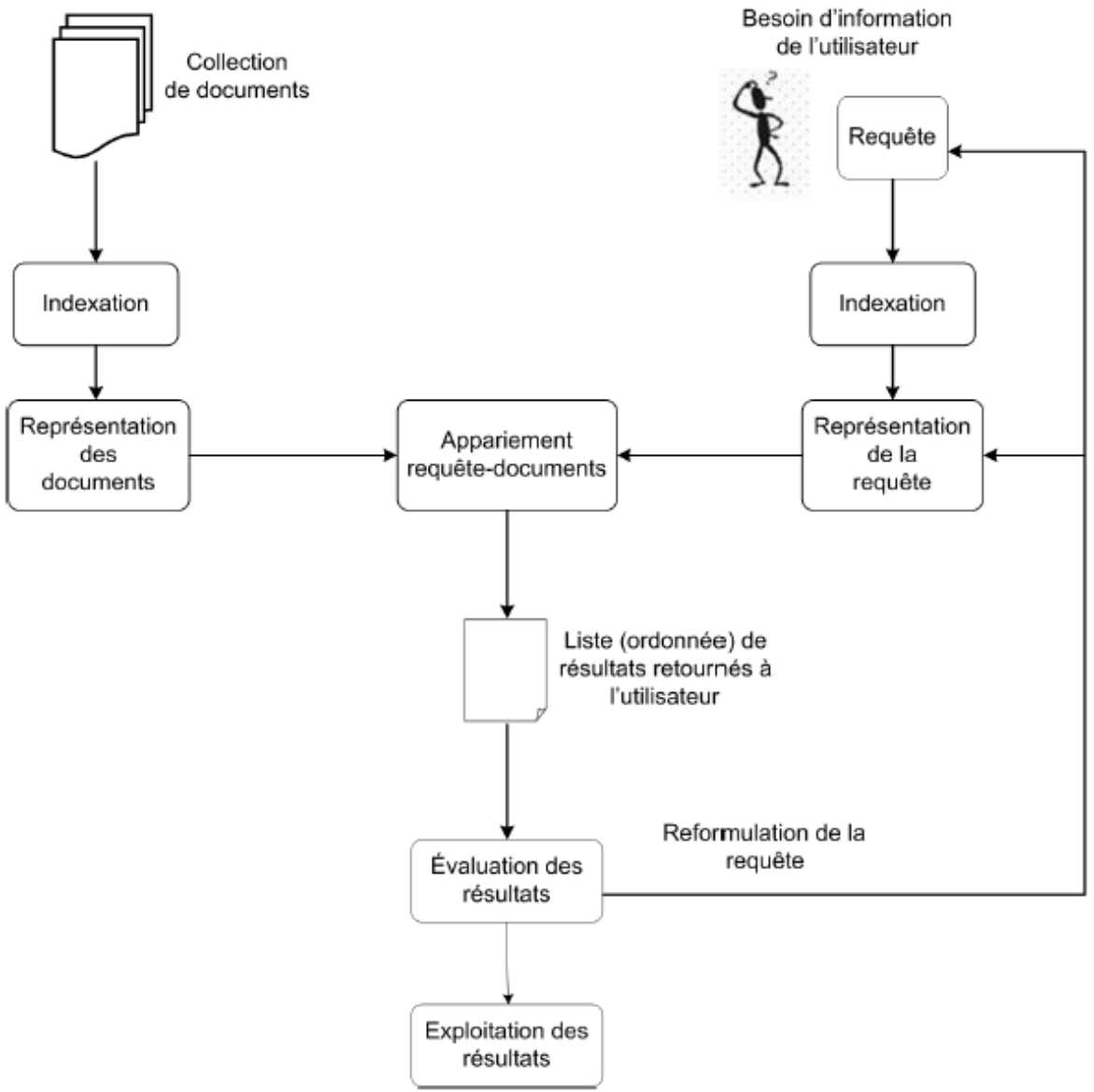
- module d'indexation de questions
- module d'indexation de documents
  $\Rightarrow$ identifier les idées majeures, les concepts importants des textes ou des questions, par une analyse de leurs contenus.
  $\Rightarrow$ chercher des représentants de ces concepts : ensemble de mots extraits des documents et requêtes - **termes d'indexation**
- module pour mettre en correspondance des documents et requêtes, représentés par des descripteurs (index).

