Laborator nr. 4
Modul de căutare booleană

an univ. 2018 – 2019

1 Scurt breviar teoretic

1.1 Caracteristici

Reprezentarea documentului – vector de ponderi, pentru care ponderea w_{ij} este definită conform ecuaţiei de mai jos:

$$w_{ij} = \begin{cases} 1 & \text{dacă } t_i \in d_j \\ 0 & \text{altfel} \end{cases} \tag{1}$$

Reprezentarea interogării – termenii interogării (sau *cheile de căutare*) sunt combinate logic utilizând operatorii booleeni **AND**, **OR** şi/sau **NOT**.

Regăsirea documentului – se bazează pe **criteriul deciziei binare** şi pe **aritmetica mulţimilor**.

Avantaje:

- Este un model de căutare simplu, cu un formalism bine pus la punct, neambiguu.
- Poate fi implementat uşor şi poate răspunde rapid pentru interogările uzuale ale utilizatorilor.

Dezavantaje

- Datorită simplităţii, este un model foarte rigid.
- Interogările complexe nu pot fi realizate direct.
- Nu poate fi controlată cu exactitate dimensiunea exactă a răspunsului.
- Nu oferă un mecanism direct de *feedback* din partea utilizatorilor.

1.2 Principalii paşi implicaţi

1. se citeşte interogarea utilizator;
2. se izolează *operanzii* (cuvintele) de *operatori*;
3. cuvintele se procesează conform modelului utilizat în construirea index-ului corespunzător;
4. se izolează pe baza index-ului invers, pentru fiecare cuvânt în parte lista de documente ce conţin termenul respectiv;
5. se realizează, rând pe rând, operaţiile indicate **AND**, **OR** şi/sau **NOT**:
   - **AND** echivalează cu intersecţia a două mulţimi;
   - **OR** echivalează cu reuniunea a două mulţimi;
   - **NOT** echivalează cu diferenţa dintre două mulţimi;
6. rezultatul obţinut este prezentat utilizatorului.

2 Aplicaţie propusă

Implementaţi o aplicaţie care să realizeze funcţia de căutare a unui sistem de regăsire de informaţii conform **modelului boolean** ([1], *Capitolul 1: Boolean retrieval*). Operaţiile dorite sunt:

- *k1* **AND** *k2*: setul de documente care conţin atât termenul *k1*, cât şi termenul *k2* – **intersecţie de mulţimi**;
- *k1* **OR** *k2*: setul de documente care conţin fie termenul *k1*, fie termenul *k2* – **reuniune de mulţimi**;
- *k1* **NOT** *k2*: setul de documente ccare conţin termenul *k1*, dar nu conţin termenul *k2* – **diferenţă de mulţimi**.

Aplicaţia trebuie să poată procesa şi interogări ce conţin mai mult de două chei de căutare (exemplu: *k1* **AND** *k2* **NOT** *k3* **OR** *k4*). Indexul invers de test este cel determinat în cadrul laboratorului nr 3.

Bibliografie

[1] Christopher D. Manning et. al. Introduction to Information Retrieval. `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`, 2009.