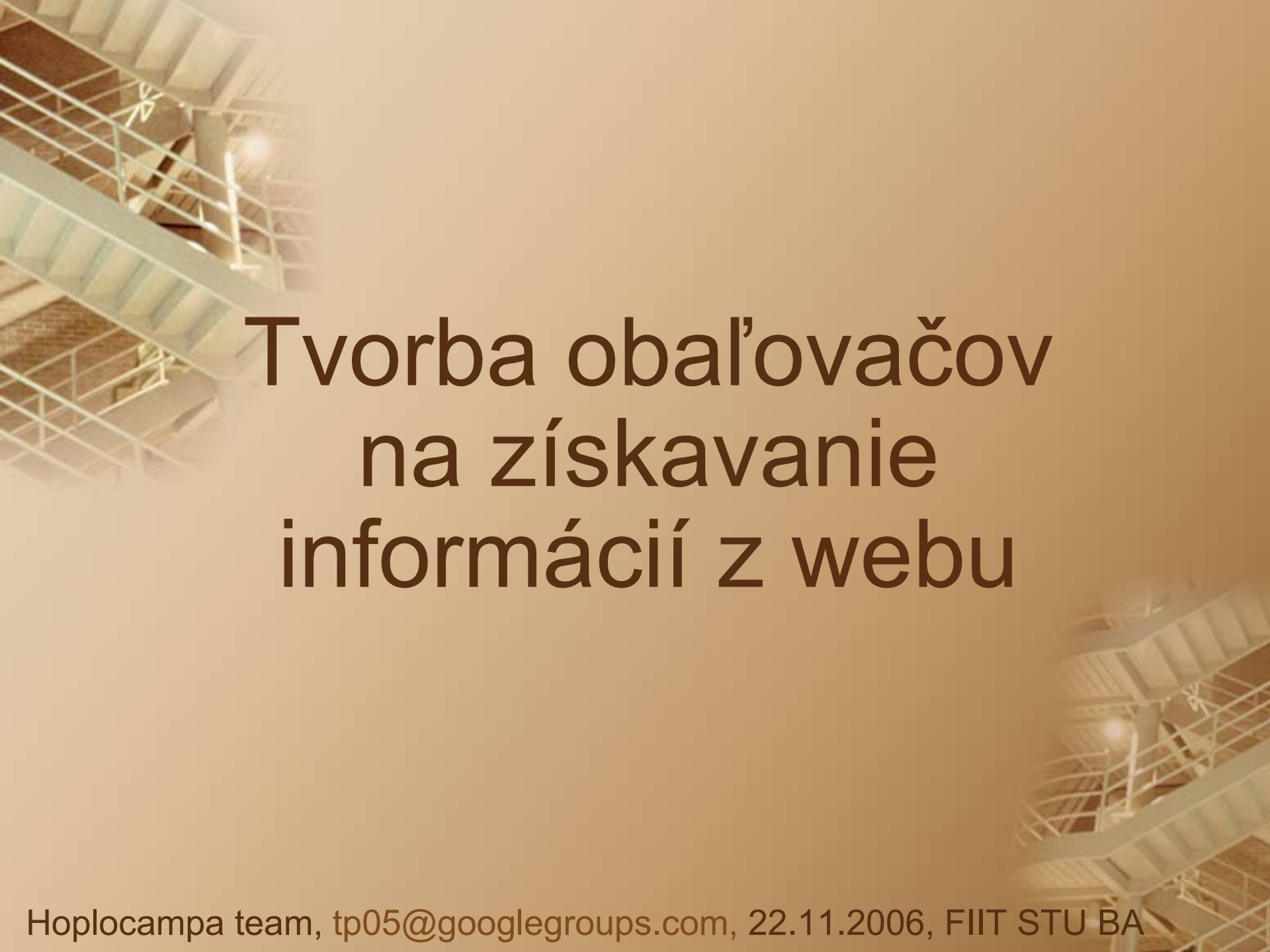




Tvorba obal'ovačov na získavanie informácií z webu

Obsah

- Úvod do obal'ovačov
- Existujúce nástroje
- Náš nástroj
 - Špecifikácia
 - Architektúra
(+ použité technológie)
- Zhodnotenie