$\Rightarrow$ liste des documents considérés comme les plus pertinents par rapport à sa requête : mesures d'évaluation ?

Principaux acteurs du processus de RI

- Le document : toute unité susceptible de constituer une réponse à une requête d'un utilisateur (texte, page web, image, séquence vidéo ou sonore, etc.).
  Pour les textes :
  - 3 vues : présentation (format), logique (structure du document), sémantique (contenu textuel)
- La requête : besoin d'information d'un utilisateur, exprimé sous forme de mots-clés (avec éventuellement des opérateurs booléens) ou de phrases en langage naturel (faisant appel à des traitements linguistiques)
- La pertinence :
  - du point de vue de l'utilisateur : jugements de pertinence de l'utilisateur - difficile à mesurer
  - du point de vue du système : calculée par le système à partir des méthodes utilisées pour comparer les documents et la requête.

Processus de RI



Indexation des documents et requêtes

- Reconnaissance des mots
- Sélection des termes d'indexation
- Pondération des termes

Reconnaissance des mots

- Tokenisation - plus ou moins difficile selon la langue traitée
  - langues agglutinantes comme allemand : plusieurs mots au sein d'une même chaine de caractères
  - langues isolantes comme chinois, vietnamien : pas de délimiteurs clairs des mots
  - aujourd'hui, 3h, après-midi, Mr J.-P. Martin, parce que, l'O.N.U
- Certains SRI s'appuient sur des modèles de *n*-grammes : pour chaque position *i* dans le texte, sélectionne une suite de *n* caractères consécutifs.
  $\Rightarrow$ Le texte est représenté par l'ensemble de ces extraits.
- Nécessité de prendre en compte du phénomène de variation linguistique (par ex. beau, belle, beaux, belles) et de la relation entre des termes (par ex. des expressions ayant des signification dépassant celle de ses mots composants pris isolément)
  appliquer une analyse linguistique des textes et des questions

Analyse linguistique (1/7)

- Identification de la langue : le web oblige à gérer le multilinguisme
- Phrases polyglottes : mélange plusieurs langues
- Identification des catégories grammaticales
  - noms / verbes / adjectifs / adverbes
  - parfois difficile : “ les poules du couvent couvent “

Analyse linguistique (2/7)

Désambiguïsation :

- Par exemple :
  - ELEVE $\rightarrow$ élève (n), élève (v), élevé (adj), élevé (pp)
  - Ce bureau ferme à cause des émeutes $\neq$ Ce bureau fermé a causé des émeutes.
  - Nous portions des portions d'avocats aux avocats.

Analyse linguistique (3/7)

Reconnaissance des mots composés :

- Expressions comme : France Telecom, le 21 juin 2008, le gouverneur de la Banque Centrale Européenne
- Prise en compte éventuelle d'un lexique spécialisé : fouille de données, carte bleue, ...
- Elaboration d'un lexique propre à l'entreprise
  en repérant les suites de formes graphiques (souvent 2 ou 3) se répétant plusieurs fois dans le corpus.

Analyse linguistique (4/7)

- Lemmatisation (mots ramenés à leur forme canonique)
  - substantifs ramenés au singulier
  - adjectifs ramenés au singulier masculin
  - flexions d'un verbe ramenées à l'infinitif
- Un dictionnaire général contient 60.000 entrées qui correspondent à 700.000 formes fléchies
- Le français, l'espagnol et l'allemand ont de nombreuses formes fléchies (conjugaisons ou déclinaisons)

Analyse linguistique (5/7)

Regroupement des variantes

- graphiques : clef = clé
- syntaxiques : complément de nom = complément nominal
- sémantiques : X achète Y à Z = Z vend Y à X
- synonymes : US = USA = Etats-Unis = Oncle Sam
- parasynonymes (mots de sens voisins) : mécontentement, colère, insatisfaction
- développement des sigles : EUR = euro ; E.D.F = EDF = Electricité de France
- métaphores : Empire du Soleil Levant (Japon), Le Pentagone (l'armée américain)

Analyse linguistique (6/7)

Regroupement des analogies

- familles de mots-dérivés : crédit / prêt / engagement / dette / emprunter / emprunteur / débiteur
- marqueurs d'intensité : peu / moins / très peu ; beaucoup / plus / très

