Crawler.NET: Komponensalapú elosztott keretrendszer a web bejárására

Hunyadi Levente és Pallos Péter

2006. november 17.

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

Motiváció

- az internet mérete rohamosan nő
- a web: szórt formában jelen lévő információ
- keresőrendszerek felértékelődése
- indexadatbázisok építéséhez webet bejáró alkalmazások

Motiváció

- friss indexadatbázis

⇓

- hatékony bejárás

⇓

- párhuzamosítás

⇓

- elosztott rendszer

⇓

- lényegesen nagyobb komplexitás

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

Tervezési célok

- skálázhatóság
- könnyű konfigurálhatóság
- átlátszó kommunikáció
- robusztusság, hibatűrés

A rendszer architektúrája

A célok megvalósítását két elkülönülő réteg biztosítja.

Keretrendszer

Általános feladatok

- kommunikáció
- életciklus-kezelés
- konfigurálás

Ráépülő alkalmazás

Konkrét feladatok

- dokumentumok letöltése
- hivatkozások kinyerése
- köztük lévő kapcsolatok nyilvántartása

Az architektúra tulajdonságai

- a keretrendszer szabványos elemeket tartalmaz, amelyek megvalósítják az általános viselkedést
- a feladat-specifikus elemeket származtatással képezzük
- az elemek között a keretrendszer laza csatolást biztosít

Előnyök:

\+ egyszerűbb és gyorsabb fejlesztés

\+ új funkcionalitással való bővíthetőség

Rendszerelem-típusok

- **Komponensek**
  a rendszer műveletvégző egységei, egy osztályt és hozzá kapcsolódó egy vagy több szálat jelentenek
- **Illesztők**
  aszinkron, üzenetalapú kommunikációt tesznek lehetővé komponensek között a folyamaton belül vagy két távoli folyamat között
- **Szolgáltatók**
  az erőforrásokhoz történő szinkronizált hozzáférést szabályozzák

Illesztők

- típusos sorok absztrakciói, üzenetsort reprezentálnak
- bemeneti és kimeneti illesztők
  - bemeneti illesztő $\longrightarrow$ fogyasztó komponens
  - termelő komponens $\longrightarrow$ kimeneti illesztő
- több-több kapcsolat a komponensek és az illesztők között; azonosítás szerepek segítségével

