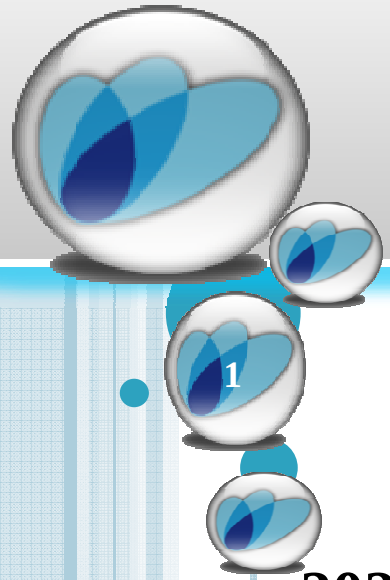
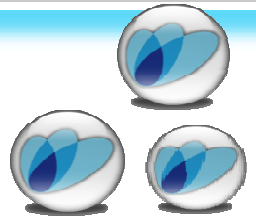
Cours Recherche D'information (Information Retrieval)

3ème Année Licence

Option : ISIL





2023-2024

Mme Z.LAAREDJ

Le modèle booléen

Plan

- Définition du modèle booléen
- Les concepts du modèle booléen.
- Principe d'Appariement du modèle booléen.
- Avantages du modèle booléen
- Limites du modèle booléen

Définition de modèle booléen

- Le modèle booléen ou logique, est historiquement le premier modèle de RI. Le modèle booléen de base (strict) repose sur la théorie des ensembles et la logique des propositions.
- Un document est représenté par une liste de termes (termes d'indexation).
- Une requête est représentée sous forme d'une équation (ou expression) logique (ou booléenne).
- Les termes d'indexation au sein des documents et des requêtes sont reliés par des **opérateurs logiques** (connecteurs logiques)[1] :

OU (∨), ET (∧) et NON (¬).

Définition du modèle booléen

- Le modèle booléen est [3] :
  - Un modèle exact match le plus commun,
  - Un document est un ensemble de termes,
  - Une requête est une expression logique formée de:
    - Termes
    - Opérateurs booléens ET / OU / NON (SAUF)

Les concepts du modèle booléen

➢ Le modèle de recherche booléen est défini par un quadruplet (T, Q, D, F) Où:

- T : Ensemble des termes d'indexation **T={t1,t2,….tn)**
- D : Ensemble des documents de la collection: **D={d1,d2,….dm)**,

  ex: d1(t1,t2,t5);
- Q : Ensemble de requêtes **Q={q1,q2,…. qk)** ,

  ex: q = t1 ∧(t2 ∨¬t3);
- F : Fonction de pertinence définie par :

$$\mathbf{F: D \times Q \longrightarrow \{0,1\}}$$

$$F(d,t) = \begin{cases} 1 \text{ si } t \in d \\ 0 \text{ sinon} \end{cases}$$

Les concepts du modèle booléen

- Modèle de documents [1]
  - Pour le modèle booléen de base, un document est représenté par ensemble de termes indépendants noté : d(t1,t2,…,tn).
  - Exemple : soient 3 documents:

    d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)

Les concepts du modèle booléen

- Modèle de requête [1]
  - Pour l'utilisateur, une requête est un ensemble de termes reliés avec des opérateurs booléens : AND (∧), OR(∨), NOT (¬)
  - L'ensemble de requêtes est noté: Q = {*q1, q2, · · · , qk*} *ou k représente le nombre de* requêtes.
  - Exemple : soit la requête q:

$$q = t1 \wedge (t2 \vee \neg t3)$$

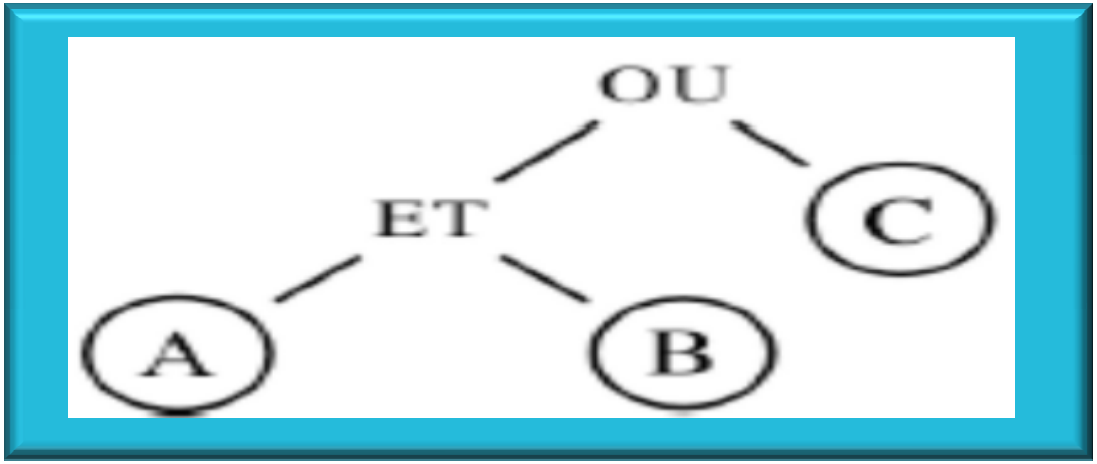
Les concepts du modèle booléen

➢ Une requête est le plus souvent formulée dans un langage de requête spécifique au système. Plusieurs types de langage de requêtes peuvent être utilisés :

- Simples
  - Ensemble de mots ou sac de termes: Bag of words
  - Une phrase ou un paragraphe en langage naturel ;
- Complexes
  - Expressions booléennes ;
  - Expressions régulières ;
  - Langage structuré précisant la valeur d'attributs tels que les noms d'auteurs, les mots du titre etc.;
  - Expression de relations de proximités pondérées entre les mots.