Analyse linguistique (7/7)

Identification des thèmes

- des termes aux thèmes de niveau 1 : chéquier / carte bleue / TIP / devises ... $\Leftrightarrow$ moyen de paiement
- des concepts de niveau 1 aux thèmes de niveau 2 : moyen de paiement / monnaie / argent ... $\Leftrightarrow$ banque

Indexation des documents et requêtes

- Reconnaissance des mots
- Sélection des termes d'indexation
- Pondération des termes

Sélection des termes d'indexation(1/2)

Sélection éventuelles des termes / thèmes

- d'après un critère statistique : élimination des termes/thèmes fréquents
- d'après un critère sémantique : sur un sujet donné
- d'après un corpus : repérage des mots à éviter et de leurs dérivations (expurgation du document)

Choisir les termes qui reflètent mieux le contenu sémantique

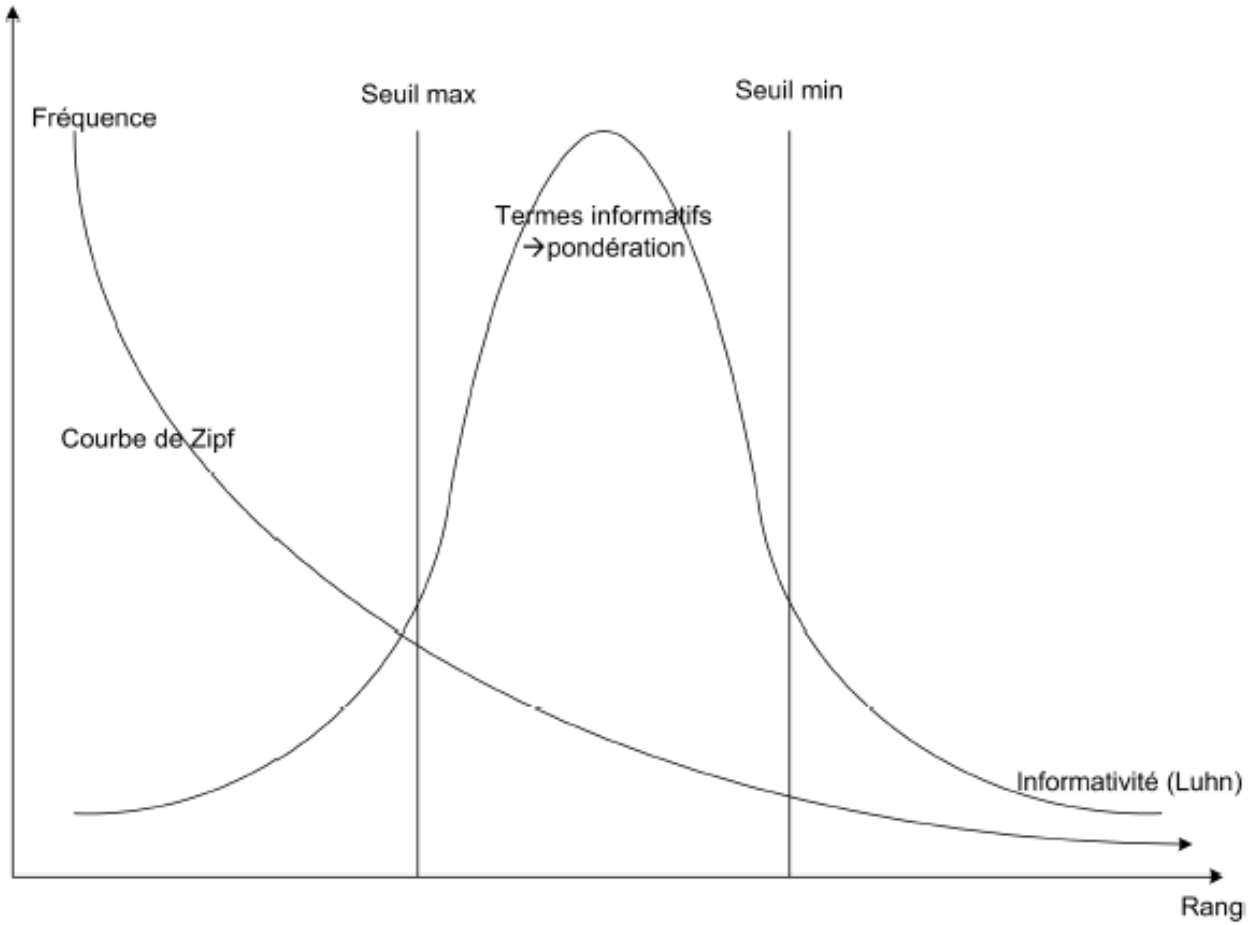
- Elimination des mots-vides (stop-lists) :
  souvent ce sont des mots "grammaticaux", ou des mots très fréquents au sein d'une collection de textes donnée (par ex. "informatique" dans un corpus spécialisé dans ce domaine)

