Web Analyzer

Programátorská dokumentace

Jan Raszyk, 2007

Organizace zdrojových souborů

Zdrojové soubory jsou rozděleny do složek core a msvs2005. První obsahuje jádro aplikace, druhá obsahuje soubor projektu (solution) pro Microsoft Visual Studio 2005 a zdrojové soubory grafického rozhraní programu pro platformu .NET.

Solution se skládá ze dvou projektů. První, wbn_console je konzolová verze, určená pro testování. Samotná aplikace je reprezentována projektem webana.

Program se skládá z modulů crawler, downloader, html_parse, sitestats a common.

Modul crawler

Tento module reprezentuje hlavní strukturu aplikace zodpovědnou za algoritmus procházení webu, správu stažených stránek a patřičného nakládání s nimi, rovněž provádí ukládání a načítání souboru projektu, správu stahování a výstup výsledných statistik. Hlavní objekt je třída Crawler, jeho hlavní funkce je Update. Ta je během stahování opakovaně volána a zajišťuje manipulaci s frontami stránek ke stažení, aktuálně zpracovávaných a v minulosti již navštívených.

Bližší popis této struktury se nachází v komentářích souboru core/crawler.h.

Modul downloader

Tento modul zajišťuje navázání spojení se serverem a stažení jedné HTML stránky z internetu. Hlavní struktura HtmlPageGetter je zodpovědná za vytvoření TCP spojení se serverem, komunikaci protokolem HTTP a zavolání parseru na načtená HTML data. Stahování a zpracovávání HTML stránky se spustí zavoláním funkce StartDownload, ta vytvoří samostatné vlánko pro stahování, aby tak byl umožněn běh několika stahování najednou. Zavoláním funkce Finished lze zjistit, zda již stahování skončilo, ať už skončilo úspěchem nebo neúspěchem.

Bližší popis se nachází v komentářích souboru core/downloader.h

Modul html_parse

Modul zajišťuje naparsování vstupního řetězce obsahujícího HTML dokument do struktury HTML_DOMTree. Ta reprezenuje seznam elementů, jejich atributů a hodnot těchto atributů.

Bližší popis se nachází v komentářích souboru core/html_parse.h

Modul sitestats

Tento modul obsahuje struktury reprezentující informace získané analýzou HTML stránek. Obsahuje strukturu PageStats, reprezentující informace o jedné HTML stránce, a strukturu SiteStats, reprezentující souhrnné informace o množině HTML stránek. Operátor += slouží k přidání statistik jedné stránky k celkovým statistikém, tj. k pomyslnému přičtení objektu typu PageStats k objektu typu SiteStats. Struktura PageStats přijímá jako parametr konstruktoru odkaz na objekt HTML_DOMTree, který v kontruktoru analyzuje a získaná data uloží členských proměnných třídy.

Bližší popis se nachází v komentářích souboru core/sitestats.h

Modul common

V tomto modulu se nacházejí obecné funkce pro manipulaci s URL adresami, například získání názvu serveru z URL adresy, získání adresy dokumentu z URL, atd., nebo funkce pro zjištění, zda řetězec splňuje zadanou wildcard masku.

Bližší popis se nachází v komentářích souboru core/common.h

Grafické rozhraní

Hlavní formulář Form1 reprezentuje úvodní obrazovku programu. Uživatel zde může buď vytvořit nový projekt analýzy, nebo načíst již existující. V prvním případě vybere název souboru projektu, vyplní nastavení (viz uživatelská dokumentace) a spustí stahování tlačítkem Next. V tom případě se inicializuje struktura crawler uživatelskými nastaveními, zobrazí se druhý formulář (Crawling) a v něm se aktivuje objekt Timer (updater). Ten opakovaně volá funkci Update objektu crawler a tím stahování postupuje, dokud funkce Update nevrátí hodnotu true, která znamená, že stahování skončilo.

Pokud uživatel otevře existující projekt, postup je podobný, struktura crawler je ale inicializována daty ze zadaného souboru.