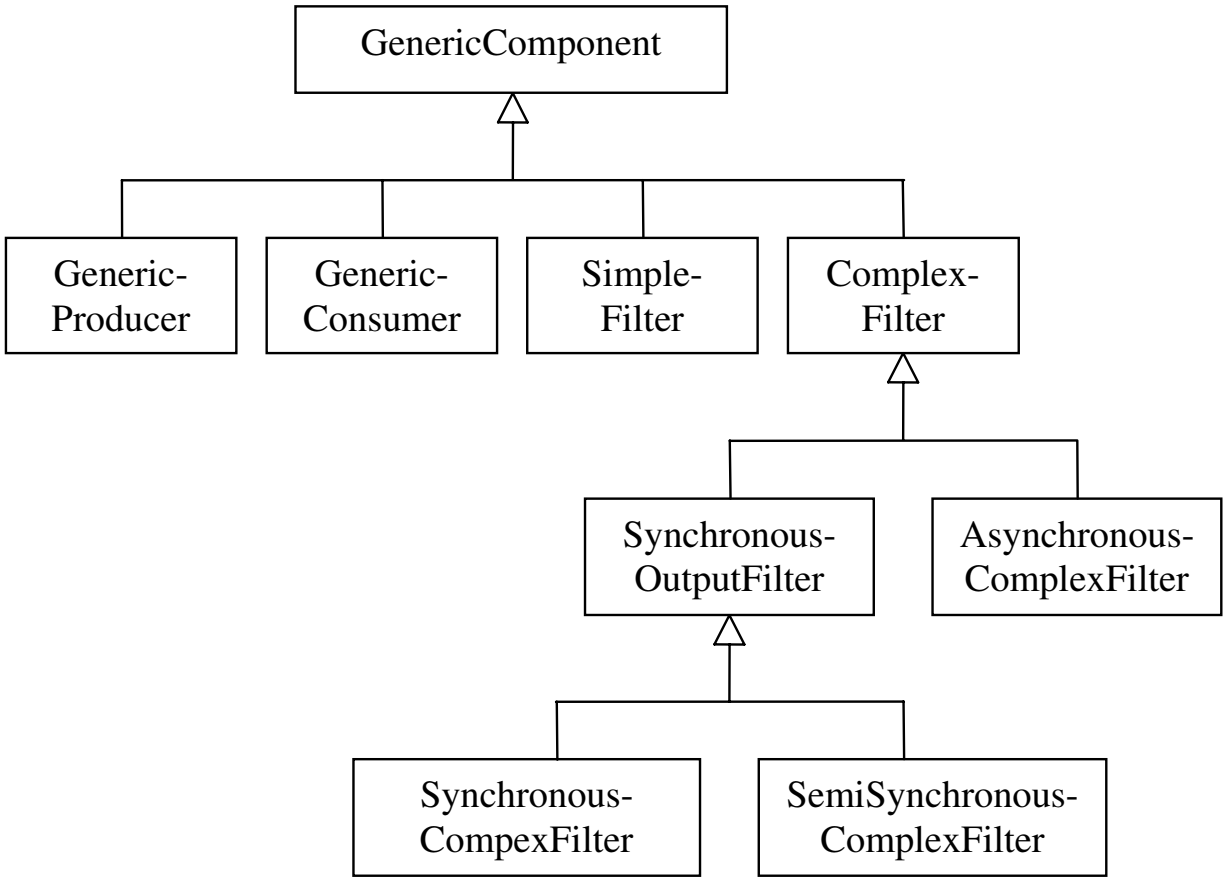
Illesztők megvalósítása

A komponensek számára átlátszó az illesztők típusa:

- **Helyi illesztő**
  típusos FIFO sor; az adatáramlás *referencia szerinti átadással*
- **Távoli illesztő**
  két külön folyamatban lévő sornak és az azokat összekapcsoló hálózati kommunikációs komponenseknek felel meg; az adatáramlás *sorosítással* (TCP protokollon keresztül)

Komponensek

A sorok kezelését a komponensek ősosztálya takarja el.



Szolgáltatók

- komponensek által közösen használt erőforrásokhoz való hozzáférés becsomagolása
- szinkronizált hozzáférés az adatforrásokhoz
- többszintű gyorsítótárazási mechanizmus *átlátszó* beépítése

Felépítés

Kliens–szerver architektúra:

- a szerver particionálja a webet és az egyes részleteinek bejárását egy-egy kliensnek utalja ki
- a kliens bejárja a web rábízott szeletét, a kifelé mutató hivatkozásokat visszaküldi a szervernek

Megvalósítás a keretrendszer alaposztályainak segítségével

Vezénylő szerver komponens

- a beérkező URL-eket tartomány, illetve hoszt alapján a felelős kliensek felé továbbítja
- amennyiben nincs felelős kliens, kijelöl egyet
- tárolja a nemrég továbbított URL-eket, kiszűrve a gyakori ismétlődéseket

