Metode de îmbunătățire a rezultatelor clusterizării documentelor HTML

Ana – Vladina Vasluianu

Rezumat

Ca urmare a evoluției tehnologiei și a metodelor de generare și stocare a unor cantități enorme de date, World Wide Web-ul a devenit o colecție imensă și nestructurată de documente distribuite, ce poate fi accesată prin intermediul internetului.

Drept urmare, principala provocare a motoarelor de căutare actuale este aceea de a oferi utilizatorilor informațiile necesare într-un timp cât mai scurt și cu un grad de relevanță cât mai bun. O problemă dificilă în acest context este reprezentată de structurarea rezultatelor oferite de motoarele de căutare în funcție de subiectul tratat.

Abordările clasice au la bază algoritmi de clusterizare a documentelor text, fiecare cuvânt din text primind în acest sens o pondere determinată după o regulă fixă.

Soluția propusă de mine urmărește identificarea unor metode de ponderare a importanței cuvintelor în funcție de eventualele formatări specifice ale acestora, pentru a obține o mai bună structurare a clusterelor rezultat.

Ponderile asociate cuvintelor vor fi direct proporționale cu gradul de relevanță pe care îl atribuie cel care a creat pagina elementelor ce alcătuiesc conținutul ei. De exemplu, dacă elementele de interes pentru utilizator vor fi formatate într-un mod diferit, puse în evidență (scrise îngroșat, italic, etc.), acestea vor primi o pondere mai mare față de situația în care cuvintele vor fi scrise normal.

Pentru a realiza acest lucru voi utiliza tehnici de tip data mining.

Analiza unor cantități mari de date și extragerea de informații cu potențial util din cadrul acestora poartă denumirea de **data mining**. Aplicarea acestei tehnici pe date Web: pagini web sau informații de structură din spatele lor se numește **web mining**. Printre metodele complexe de analiză a datelor se numără și algoritmi de clusterizare.

Metodele de clusterizare se clasifică în 3 mari categorii: metode bazate pe tehnici de partiționare, metode bazate pe tehnici de aglomerare/divizare ierarhică și metode bazate pe densitate.

În acestă lucrare am optat pentru optimizarea rezultatelor clusterizării agloritmului DBSCAN (metodă de clusterizare bazată pe densitate) datorită capacității acestui algoritm de a identifica clustere de dimensiuni și forme diferite și de a descoperi valorile extreme din cadrul seturilor analizate. Optimizarea rezultatelor algoritmului se va realiza prin procesarea documentelor HTML ținând cont de modului de formatare a cuvintelor în cadrul paginilor.