Crawler.NET: Elosztott webrobotok koordinálása és vezérlése

Hunyadi Levente

2007. június 19.

Crawler.NET
A component-based distributed framework for web traversal

Bevezetés
Motiváció
Célok
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés

Bevezetés

Motiváció

Bevezetés
Motiváció
Célok
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés

- a web: szórt formában jelenlévő információ
- gyorsan változó tartalom
- friss indexadatbázishoz a web gyakori bejárása szükséges

Motiváció

Bevezetés
Motiváció
Célok
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés

Központosított

+ átlátható modell
+ könnyű vezérlés
− szűk keresztmetszetek
− gyenge hibatűrés

Elosztott

+ csökkenő távolság a bejárandó tartalomtól
+ kisebb hálózati terhelés
+ skálázható
− összetettebb együttműködést igényel

Crawler.NET
A component-based distributed framework for web traversal

Tervezési célok

Bevezetés
Motiváció
Célok
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés

- kihasználni az elosztott rendszerből fakadó előnyöket
- elrejteni a komplexitást

Bevezetés

Architektúra

Felépítés

Építőkövek

Laza csatolás

Építőkövek

Referencia-megvalósítás

Értékelés

Architektúra

A rendszer felépítése

Bevezetés
Architektúra
Felépítés
Építőkövek
Laza csatolás
Építőkövek
Referencia-megvalósítás
Értékelés

Kétrétegű:

Keretrendszer
Általános feladatok

- kommunikáció
- illesztés
- konfigurálás
- életciklus-kezelés
- felügyelet

Ráépülő alkalmazás
Konkrét feladatok

- dokumentumok letöltése
- hivatkozások kinyerése
- hivatkozáskapcsolatok nyilvántartása

A keretrendszer építőkövei

Bevezetés
Architektúra
Felépítés
Építőkövek
Laza csatolás
Építőkövek
Referencia-megvalósítás
Értékelés

- **Komponensek**: a rendszer műveletvégző egységei
- **Illesztők**: üzenetalapú kommunikációt tesznek lehetővé
- **Szolgáltatók**: az erőforrásokhoz történő hozzáférést szabályozzák

$\Longrightarrow$ az építőkövek között laza csatolás, kötés futási időben

Laza csatolás

Bevezetés
Architektúra
Felépítés
Építőkövek
Laza csatolás
Építőkövek
Referencia-megvalósítás
Értékelés

Előnyök:

- deklaratív szemlélet:
  rendszer = komponensek + szolgáltatók + összekapcsolási mód + paraméterezés
- nincs kötött elemkészlet, könnyen kiterjeszthető
- egyszerűbb és gyorsabb fejlesztés, csak funkcionális aspektusok

Bevezetés

Architektúra

Építőkövek

Illesztők

Komponensek

Szolgáltatók

Referencia-megvalósítás

Értékelés

Építőkövek

Illesztők

Bevezetés
Architektúra
Építőkövek
Illesztők
Komponensek
Szolgáltatók
Referencia-megvalósítás
Értékelés

- a laza csatolást biztosítják
- típusos (FIFO) sorok absztrakciói, üzeneteket továbbítanak
- bemeneti és kimeneti illesztők, adatáramlás:
  - bemeneti illesztő $\longrightarrow$ fogyasztó komponens
  - termelő komponens $\longrightarrow$ kimeneti illesztő
- több-több kapcsolat a komponensek és az illesztők között $\Longrightarrow$ azonosítás szerepek segítségével