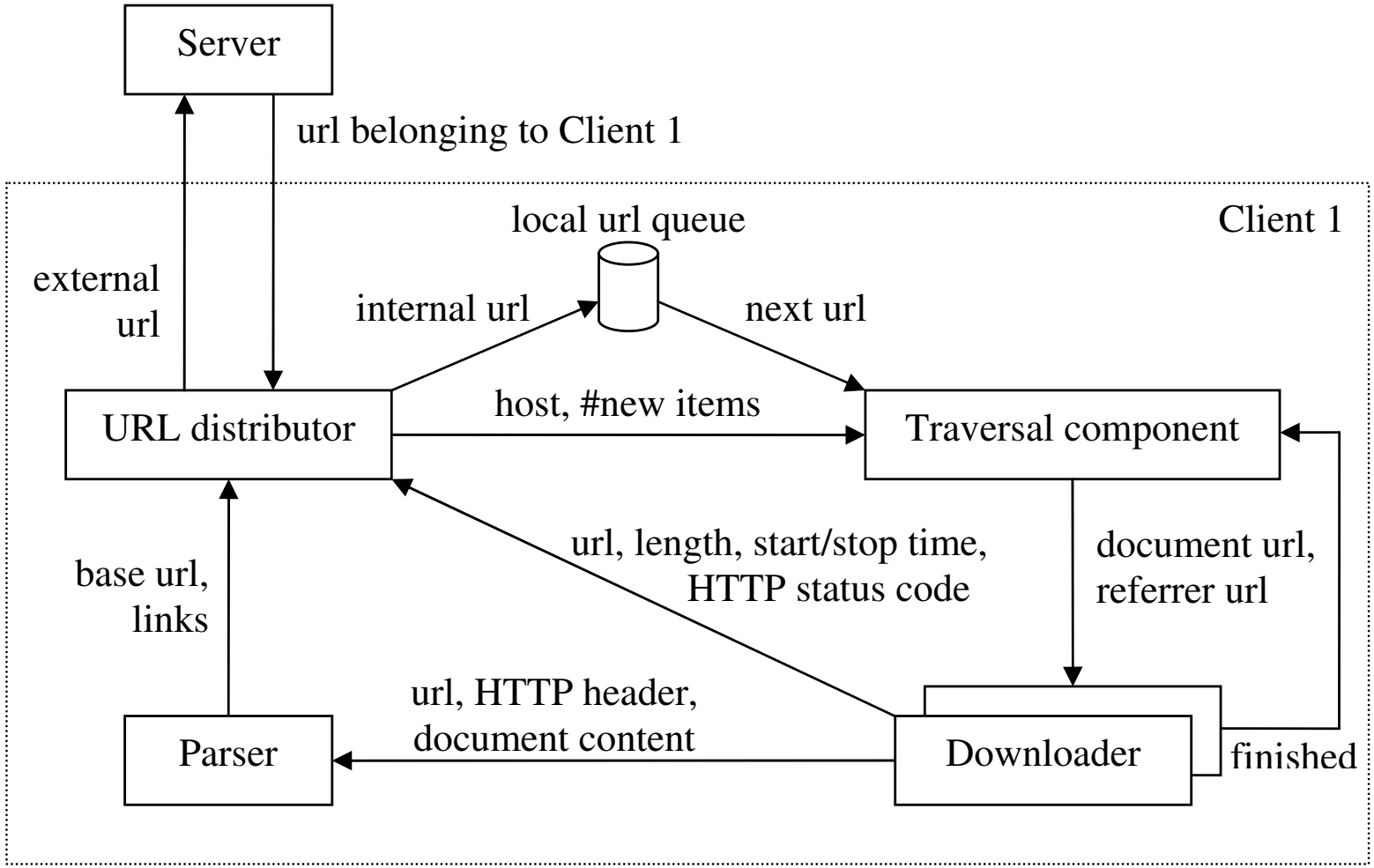
Vezénylő szerver komponens

A vezénylés során a kicserélt adatmennyiség korlátozott:

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

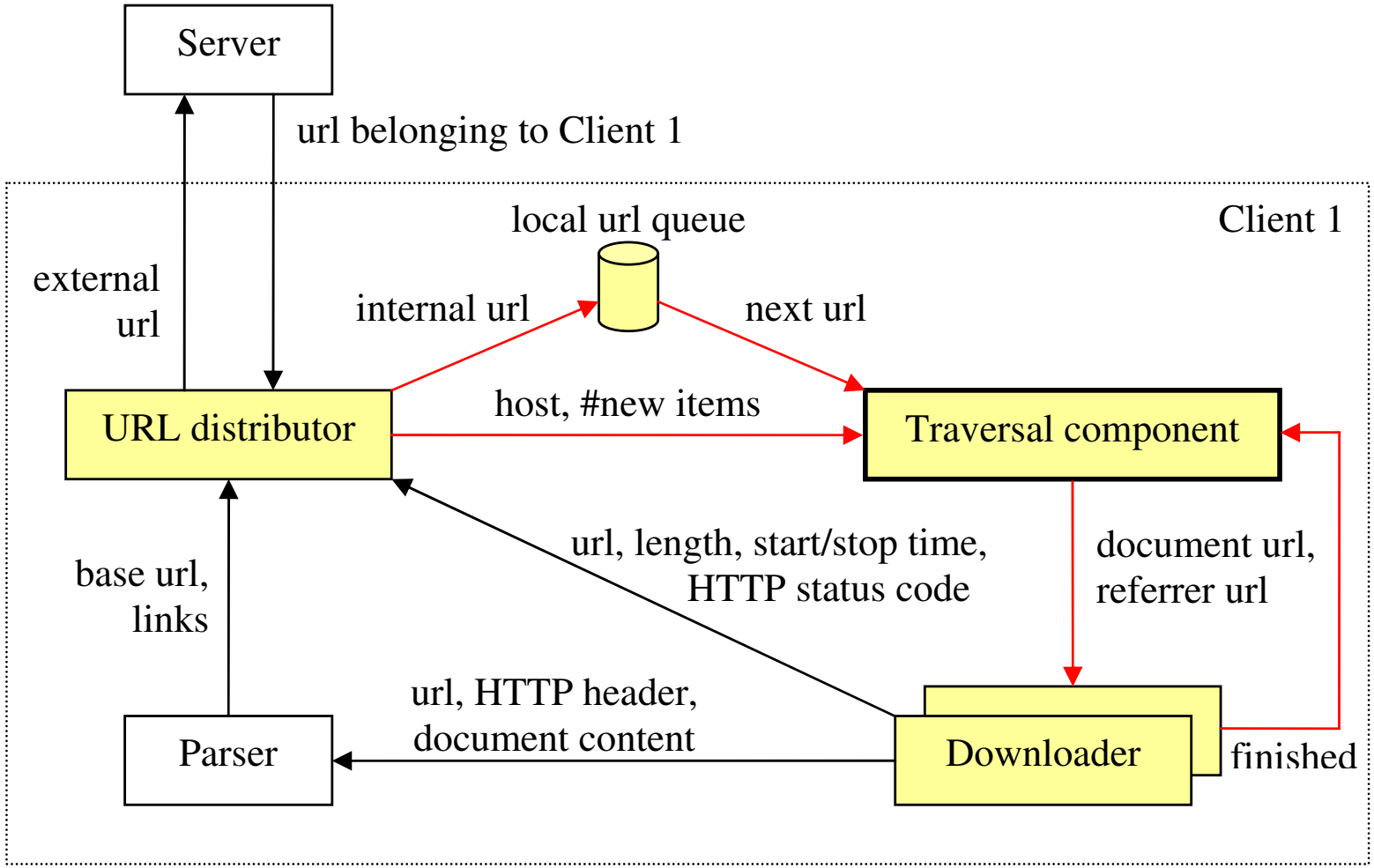
Terheléselosztás vezénylőpéldányok között: URL-ek szétosztása az URL hoszt név hash értéke alapján

Egy kliens alapkiépítése



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

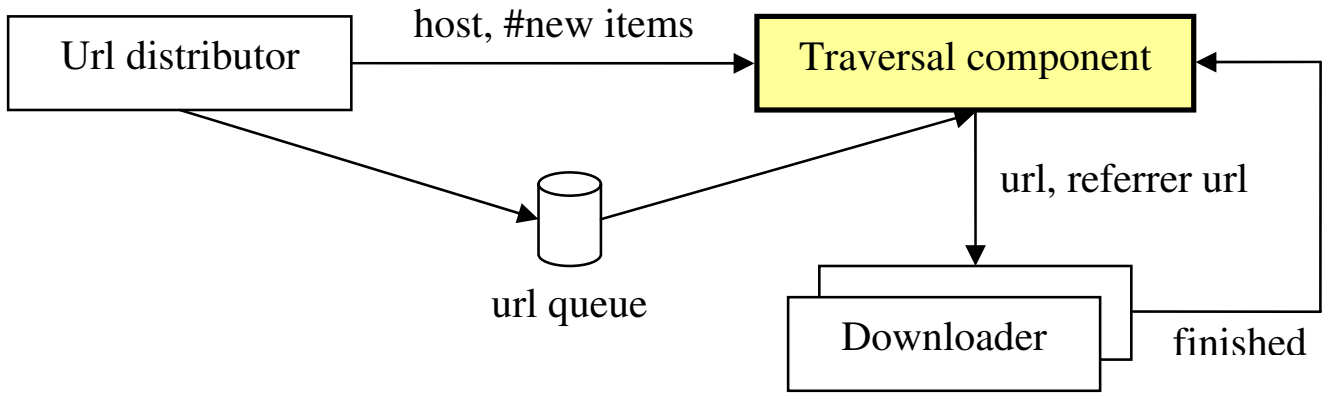
Bejáró komponens



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

Bejáró komponens

- letöltésre váró URL-ek tárolása *perzisztens tár*ban
- *értesítés* új URL-ek érkezéséről vagy egy kiszolgáló felszabadulásáról
- *szélességi* vagy *relevancia alapú bejárás* alapján a következő dokumentum kiválasztása