Sélection des termes d'indexation(2/2)

Choisir les termes qui reflètent mieux le contenu sémantique (suite)

- Analyse basée sur les fréquences d'occurences des mots : loi de Zipf, conjecture de Luhn
  - Les termes rangés par ordre décroissant de leur fréquence d'apparition dans un texte
  - Loi de Zipf (1949) : rang * (fréquence du terme / nombre de termes) = constante
  - Conjecture de Luhn (1978) : seuil min, max
  - Salton (1975) propose d'utiliser l'intervalle [|C|/100, |C|/10] où |C| - nombre de mots dans la collection

Loi de Zipf et conjecture de Luhn



Indexation des documents et requêtes

- Reconnaissance des mots
- Sélection des termes d'indexation
- Pondération des termes

Pondération des termes

Combinaison de 3 facteurs

- Pondération locale : importance du terme dans le document
- Pondération globale : représentativité du terme dans l'ensemble de la collection de documents
- Facteur de normalisation : prise en considération la taille du document

Pondération locale (1/2)

- Poids d'un terme t_i dans un document d_j : $w(t_i, d_j)$.
- Facteur binaire = 1 si le terme est présent, = 0 si le terme est absent
- Facteur de fréquence *tf* : nombre d'occurrences d'un terme donné dans le document
- Facteur de fréquence normalisée : prend en compte la fréquence d'un terme dans le document et mesure son importance relativement aux autres termes du document

$$w(t_i, d_j) = \frac{tf(t_i)}{\max_{t \in d_j} tf(t)}$$

Pondération locale (2/2)

- Facteur logarithmique : ne pas accorder plus d'importance à un document qui possède un grand nb de fois un des termes de la requête par rapport à un document qui contient peu de fois plusieurs termes de la requête

$$w(t_i, d_j) = \begin{cases} 1 + \log(tf(t_i)) \text{ si } tf(t_i) > 0 \\ 0 \text{ sinon} \end{cases}$$

- Facteur augmenté : réduire les différences entre les valeurs associées aux termes du document

$$w(t_i, d_j) = \begin{cases} 0,5 + 0,5 \frac{tf(t_i)}{\max_{t \in d_j} tf(t)} \text{ si } tf(t_i) > 0 \\ 0 \text{ sinon} \end{cases}$$

Pondération globale

- Mesure l'importance d'un terme au sein de l'ensemble des documents de la collection
- Un terme apparaissant dans peu de documents est considéré comme plus discriminant $\Rightarrow$ priviligé
- Fréquence documentaire inverse *idf* (*inverse document frequency*)

$$idf(t_i) = log\frac{N}{n_i}$$

n_i nombre de documents contenant le terme t_i, N nombre total de documents

- Une variante de cette mesure :

$$idf\_prob(t_i) = log\frac{N - n_i}{n_i}$$

Normalisation

- Intégration de la taille des documents aux formules de pondération
- Deux facteurs souvent utilisés (normalisation par le cosinus) :

$$\frac{1}{\sum_{i=1}^{n}(l_i \cdot g_i)} \quad \text{et} \quad \frac{1}{\sqrt{\sum_{i=1}^{n}(l_i \cdot g_i)^2}}$$

l_i et g_i pondérations locales et globale du i-ième terme et n nombre de termes d'indexation.