Illesztők megvalósítása

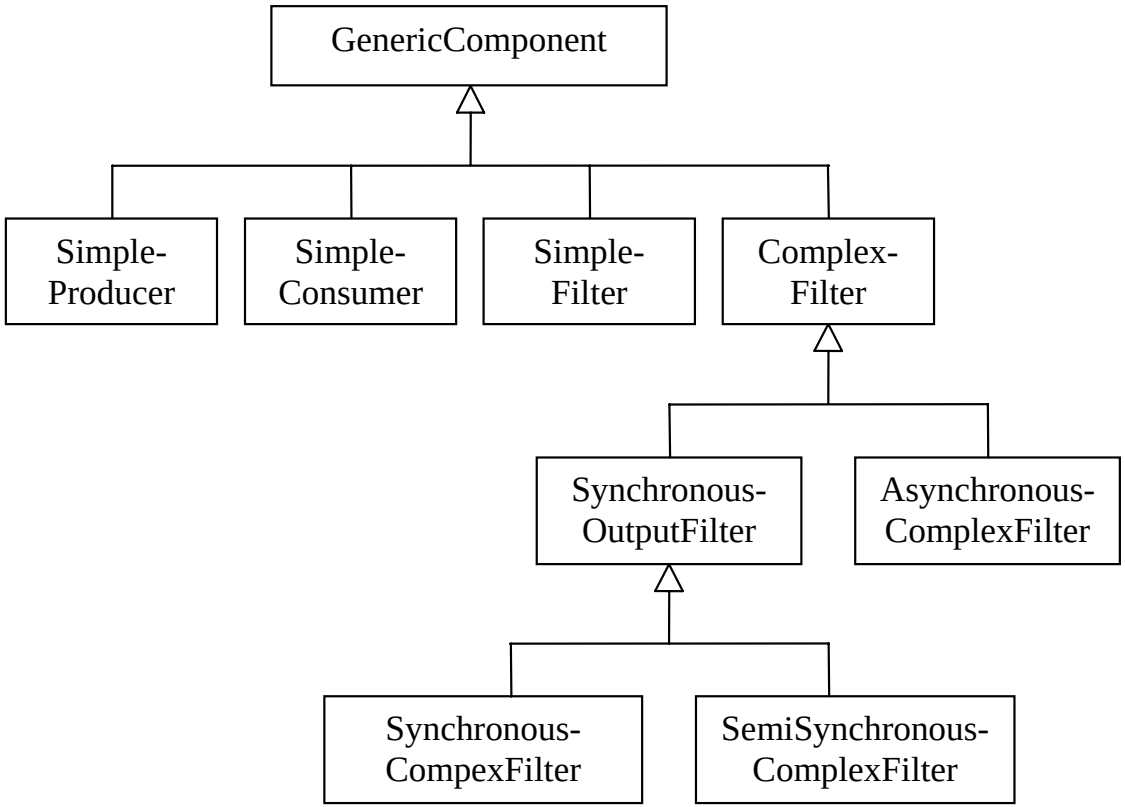
Bevezetés
Architektúra
Építőkövek
Illesztők
Komponensek
Szolgáltatók
Referencia-megvalósítás
Értékelés

- **Helyi illesztő**
  adatáramlás *referencia szerinti átadással*
- **Távoli illesztő**
  adatáramlás *sorosítással* (TCP protokollon keresztül)

Komponensek

Bevezetés
Architektúra
Építőkövek
Illesztők
Komponensek
Szolgáltatók
Referencia-megvalósítás
Értékelés

Általános viselkedés az ősosztályban, specifikus viselkedés származtatással



Komponensek

Bevezetés
Architektúra
Építőkövek
Illesztők
Komponensek
Szolgáltatók
Referencia-megvalósítás
Értékelés

Alapszolgáltatások:

- példányosítás és kötés XML leíró alapján
- beállítások módosítása külső (grafikus) felületről
- adatáramlás-vezérlés
- életciklus-kezelés: leállítás és újraindítás

Szolgáltatók

Bevezetés
Architektúra
Építőkövek
Illesztők
Komponensek
Szolgáltatók
Referencia-megvalósítás
Értékelés

- feladatuk: komponensek közötti adatmegosztás
- csomagolók: szabályozzák az erőforrásokhoz való hozzáférés
- fajtái:
  - service provider: folyamat-szintű erőforrások
  - data provider: hátterében állomány vagy adatbázis
- többszintű gyorsítótárazási mechanizmus *átlátszó* beépítése

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Felépítés
Vezénylés
Grafikus felület
Értékelés

Referenciamegvalósítás

Crawler.NET
A component-based distributed framework for web traversal

Felépítés

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Felépítés
Vezénylés
Grafikus felület
Értékelés

Mester–szolga (master–worker) architektúra:

- a mester particionálja a webet és az egyes részleteinek bejárását egy-egy szolgának utalja ki
- a szolga bejárja a web rábízott szeletét, a kifelé mutató hivatkozásokat visszaküldi a mesternek

Megvalósítás a keretrendszer alaposztályainak segítségével $\Longrightarrow$ elrejtett elosztott megvalósítás

