IntelliWrapper

…skrotenie webu

Hoplocampa team
tp05@googlegroups.com

13.6.2007

Obsah prezentácie

- Motivácia a ciele
- Architektúra
- Prezentačná vrstva
- Aplikačná vrstva
  - Akcie
  - Metóda získavania dát
  - Učenie
  - Výstupy
- Zhodnotenie

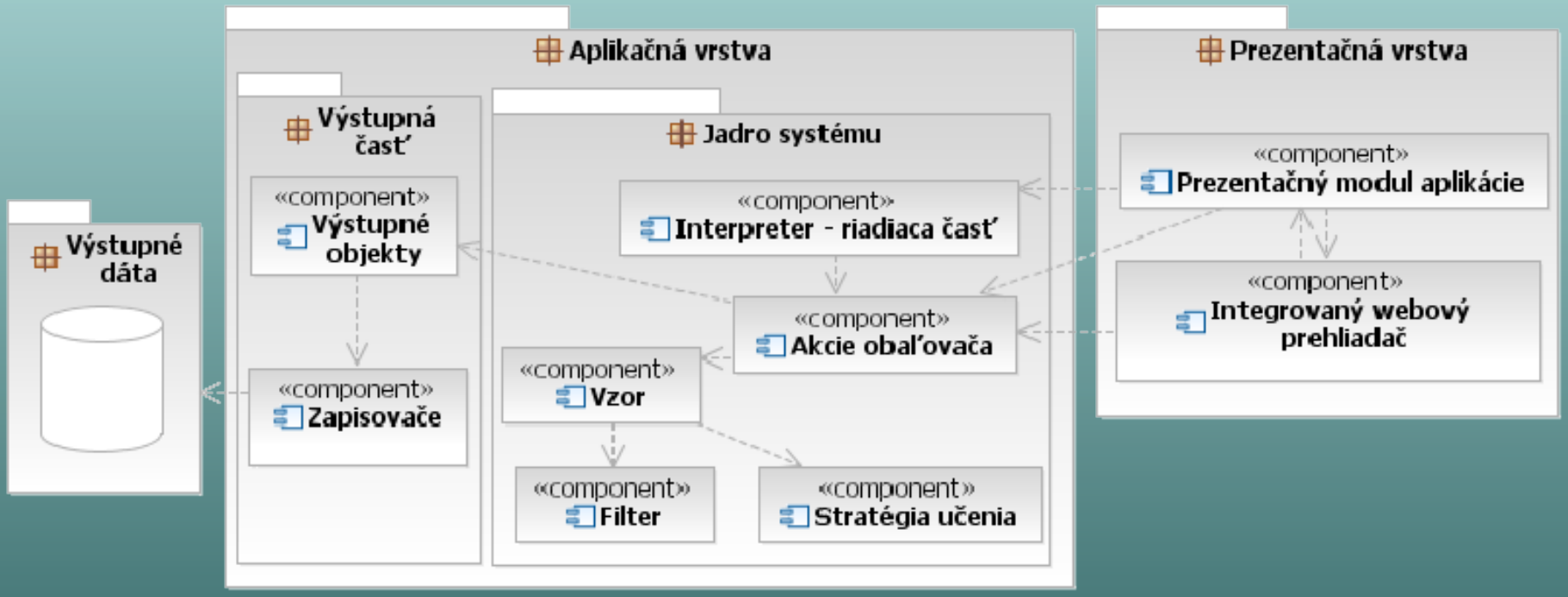
Motivácia

- Výber relevantných informácií z webu
  - Preinformovanosť
  - Dezorientácia
- Eliminácia nežiaduceho obsahu
  - Reklamné pásy
  - Nepotrebný mediálny obsah
- Archivácia dát v rozličných prostrediach
  - Multiplatformnosť