Úvod

- web ~ kvantum informácií, preinformovanosť
- získavanie relevantných dát:
 - personalizácia
(adaptívne hypermédiá)
 - obalovač
(angl. wrapper)
- prípady použitia

Čo je obal'ovač

obal'ovač (web stránok) - program, ktorý vykonáva automatizovaný výber relevantných informácií z neštruktúrovaného kódu jazyka HTML zameraného najmä na prezentáciu obsahu webu pre človeka

Zadanie projektu

- vytvorenie architektúry rozšíriteľnej o viaceré typy reprezentácií dokumentov a rôzne typy vzorov
- rozšírenie existujúceho aplikačného rámca o:
 - možnosť tvorby obalovača pomocou prehliadača
 - návrh a realizáciu učenia obalovača

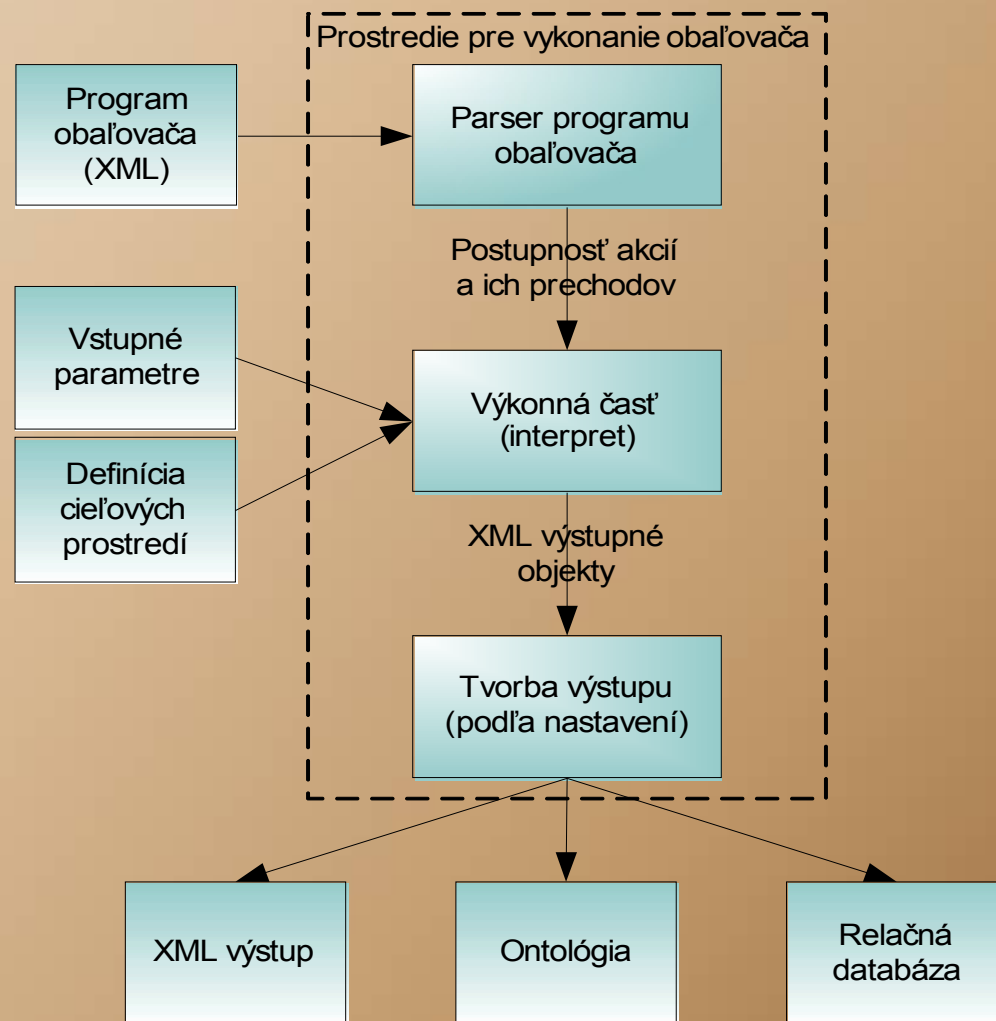
Existujúce nástroje

		
<i>Spôsob tvorby</i>	- RoboMaker - definovanie krokov robota - ModelMaker - modelovanie vstupných a výstupných dát	- vytvorenie stromu aplikácií vzorov na dokument - (úprava výstupu – XML Tool)
<i>Spôsob reprezentácie obaľovača</i>	konečný automat	program Elog
<i>Podpora učenia</i>	nie	nie

