Motor de căutare semantic

Anca-Nicoleta Ciubotaru

Rezumat

Cantitatea datelor nestructurate a avut de-a lungul timpului o creştere exponenţială, iar regăsirea informaţiilor de interes s-ar realiza cu dificultate fără o metodă eficientă de indexare, ordonare şi clasificare.

Tema proiectul constă în crearea unui motor de căutare ce ia în considerare caracterul semantic al cuvintelor, facilitând procesul de regăsire şi clasificare a informaţiilor pe baza sensului interogării efectuate de utilizator. Diferenţa faţă de implementările motoarelor de căutare clasice, care redau informaţiile independent de sensul termenilor din interogare, constă în analiza contextului în care aceştia apar. Unul dintre avantajele acestei abordări este faptul că în urma unei interogări date de utilizator, informaţiile sugerate de motorul de căutare vor aparţine unui întreg domeniu de referinţă, documentele cu caracter similar fiind considerate relevante pentru utilizator.

Motorul de căutare semantic realizează următoarele sarcini: vizitarea paginilor de pe World Wide Web, salvarea conţinutului, acolo unde este permis, indexarea şi generarea sensurilor cuvintelor pe baza contextului în care acestea apar. În urma introducerii unei interogări, utilizatorul va obţine rezultate constând în pagini relevante din punct de vedere a sensului, dar şi reprezentarea vectorială a cuvintelor reprezentative.

Arhitectura este modulară, fiecare modul având un scop clar definit în cadrul ansamblului din care face parte. Una dintre componentele de bază este un modulul ce realizează căutarea, selecţia şi indexarea documentelor în vederea minimizării timpului de regăsire. Pentru reducerea timpului de prelucrare, etapele sunt implementate prin intermediul modelului de programare *Map Reduce*, ce realizează eficient procesarea paralelă a unor cantităţi mari de date. Modulul ce realizează procesarea informaţiilor prin generarea sensurilor cuvintelor raportate la contextul în care apar utilizează modelul arhitectural *Skip Gram,* ce determină reprezentările vectoriale ale cuvintelor prin intermediul unei reţele neuronale. Astfel, fiecare cuvânt are asociat un set de vectori n-dimensionali pe baza cărora se vor putea identifica sinonimele şi sensurile cuvintelor, luând ca unitate de măsură distanţa cosinus dintre aceştia. Modulul ce realizează legătura dintre utilizator şi nucleul de procesare îi permite acestuia să obţină informaţii de interes. Astfel, acesta va vizualiza câmpul vectorial aferent rezultatelor obţinute implementat prin intermediul algoritmului *t-distributed Stochastic Neighbor Embedding* (t-SNE) ce are ca scop reducerea numărului de dimensiuni ale vectorilor n-dimensionali şi generarea grafică în două sau trei dimensiuni.

Modul de căutare a informaţiilor pe World Wide Web are un caracter desprins din limbajul natural şi deşi cuvintele cheie sunt importante în vederea obţinerii informaţiilor de interes, contextul în care acestea apar este necesar a fi luat în considerare în procesul de construire a unui motor de căutare care să răspundă cerinţelor date de utilizator.

Îndrumător: Conf. dr. ing. Mihai Horia Zaharia