Ciele projektu

- Nadväznosť na existujúci projekt
  - WrapperSuite
- Zjednodušenie procesu tvorby obaľovača
  - Pohodlné používateľské prostredie
    - Integrácia webového prehliadača
  - Skrátenie času tvorby
    - Automatizácia prostredníctvom učenia
- Rozšíriteľnosť

Architektonický návrh



Základné pojmy

- (Sub)dokument
  - Zdroj dát – z čoho extrahovať
- Kontext
  - Obsahuje subdokument a premenné
- Filter
  - Predpis, čo extrahovať
- Stratégia učenia
  - Generovanie filtra z príkladov
- Vzor
  - Zapuzdruje filter, príklady a stratégiu učenia

Používateľské prostredie

- GUI pre tvorbu programu obaľovača
- Definovanie postupnosti akcií
  - Tvorba
  - Mazanie
- Editácia vlastností akcií
- Spúšťanie programu obaľovača
  - Interpreter
- Prístup k webovému prehliadaču

Webový prehliadač

- Interaktivita výberu dát
  - Jednoduché označovanie myšou
  - Pozitívne a negatívne príklady
  - Automatizácia pomocou učenia
- Vizualizácia elementov na stánke
  - Priebežné zobrazovanie „aktívneho“ elementu pod kurzorom
  - Označenie vybratých príkladov rámom
  - Identifikácia aktívnej časti stránky

Prezentačná vrstva – realizácia

- Knižnica Swing
- Hierarchia tried GUI odráža hierarchiu tried akcií
- Volanie dialógov akcií pomocou *reflexie*
- Použitie návrhového vzoru *abstraktná továreň* pre kompozíciu obsahu dialógov
  - PatternDialog