Felépítés

Bevezetés

Architektúra

Építőkövek

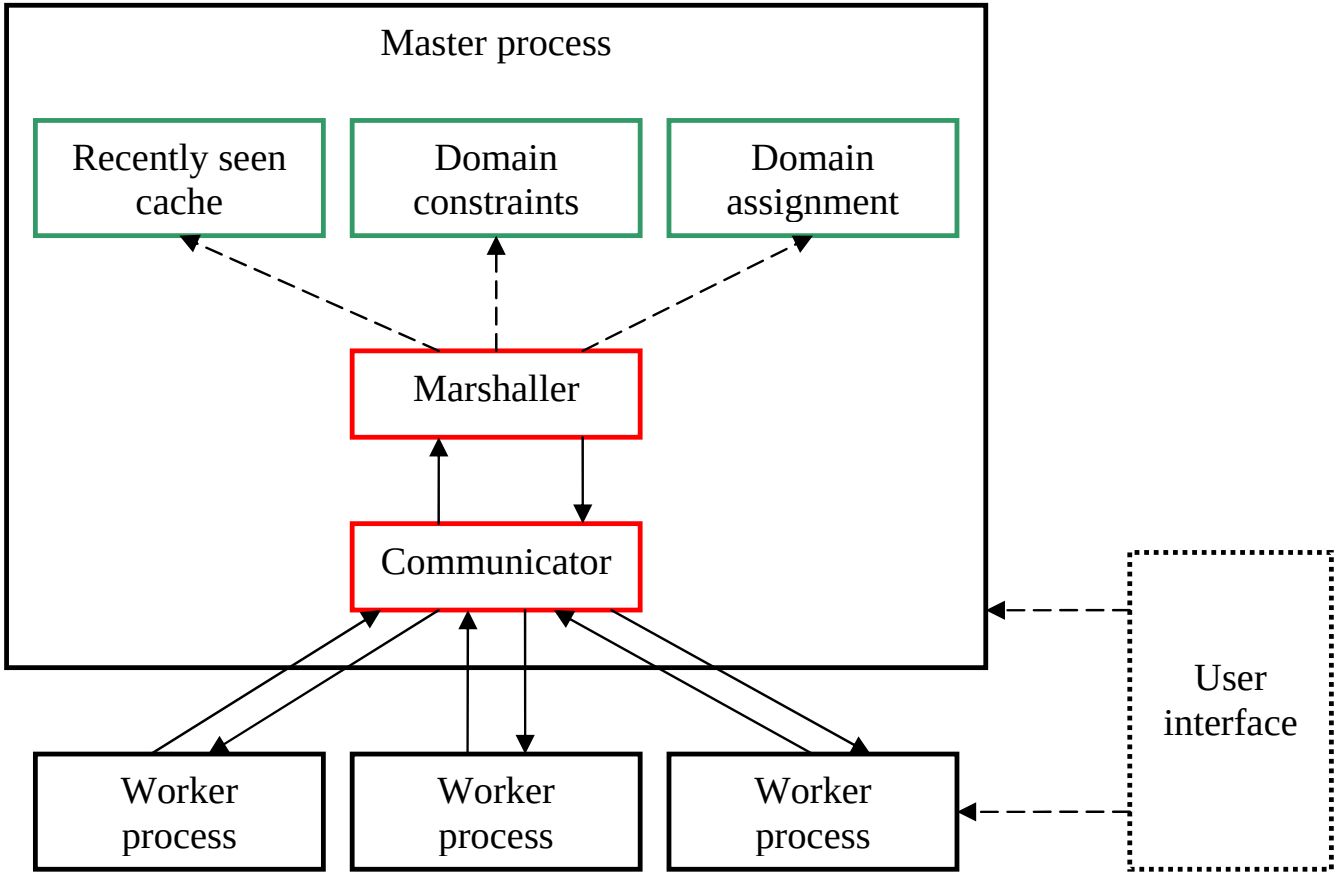
Referencia-megvalósítás

Felépítés

Vezénylés

Grafikus felület

Értékelés



Crawler.NET
A component-based distributed framework for web traversal

Vezénylő komponens

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Felépítés
Vezénylés
Grafikus felület
Értékelés

- a mester legfontosabb alkotóeleme
- a beérkező URL-eket reguláris kifejezésekhez hasonló szabályok alapján szűri
- a szűrt URL-eket tartomány alapján továbbítja
- szükség esetén kijelöl új felelős szolgát
- tárolja a nemrég továbbított URL-eket, kiszűrve a gyakori ismétlődéseket

Vezénylő komponens

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Felépítés
Vezénylés
Grafikus felület
Értékelés

Korlátozott kicserélt adatmennyiség:

- *lokalitási elv*: a hivatkozások kb. 10%-a mutat csak hoszton kívülre
- *kötegelt átvitel*: az ismeretlen hoszthoz tartozó URL-eket a szolga csoportosan küldi át a mesternek és fordítva
- nemrég látott URL-ek automatikus eldobása

A mestert és a szolgákat tipikusan egy-egy gépcsoport realizálja.