Existujúce nástroje

		
<i>Typy akcií</i>	Navigačné, extrakčné, iteračné (ForEachTag), vyplňanie formulárov	extrakčné, navigačné, použitie rekurzie
<i>Spôsob selekcie subdokumentu</i>	- Tag Finders – určenie dokumentu - filter (XPath, regulárny výraz)	filtre – XPath, regulárny výraz
<i>Ošetrenie chýb</i>	- obsluha chyby - report s ukončením - ošetrenie v predchádzajúcom kroku - ignorovanie	- priame nie je - verifikácia

Existujúce nástroje WrapperSuite



Existujúce nástroje

WrapperSuite

- **WrapperDesigner** – grafické rozhranie (GUI) pre vytváranie obalovačov s podporou interpretácie a ladenia
- **Wrap** – nástroj pre interpretáciu obalovačov z príkazového riadku

Špecifikácia

Ciele projektu

- Zameranie sa na používateľa
- Systém pomáha pri tvorbe obalovačov
- Zjednodušenie jazyka alebo niekedy je menej viac
- Jednoduchá reakcia na zmeny dokumentov

Špecifikácia

Použitie nástroja

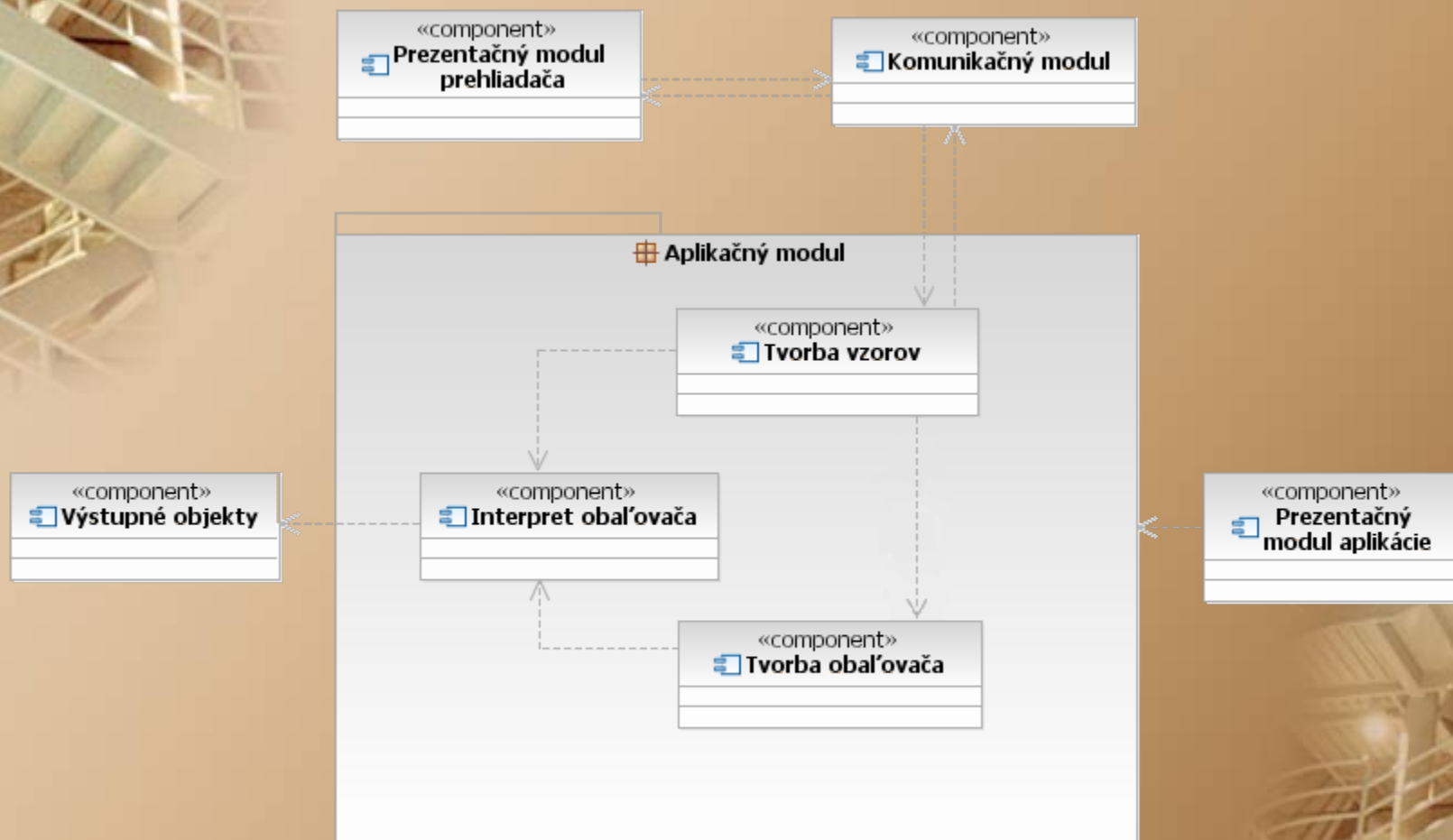
- Lokálny kontext akcií
- Práca nad subdokumentom
- Výstupné objekty
- Vzory, filtre, príklady
- Stratégie učenia