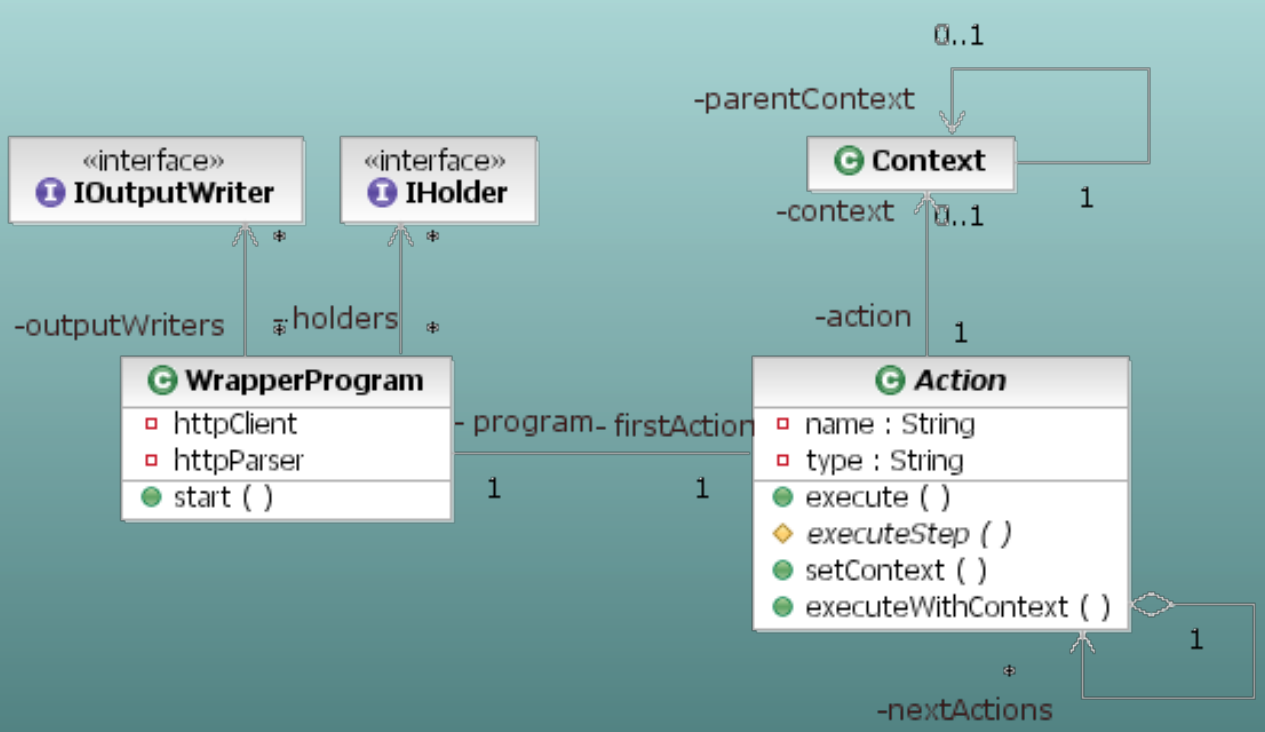
Prezentačná vrstva – realizácia

- JRex
  - Integrovaný prehliadač
  - Komponent webového prehliadača pre Javu
  - Postavené na Mozilla technológii
  - API nad zobrazovacím jadrom GECKO
  - Spracovanie a zobrazovanie DOM dokumentu

Jadro – základné princípy

- Akcie – základná funkcionalita
  - Stromová štruktúra
- Lokálny kontext
  - Princíp dekompozície
- Vzory
  - Lokalizácia extrahovaných dát
- Výstupné objekty