Les concepts du modèle booléen

- Une requête est une expression booléenne qui peut se modéliser avec un arbre où les feuilles sont des termes (pondérés ou non), et les nœuds internes sont des opérateurs AND (∧), OR(∨), NOT (¬)[5].
- Par Exemple : la requête q=(A et B) ou C est représentée par l'arbre de la figure suivante:



Arbre de la requête (A et B) ou C [4]

Principe d'Appariement du modèle booléen

- Avec le modèle Booléen, le module de recherche mis en œuvre consiste à effectuer des opérations sur l'ensemble de documents afin de réaliser un appariement exact avec l'équation de la requête.
- L'appariement exact est basé sur la présence ou l'absence des termes de la requête dans les documents.
- La décision binaire sur laquelle est basée la sélection d'un document ne permet pas d'ordonner les documents renvoyés à l'utilisateur selon un degré de pertinence.
- La fonction de correspondance est basée sur l'implication logique en logique des propositions (logique d'ordre 0)[5].

Principe d'Appariement du modèle booléen

- Appariement Exact est basé sur la présence ou l'absence des termes de la requête dans les documents.
- À cette étape d'Appariement, le modèle booléen permet de calculer **la fonction de pertinence** appelée également score de similarité ou **Relevance Status Value (RSV)** pour un couple (document, requête). Cette mesure est notée RSV (di, qj) et retourne le score de similarité du document di par rapport à la requête qj[5].

Principe d'Appariement du modèle booléen

➢ La correspondance RSV(d, q) entre une requête et un document est déterminée de la façon suivante:

$R(d, t_i) = 1$ *si* $t_i \in d$; *0 sinon.*

$R(d, q_1 \wedge q_2) = 1$ *si* $R(d, q_1) = 1$ *et* $R(d, q_2) = 1$; *0 sinon.*

$R(d, q_1 \vee q_2) = 1$ *si* $R(d, q_1) = 1$ *ou* $R(d, q_2) = 1$; *0 sinon.*

$R(d, \neg q_1) = 1$ *si* $R(d, q_1) = 0$; *0 sinon.*

Principe d'Appariement du modèle booléen

- Le score de similarité RSV(d,q)
  - La mesure RSV(d,q) est une fonction de correspondance (de similarité, de similitude) entre un document et une requête.
  - Dans le modèle booléen de base, tous les documents qui satisfont une requête sont retrouvés (généralement classés dans un ordre chronologique).
  - Ils ne sont pas classés selon leur pertinence . Cela est dû au fait qu'un document satisfait une requête ou ne la satisfait pas (1 ou 0).

Principe d'Appariement du modèle booléen

- Le modèle booléen peut être utilisé efficacement sur des collections de documents spécialisés dans le cas où les utilisateurs ont une bonne connaissance du vocabulaire associé à ces collections.
- Pas pour un large public.

Avantages du modèle booléen

- Simple à mettre en œuvre
- Transparent pour l'utilisateur dans la mesure où les documents sont retournés selon un appariement exact avec le besoin en information.
- La clarté conceptuelle des systèmes booléens[2].

Limites du modèle booléen

- Tous les termes dans un document ou dans une requête étant pondérés de la même façon simple (0 ou 1) c à d, indexation binaire,
- La sélection d'un document est basée sur une décision binaire,
- Pas d'ordre pour les documents sélectionnés,
- La formulation ou écriture d'une requête booléenne est difficile pas toujours évidente pour beaucoup l'utilisateurs (simple utilisateur),
- Problème de collections volumineuses : le nombre de documents retournés peut être considérable,
- Le modèle ne permet pas de retourner un document s'il ne contient qu'une partie des mots de la requête (si le connecteur ET est utilisé)[2,3] .

Références bibliographiques

[1] https://www.irit.fr/~Mohand.Boughanem/slides/RI/chap4-mod-bool-vect.pdf

[2] http://univ.encyeducation.com/uploads/1/3/1/0/13102001/mi3an10-recherche_information.pptx

[3] https://lipn.univ-paris13.fr/~rozenknop/Cours/MICR_REI/Seance6/modeles-RI-1.pdf

[4]https://tel.archivesouvertes.fr/file/index/docid/785143/filename/2006_these_A_mercier_417I.pd

[5] https://www.youtube.com/watch?v=BDi3drDPibY