Grafikus felület

Bevezetés

Architektúra

Építőkövek

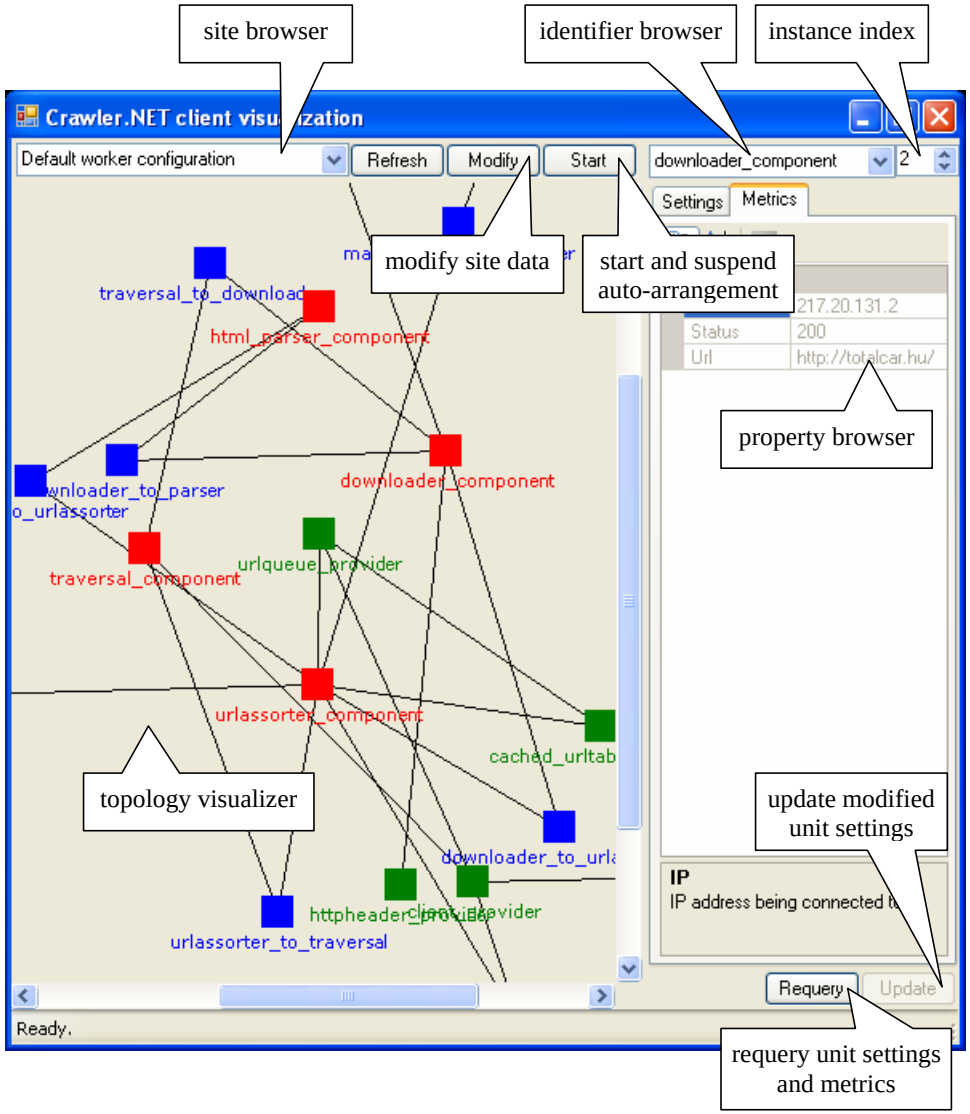
Referencia-megvalósítás

Felépítés

Vezénylés

Grafikus felület

Értékelés



Crawler.NET
A component-based distributed framework for web traversal

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés
Összefoglalás

Értékelés

Összefoglalás

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés
Összefoglalás

- keretrendszer az általános feladatokra
- laza csatolású, feladatspecifikus komponensek
- nyitott, kiterjeszthető, skálázható architektúra
- deklaratív konfigurálhatóság, testreszabhatóság
- átlátszó gyorsítótárazási mechanizmusok támogatása
- grafikus felület

A megvalósított rendszer forráskódja elérhető a SourceForge.net-en

Crawler.NET
A component-based distributed framework for web traversal

Bevezetés
Architektúra
Építőkövek
Referencia-megvalósítás
Értékelés
Összefoglalás

Köszönöm a figyelmet!

Crawler.NET
A component-based distributed framework for web traversal