Extrakcia informácií zo štruktúrovaných webových zdrojov

Peter Kál
UPJS, Košice, Prírodovedecká fakulta

Internet v súčasnosti poskytuje takmer neobmedzené množstvo informácií. Vzhľadom na ich objem a obsah vzniká potreba automatizovaných prístupov detekcie a extrakcie relevantných informácií. V práci je navrhnutá séria algoritmov, ktoré slúžia na vyhľadanie informácií a ich následné uloženie v zrozumiteľnej forme za účelom ich budúceho spracovania. Na odhalenie sémantiky internetových stránok je použitá diferenčná metóda, ktorá produkuje anotácie stránok. Praktické použitie navrhnutých algoritmov je overené implementáciou extrakčného nástroja. Navrhnutý nástroj je určený hlavne pre prácu nad produktovými stránkami internetových katalógov, ktoré využívajú pevne definovanú štruktúru. V závere bola funkčnosť navrhnutého riešenia overená sériou štandardných testov. Získané výsledky poskytujú priestor pre porovnanie nového riešenia s už existujúcimi.