Sisteme de regăsire a informațiilor. Metode avansate de analiză a datelor

Roxana Ungureanu

Rezumat

Tehnologia de astăzi ne permite să producem și să stocăm un volum imens de date. Identificarea de tipare, tendințe și anomalii în seturile de date și sintetizarea lor în modele clare este una din provocările erei informațiior.

Problema pe care mi-am propus să o adresez este optimizarea rezultatelor pe care le oferă un motor de căutare care gestionează date pe Web. Provocările domeniului derivă din faptul că Web-ul este o colecție imensă de documente, complet nestructurată, nesupravegheată în dezvoltare și cu un conținut dinamic și interactiv. Un sistem de căutare performant la ora actuală ar trebui să vină în ajutorul clienților prin a grupa informațile oferite utilizatorilor.

În acest context, studiul pe care l-am realizat vizează optimizarea funcționalității motoarelor de căutare și creșterea performanțelor acestora tocmai prin identificarea unui sens a rezultatelor oferite.

Principalele componentele utilizate în această lucrare vizează domeniul *data mining*. După cum bine se știe, **data mining** înseamnă **identificarea informațiior cu potențial util, netriviale, din colecții mari de date**. Aplicarea acestor tehnici peste date de tip Web, respectiv, pe pagini Web sau pe informații de structură din spatele acelor pagini se numește ***Web mining***. O principală metodă de analiză care atinge scopul propus, este metoda de clusterizare.

Clusterizarea presupune identificarea unui grup finit de grupuri în cadrul colecției de entități, în cazul de față, aceste entități sunt documentele Web.

Principalele metode de clusterizare se bazează fie pe tehnici de partiționare, fie pe tehnici de aglomerare/divizare ierarhică, fie pe densitate. În această lucrare am scos în evidență avantajele și dezavantajele a trei metode de clusterizare a documentelor: clusterizare prin partiționare (K-medoids), clusterizare ierarhică (DIANA) și clusterizare bazată pe densitate (DBSCAN).

Rezultatele de performanță pentru fiecare dintre cele trei metode au fost obținute prin aplicarea unor implementări proprii ale algoritmilor K-medoids, DIANA și, respectiv, DBSCAN, pe un set de documente text. Rezultatele obținute par să favorizeze metodele bazate pe densitate.

În plus, pentru a facilita utilizarea clusterelor, am implementat trei metode de etichetare automată a clusterelor folosind două metode de selecție a trăsăturilor și o metodă proprie, asemănătoare cu metoda bazată pe frecvențe.