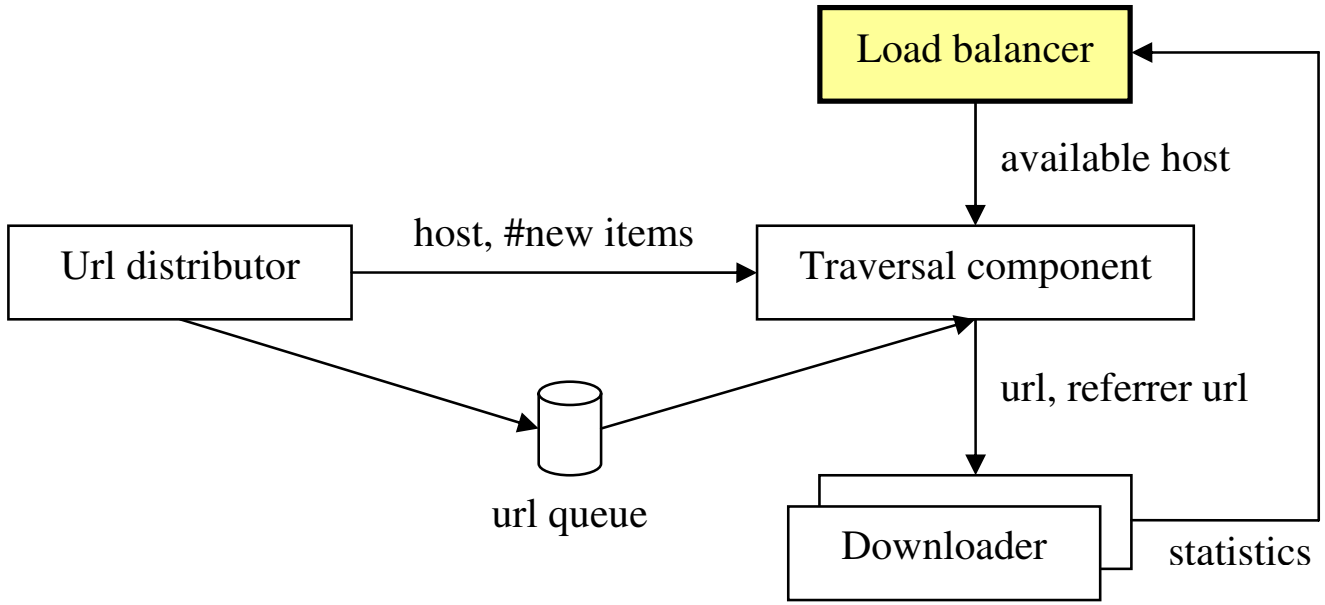


Crawler.NET
A component-based distributed framework for web traversal

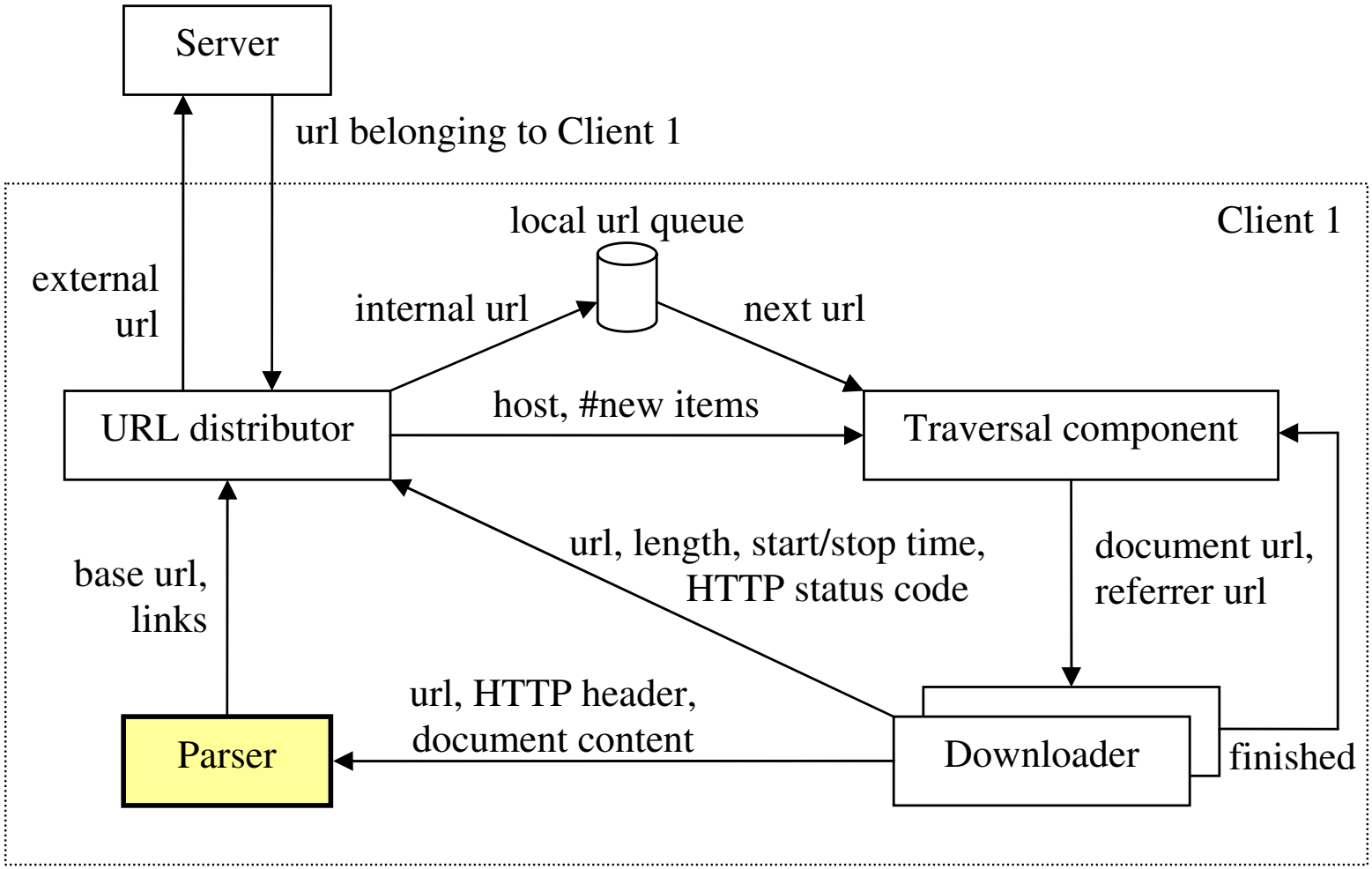
Terheléselosztó komponens

- megakadályozza a webkiszolgálók túlterhelését
- szoros együttműködés a bejáró komponenssel
- kézzel konfigurálható a kérések gyakorisága globálisan és hosztonként is
- dinamikusan változtatható a kérések időköze a visszajelzésként kapott válaszidő és sebesség alapján