Jadro – diagram tried



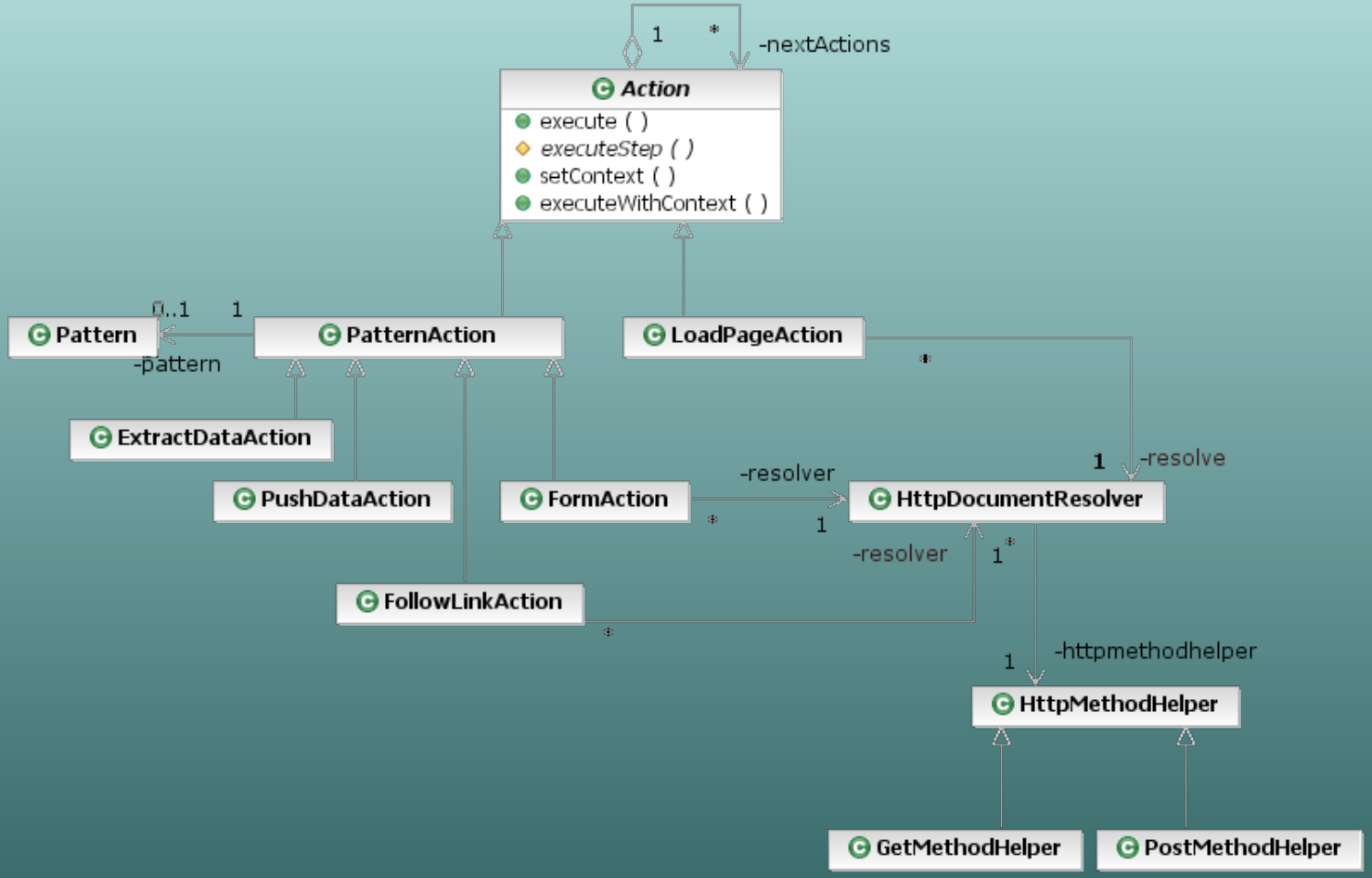
Akcie obaľovača

- Hierarchické usporiadanie v stromovej štruktúre
- Dva typy akcií:
  - Navigačné
  - Extrakčné
- LoadPageAction
  - Načítanie úvodnej stránky
- FollowLinkAction
  - Nasledovanie odkazu („next“ odkazy)

Akcie obaľovača

- ExtractDataAction
  - Extrakcia dát zo subdokumentov
  - Relativizácia kontextov akcií a výstupných objektov
- PushDataAction
  - Zapisovanie dát do nastavených výstupných objektov
- FormAction
  - napĺňanie a odosielanie formulárov

Model akcií systému

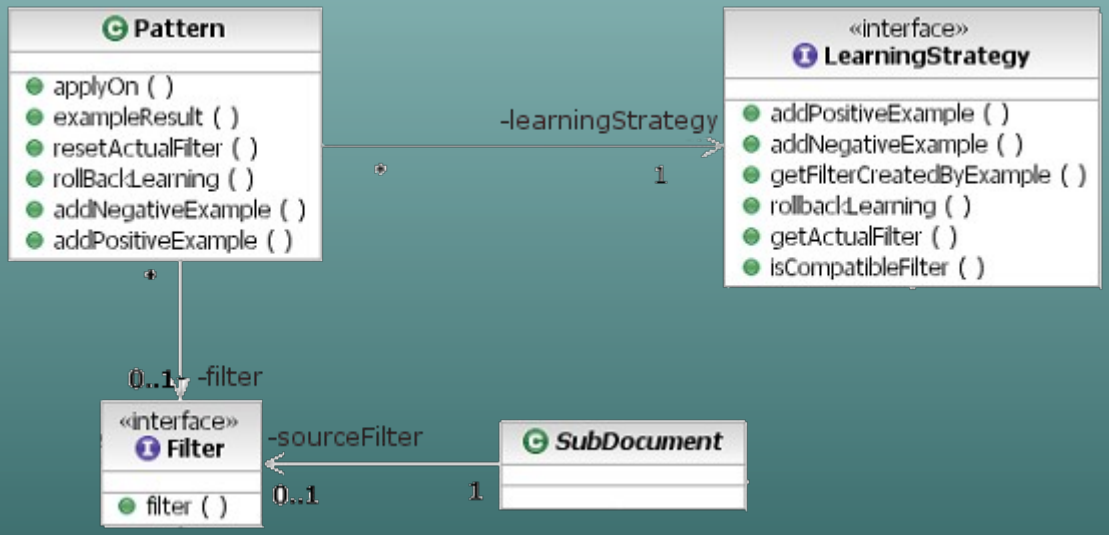


Dokumenty

- Rozličná reprezentácia zdrojových dát
- Umožňujú rozšíreteľnosť aplikácie obaľovača
- Implementované typy:
  - XmlSubdocument
  - StringSubdocument
  - AttributeSelectionSubdocument
- Flexibilita zabezpečená *adaptačným rámcom*