Špecifikácia

Použitie nástroja

- Ako to všetko spojiť?
- Integrácia s existujúcou aplikáciou
 - vytvorené rozhrania
 - jazyk obalovača
 - spracovávanie chýb
 - výstupné zapisovače

Architektúra



Tvorba obaľovača

modul Tvorba obaľovača

- Vstupy a výstupy modulu
- Vlastnosti obaľovača
 - Akcie
 - Relácie medzi akciami
 - Vstupné premenné
 - Štartovacia akcia
 - HTTP klient a parser
- Proces tvorby

Akcie obalovača

modul Tvorba obalovača

- Definícia akcií
 - Typ akcie
 - Atribúty akcie
- Navigačné akcie
- Extrakčné akcie
- Odoslanie formulára

Učenie

modul Tvorba vzorov

- Učenie s učiteľom
 - zadávanie príkladov
 - interakcia s používateľom
 - netreba ručné zadávanie vzorov
 - generalizácia
 - dodržať podmienky príkladov
 - zahrnúť všetky pozitívne príklady
 - nezahrnúť žiadne negatívne príklady

Učenie

modul Tvorba vzorov

- Základný prístup
 - prístup k príkladu ako k reťazcu (s ohľadom na XPath)
- Metóda s použitím zložených filtrov
 - jednoduché filtre
 - redukcia filtrov
 - tvorba zloženého filtra

$$\bullet \quad C = C_1 \wedge C_2 \wedge (\neg C_3)$$

- Prezentačný modul aplikácie

- GUI rozhranie medzi používateľom a obal'ovačom
- Úlohy modulu
 - distribúcia údajov pre tvorbu a modifikáciu obal'ovačov
 - riadenie obal'ovača
 - zobrazenie získaných údajov z webových stránok

Prezentačný modul prehliadača

- Rozhranie medzi používateľom a prehliadačom Mozilla Firefox
- Mozilla Firefox rozšírenie
 - technológia XUL, JavaScript
- Úlohy modulu
 - distribúcia údajov získaných od používateľa
 - adresa koreňového dokumentu (URL)
 - pozitívne a negatívne príklady pri procese učenia vzorov
 - zobrazenie subdokumentu

Komunikačný modul

- Použitie technológie XPCOM
 - multiplatformový komponentový objektový model
 - úzko spojený so zobrazovacím jadrom Gecko
- Komunikácia pomocou rozhraní:
 - prenášanie subdokumentov
 - prenášanie filtrov
 - ...

Sumár

čo sme zatiaľ urobili

- Analyzovali problémovú oblasť
 - existujúce riešenia
 - minuloročný aplikačný rámec
- Navrhli zmeny a rozšírenia, ktoré
 - z existujúceho riešenia vytvorila ľahšie rozšíriteľný aplikačný rámec
 - poskytnú používateľom požadovanú mieru interakcie
 - pridajú obalovaču prvky umelej inteligencie

Plán

čo ešte len urobíme

- Tvorba prototypu:
 - kľúčové úpravy jadra existujúceho aplikačného rámca
 - rozhranie dočasne nahradzujúce rozšírenie prehliadača
 - implementácia vybraných vzorov
 - implementácia jednoduchšej stratégie učenia sa