Terheléselosztó komponens



Dokumentumelemző komponens



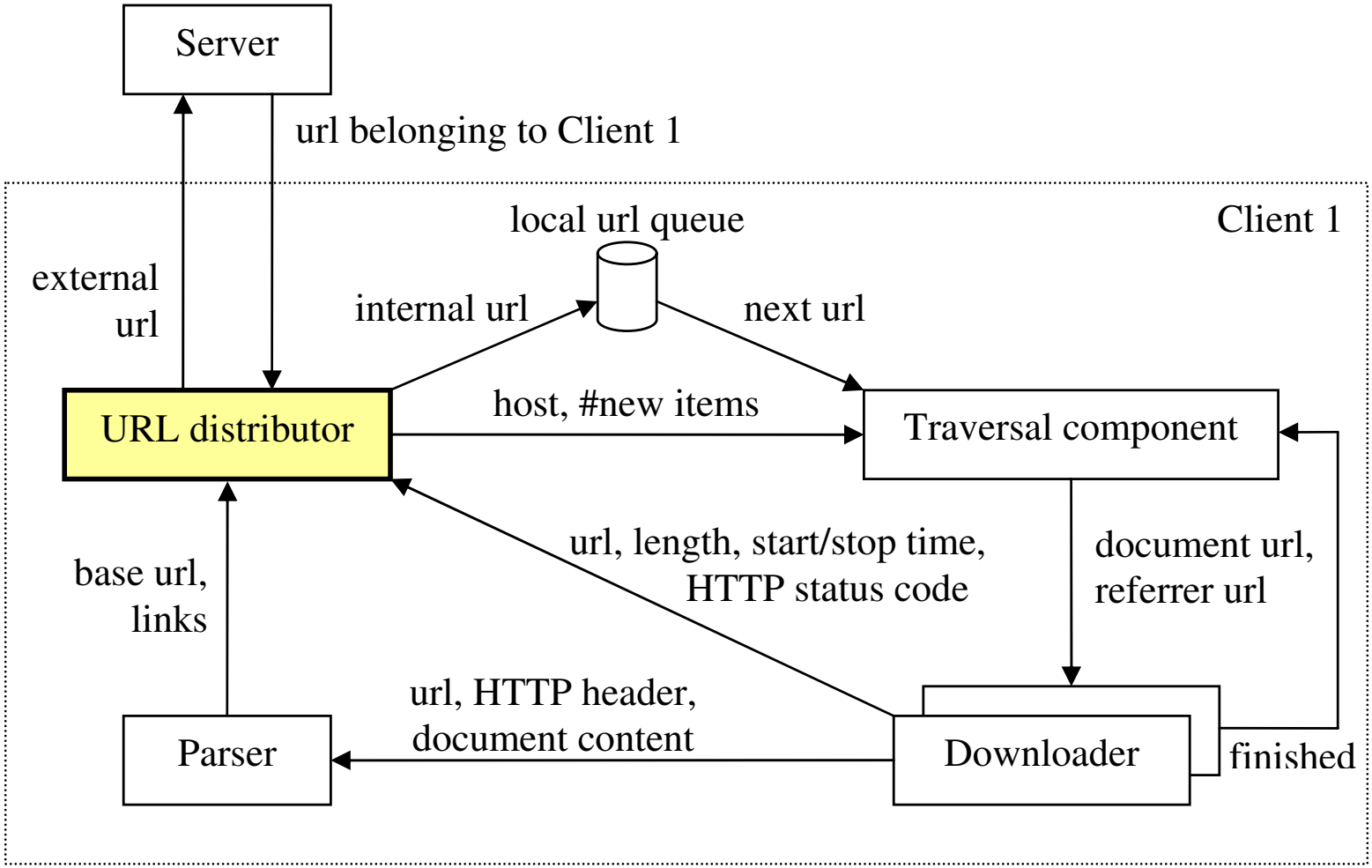
Crawler.NET
A component-based distributed framework for web traversal

Dokumentumelemző komponens

Elsődleges feladata: hivatkozások kinyerése a dokumentumokból.

- nem szabványos HTML állományok
- hivatkozások sok különféle formában
- teljesítménybeli megfontolások (pl. szkriptek értelmezése)
- sokféle dokumentumtípus és -kódolás kezelése
  - egyedi elemzők (saját fejlesztés vagy open-source) vagy
  - IFilter szűrők segítségével

URL elosztó komponens



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

URL elosztó komponens

- konfigurációs állományban megadhatók a letöltendő és kihagyandó URL-minták
- robotokra vonatkozó korlátozások (`robots.txt`, `<meta>` címkék, `rel` attribútum)
- keresőcsapdák észlelése és elhárítása:
  - URL tartomány kézi tiltása menet közben
  - szűrés relevancia alapján
  - mennyiségi korlátozás

Továbbfejlesztési lehetőségek

- konfiguráció és folyamatvezérlés grafikus felületről
- dinamikus, futási idejű átkonfigurálhatóság
- nagyméretű adathalmazok tárolása általános célú adatbázisok helyett struktúrált állományokban
- szimuláció és mérés nagyméretű dokumentumhalmazokon

Összefoglalás

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

A megvalósított rendszer forráskódja elérhető a SourceForge.net-en.

`http://sourceforge.net/projects/webcrawler`

Köszönjük a figyelmet!

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