Sisteme de regăsire a informaţiilor – Metode de indexare şi căutare a datelor

Anca Farcaş

Rezumat

Tema abordată în lucrarea de diplomă este **procesul de indexare şi căutare în motoarele de căutare.** Indexarea poate fi privită ca un proces ce oferă o relaţie de ordine asupra datelor pe care le analizăm. În această lucrare se va pune accent pe două tipuri de indexare şi anume, **indexarea directă** şi **indexare inversă**.

Indexarea reprezintă procesul prin care datele unei colecţii sunt reorganizate în scopul de a fi regăsite uşor şi precis. **Indexarea directă** are ca scop determinarea indecşilor relativ la document. I**ndexare inversă** are drept scop determinarea documentelor relativ la index. Aceasta reprezintă componenta de bază a oricărui motor de cătuare, utilă în determinarea relevanţei unui cuvânt la colecţia de documente în care se face căutarea.

Printre etapele principale ale indexării directe, se numără **pre-procesarea** şi construirea **indecşilor adecvaţi**. Pentru preprocesare, a fost considerată componenta de analiza a formei de bază a unui cuvânt. Pentru a aduce cuvântul la forma sa canonică au fost construiţi doi arbori lingvistici care comnică între ei, unul din aceştia fiind populaţi cu toate cuvintele limbii engleze, iar al doilea cu formele neregulate a substantivelor, verbelor, adverbelor (exemplu: man - men, mouse - mice), fiecare cuvânt din arborele neregulat având legătură cu forma de bază a cuvântului ce îi coincide în arborele dicţionar. Scopul acestei abordări este de a soluţiona acele cazuri în care algoritmi clasici de stemming (*Porter*, *Lovins* şi *Paice*) nu oferă rezultate satisfăcătoare. Această formă de preprocesare este aplicată atât în faza de construire efectivă a indexului, cât şi în componenta de căutare.

Fiecare fişier din colecţia de documente deja indexate direct, vor fi trecute prin procesul de **indexare inversă**, proces care va ajuta la cautarea rapidă a unui cuvânt în întrega colecţie de fişiere.

Pentru componenta de căutare, accentul s-a pus pe **modulu de căutare** de tip **boolean** şi cel bazat pe **reprezentarea vectorială**.

În ceea ce priveşte **modelul boolean** de căutare, reprezentarea documentului se realizează sub formă de ponderi. Termenii interogării sunt combinaţi logic utilizând operatorii booleani AND, OR şi NOT, iar regăsirea documentelor căutate se bazează aritmetica mulţimilor şi a criteriului deciziei binare. Acest modul de căutare boolean a ajutat la restrângerea colecţiei de documente pentru interogarea utilizatorului.

Pentru **modulul de căutare bazată pe reprezentarea vectorială**, reprezentarea documentului, cât şi a interogării, constă în vectori de ponderi şi se obţine rapid dintr-un **index direct cantitativ**. Regăsirea documenetelor de interese se bazează pe distanţa dintre două documente calculată ca fiind unghiul dintre doi vectori n-dimensionali. Acest modulul, bazat pe reprezentarea vectorială a oferit seturi relevante de documente pentru cererea utilizatorului.