Vzory akcií

- Pozostávajú z dvoch častí:
  - Filter
  - Stratégia učenia



Filter

- Realizácia výberu časti dokumentu
- Typy podľa reprezentácie spracovávania:
  - XPathFilter
  - CompoundXPathFilter
  - AttributeSelectionFilter
- Mechanizmus filtrovania:
  - Vstup: dokument
  - Výstup: zoznam vyfiltrovaných subdokumentov

Učenie

- Automatizácia tvorby obaľovača
- Generovanie filtra na základe (málo) príkladov
  - Pozitívne
  - Negatívne
- Podpora viacerých stratégií
  - Jednoduchá rozšíriteľnosť

Stratégie učenia

- Opakujúca stratégia
  - *Repeating XPath Learning Strategy*
  - spracováva XPath výrazy
  - vyberá naposledy zadaný príklad

Krok	Vstup	Výstup
1.	/HTML/BODY/TABLE[1]/TR[1]/TD[1]	/HTML/BODY/TABLE[1]/TR[1]/TD[1]
2.	/HTML/BODY/TABLE[1]/TR[2]/TD[1]	/HTML/BODY/TABLE[1]/TR[2]/TD[1]

Stratégie učenia

- Jednoduchá stratégia
  - *Simple XPath Learning Strategy*
  - spracováva XPath výrazy
  - jednoduchá generalizácia XPath výrazu
    - odstraňovanie odlišných indexových podmienok
    - odstraňovanie odlišujúcich sa postfixov výrazu

/HTML/BODY/TABLE[1]/TR[1]/TD[2]/
/HTML/BODY/TABLE[1]/TR[2]/TD[2]/
/HTML/BODY/TABLE[1]/TR/TD[2]

Stratégie učenia

- Stratégie výberu na základe atribútov
  - *Attribute Selection Learning Strategy*
  - metóda zhlukovania – k-means algoritmus
  - rozdelenie všetkých elementov na stránke do skupín (zhlukov) na základe podobnosti dostupných atribútov
    - Názov elementu, hodnota Class atribútu, hĺbka v DOM strome, index (medzi rovnomernými súrodencami)
  - pri výbere pozitívneho príkladu sa získajú všetky elementy v rovnakom zhluku

Výstupné objekty

- Udržiavanie extrahovaných dát v stromovej štruktúre
  - následný zápis pomocou zapisovačov
- Typy výstupných objektov:
  - DOM – udržiavanie výstupného dokumentu zapisovanie nových dát
  - Relative – vytvorenie relatívneho výstupného objektu s odkazom na výstupný objekt rodiča, zápis je delegovaný na rodiča

Zhodnotenie

- Zmenili sme jadro aplikácie
- Zaviedli sme
  - Vzory asociované k akciám
  - Pojem lokálneho kontextu
  - Stratégie učenia, ktorými sa vytvárajú filtre
- Používame integrovaný webový prehliadač
- Aplikácia je jednoducho rozšíriteľná o ďalšiu funkcionalitu

Možnosti ďalšej práce

- Rozšírenie používateľské prostredie aplikácie
- Zlepšenie výkonnosti integrovaného prehliadača
- Doplnenie stratégii učenia, ktoré by pracovali aj s inou reprezentáciou dokumentu
- Implementácia zapisovačov do nových výstupných prostredí

Ďakujeme za pozornosť!

Lúči sa Hoplocampa team...