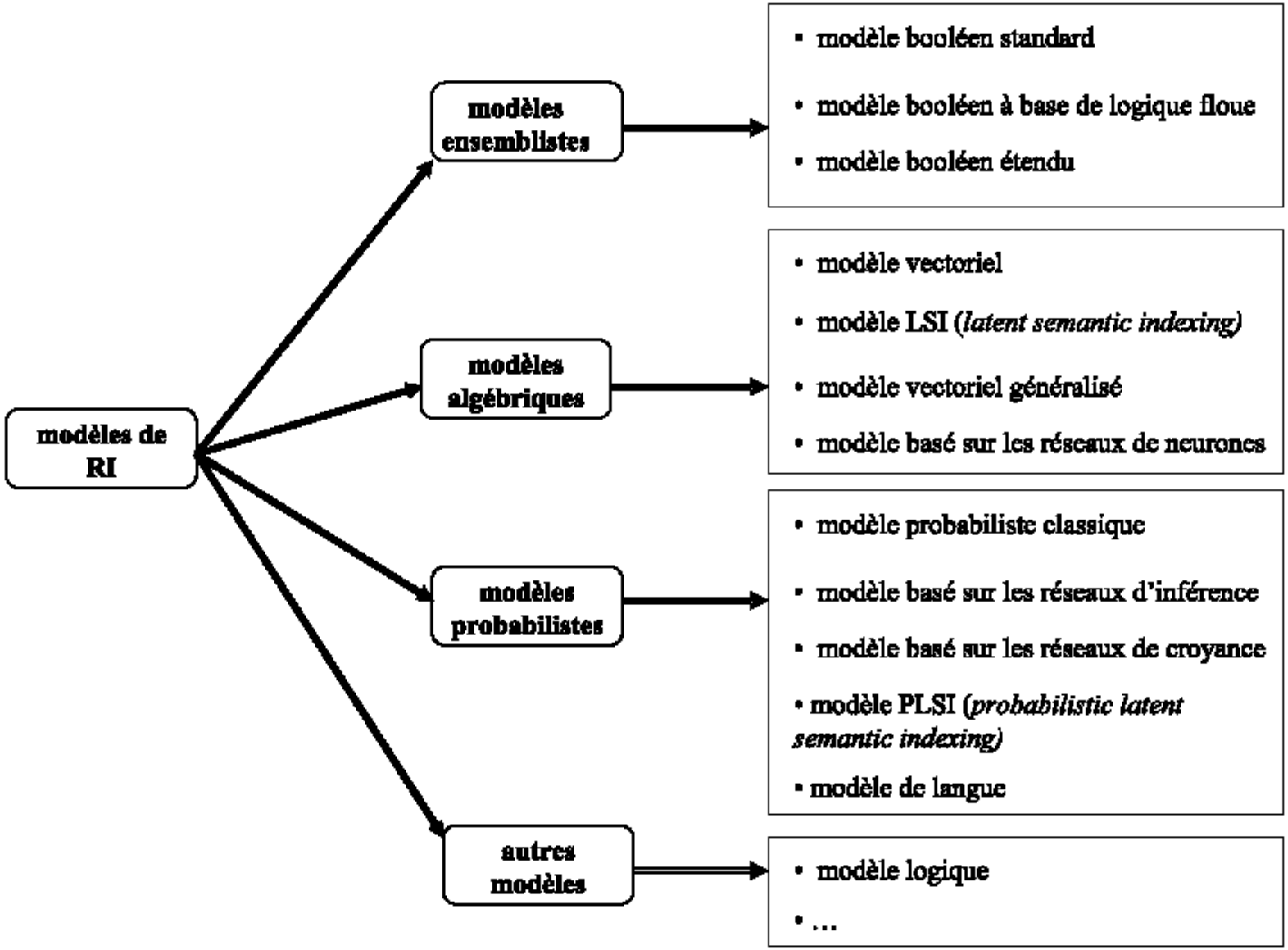
Combinaison des pondérations

- $w(t_i, d_j) = l_i \times g_i \times n_i$
- Chaque document (et requête) associé à un ensemble de termes pondérés - descripteur représentatif de son contenu.

Processus de recherche des documents pertinents

- Créer une représentation interne des textes et des questions à partir de leurs termes d'indexation
- Comparer ces représentation
- Calculer le score qui détermine le degré de pertience (système) du document d par rapport à la requête q : $RSV(d, q)$
  (*Retrieval Status Value*)

Principaux modèles



Evaluation d'un SRI

- On se place en RI textuelle: collection de documents, requêtes, réponses aux requêtes
- Problème :
  - Comment évaluer la qualité d'un SRI?
  - Comment savoir de 2 SRI lequel est le meilleur?

