IntelliWrapper

...skroťte web!

- Strácate sa v neprehľadných webových stránkach?
- Prekáža Vám prítomnosť informácií, ktoré sú Vám nanič?
- Potrebujete získať dáta v spracovateľnej podobe?

Riešením je IntelliWrapper:

IDE pre návrh vlastných obaľovačov

- **Tvorba programu obaľovača**
  IntelliWrapper je prostredím pre vytvorenie programu obaľovača – súboru pokynov a pravidiel (tzv. akcií), ktoré opisujú spôsob získania požadovaných dát z webových stránok.
- **Spúšťanie programu obaľovača**
  Integrálna súčasť aplikácie – Interpreter – umožňuje spustenie programu obaľovača a sledovanie a riadenie priebehu extrakcie. Spustenie je možné v dvoch módoch – štandardnom a ladiacom, ktorý produkuje detailnejšie informácie o aktuálne vykonávaných akciách.
- **Ukladanie/otváranie**
  Vďaka serializácii programu obaľovača je možné uložiť ho na disk a v prípade potreby kedykoľvek otvoriť. Odpadá tak nutnosť vytvárania postupnosti rovnakých akcií znova a znova.

Široké spektrum prístupov k tvorbe obaľovača

- **Viacero spôsobov reprezentácie zdrojových dát**
  IntelliWrapper ponúka viacero prístupov k reprezentácii informácií na webových stránkach. Rôzne reprezentácie umožňujú použitie rozličných mechanizmov na selekciu extrahovaných informácií. Na webové stránky je zatiaľ možné pozerať sa ako na DOM elementov alebo ako na jednoduchý textový reťazec. Ďalšie možnosti sú otvorené.
- **Rôzne stratégie učenia**
  Pri výbere dát, ktoré chceme zo stránok extrahovať, je možné voliť z viacerých stratégií učenia. Každá stratégia obsahuje špecifický spôsob zovšeobecnenia používateľom definovanej vzorky príkladov.

Integrovaný webový prehliadač

- **Pohodlné používateľské prostredie**
  Integrovaný webový prehliadač umožňuje tvorbu programu obaľovača v používateľovi prirodzenom prostredí. Je odbúraná potreba konfigurovať akcie obaľovania manuálne – je zabezpečená plná interaktivita pomocou niekoľkých kliknutí myšou.
- **Jednoduché označovanie príkladov**
  V procese selekcie dát na extrakciu vstavaný prehliadač vizualizuje elementy nachádzajúce sa pod kurzorom myši, čím uľahčuje používateľovi výber relevantných oblastí. Pri prehliadaní je zreteľne odlíšená aktívna časť HTML dokumentu a nie je možné označiť dáta, ktoré sú pre danú etapu extrakcie neprístupné.

Postavené na Mozilla technológii

- **Zobrazovacie jadro GECKO**
  IntelliWrapper využíva najnovšie technológie pre vykresľovanie webového obsahu.
- **JRex komponent**
  Komponent integrovaného prehliadača obsahujúci API pre vnorenie zobrazovacieho jadra do Java aplikácie. Webový prehliadač je súčasťou distribučného balíka IntelliWrapper a nie je nutné používať nástroje tretích strán.

Inteligentná automatizácia výberu dát

- **Zovšeobecňovanie vybratých príkladov**
  Jedným zo základných cieľov aplikácie IntelliWrapper je skrátenie času stráveného tvorbou programu obaľovača. Prostriedkom na jeho dosiahnutie je zavedenie *učenia* na základe príkladov – používateľ zvolí niekoľko vzoriek toho, čo extrahovať, a IntelliWrapper dokončí prácu za neho.
- **Rôzne prístupy k zovšeobecňovaniu príkladov**
  IntelliWrapper dovoľuje použiť viacero stratégií učenia – mechanizmov generalizácie zvolených príkladov. Prítomnosť stratégií zabezpečuje potrebnú flexibilitu v procese obaľovania, keď sú používateľovi ponúknuté viaceré alternatívy orientácie sa v chaotickom a neprehľadnom svete HTML elementov.

Výstupy v rôznych formátoch

- **Podpora niekoľkých výstupných prostredí**
  Široké spektrum možností použitia IntelliWrapper-a sú zabezpečené najmä prítomnosťou zapisovačov rôznych typov. Tie umožnia získané dáta uložiť do XML súborov, relačnej databázy alebo ontológie.

Hoplocampa team