

Extrakcia dát na základe kategorizácie webových stránok

Bc. Pavol Rajzák, Prírodovedecká fakulta UPJŠ



**Automatická
kategorizácia
a extrakcia**

Lenovo Thinkpad T430



Prices

\$879 ([view](#))
\$899 ([view](#))
\$929 ([view](#))
\$999 ([view](#))

Processor: **Intel® Core™ i7**
Display: **14" HD or HD+ display**
Weight: **Starting at 4.77 lbs.**
Battery: **Up to 30 hours**
Storage: **500GB HDD, 256GB SSD**
RAM: **Up to 16GB [2x 8GB]**

Pros: Extremely long battery life; Comfortable Precision keyboard; Solid performance; ThinkVantage apps useful and intuitive; Runs cool

Cons: Relatively dim display; Small touchpad

The Verdict: A superb new keyboard and more than 13 hours of battery life make the ThinkPad T430 an excellent business laptop.

(ilustračný obrázok)

Využitie kategorizácie pri extrakcii informácií z webových stránok

Hlavná myšlienka navrhovaného systému je predspracovať webové stránky a vybrať vhodný **extrakčný mechanizmus**, ktorý by dokázal najlepšie spracovať informácie z webovej stránky.

Klasifikácia pozostáva z viacerých krokov a berie do úvahy ako štruktúru, tak aj obsah.

Výsledkom práce je systém, ktorý automaticky spracováva a extrahuje informácie o produktoch. Tieto informácie sú prezentované používateľovi s možnosťou uloženia do databázy a ďalšieho strojového spracovávanía.

Motivačný príklad: Používateľ si chce kúpiť nový notebook a hľadá vhodný model na internete.

Klasická metóda:

Používateľ vyhľadáva informácie prostredníctvom vyhľadávачa. Prejde si niekoľko recenzií, používateľských hodnotení stránky s detailami produktu, porovnávače cien. Každá informácia sa nachádza na samostatnej stránke, čo je veľmi náročné na čas.

Využitím myšlienky nášho systému:

Používateľ zadá kľúčové slová do nášho vyhľadávачa. Systém v reálnom čase kategorizuje stránky a na základe priradených kategórií extrahuje informácie, ktoré sú používateľovi ponúknuté na jednej sumarizačnej stránke. Tieto informácie sú získané zo stránok vyhľadávania v prvom prípade.

<?xml?>