Approche d'évaluation

- Evaluation par *benchmarking* : mais les performances sont valides sur ce seul *benchmark*
- Collection de test: mais représentativité des thèmes, de la taille du corpus etc. difficile
- Par ailleurs, la construction des requêtes pour l'évaluation est difficile.

Problème

- La pertinence d'un document
  - dépend de l'utilisateur,
  - dépend des besoins actuels de l'utilisateur,
  - évolue dans le temps

Evaluation classique – hypothèses simplificatrices

- Pertinence binaire : un document est considéré soit pertinent, soit non pertinent. Aucune graduation de pertinence n'est prise en compte dans la tâche d'évaluation ;
- Jugements de pertinence absolus : aucune remise en question de ces jugements n'est possible ;
- Non-additivité : chaque document est évalué indépendamment des autres membres de la collection. La pertinence d'un document ne dépend pas des autres textes répondant à la requête ;
- Absence de mémoire : un document reste pertinent même si un autre au contenu similaire a déjà été présenté à l'utilisateur.

Matrice de confusion

Réalité	Prédit pertinents	Prédit non-pertinents
Pertinents	VP	FN
Non pertinents	FP	VN

Mesures classiques

- Rappel : $R = VP/(VP + FN)$
- Précision : $P = VP/(VP + FP)$
- Bruit : $B = 1 - P$
- Silence : $S = 1 - R$
- F-mesure = $P * R/(\alpha P + (1 - \alpha)R)$, $(\alpha \in [0, 1])$
  $F_1 = 2 * P * R/(P + R)$ $(\alpha = 0.5)$

Mesures complémentaires

- Cas limites:
  - Aucun document sélectionné
  - Tous les documents sélectionnés
- Précision moyenne interpolée : moyenne des précisions obtenues en fixant R
- Moyenne : MAP (*Mean Average Precision*) moyenne des précisions obtenues après chaque document pertinent retrouvé

Mesures orientées utilisateur (1)

- Mesures subjectives:
  - Novelty ratio : capacité à trouver des nouveaux documents
  - Coverage ratio : capacité à retrouver des documents déjà vus

Mesures orientées utilisateur (2)

- D'autres critères
  - le temps mis par le système pour fournir des réponses à l'utilisateur ;
  - l'effort effectué par l'utilisateur pour obtenir l'information recherchée (e.g. le nombre de requêtes qu'il a dû formuler avant d'avoir le résultat recherché) ;
  - la qualité de la présentation des résultats par le système (e.g. à partir de la liste de résultats fournis par le système, combien de documents l'utilisateur a-t-il dû parcourir avant de trouver le document recherché ?).

Conclusion pour la RI

- Pas de mesure idéale
- Beaucoup de questions :
  - Adéquation des tâches d'évaluation aux problèmes réels ?

Plan

Applications

- Catégoriation de textes
  - Classes de documents en fonction de leurs thèmes ou leur forme, permettant l'indexation automatique ou le filtrage automatique de documents
  - Détection de spam
  - Analyse de sentiments, d'opinions ...
  - Extraction de relations entre des entités : reconnaissance de entités (nommés), classifier leurs relations ...
- Classification de textes : apprentissage non supervisé
  - Partitionner les documents en groupes homogènes $\Rightarrow$ interpréter la nature des groupes
- Analyse des tendances, résumé automatique

Représentation de données textuelles

- Représentation des documents par les sacs de mots (BOW) : vecteurs multidimensionnels
  - Réduction de la dimensionnalité grâce aux ontologies (regrouper des mots synonymes, utiliser les relations des concepts)
  - Réduction de dimension avec LSA (latent semantic analysis)
- Plongement de mots (word embedding) : chaque mot est représenté par un vecteur de nombres réels reflétant son contexte

https://eric.univ-lyon2.fr/ricco/cours/slides/TM.A%20-%20introduction%20text%20mining.pdf
https://tutoriels-data-mining.blogspot.com/search/label/Text%20Mining
http://wordvec.colorado.edu/word_embeddings.html
https://www.tensorflow.org/tutorials/representation/word2vec