Lehrbeauftragte:
DI Peter Meerwald MSc
DI Robert Praxmarer

Team:
Angerer Theresa
Czernik Thomas
Frick Matthias
Havranek Ivo

Web Spider Dokumentation

In der Lehrveranstaltung

Erweiterte Konzepte der Programmierung

MultiMediaTechnology
Bachelorstudiengang
mmt

Salzburg im SS2010, am 18.04.2010

Inhaltsverzeichnis

1. Broken Links Liste und Statistik

2. Programmarchitektur

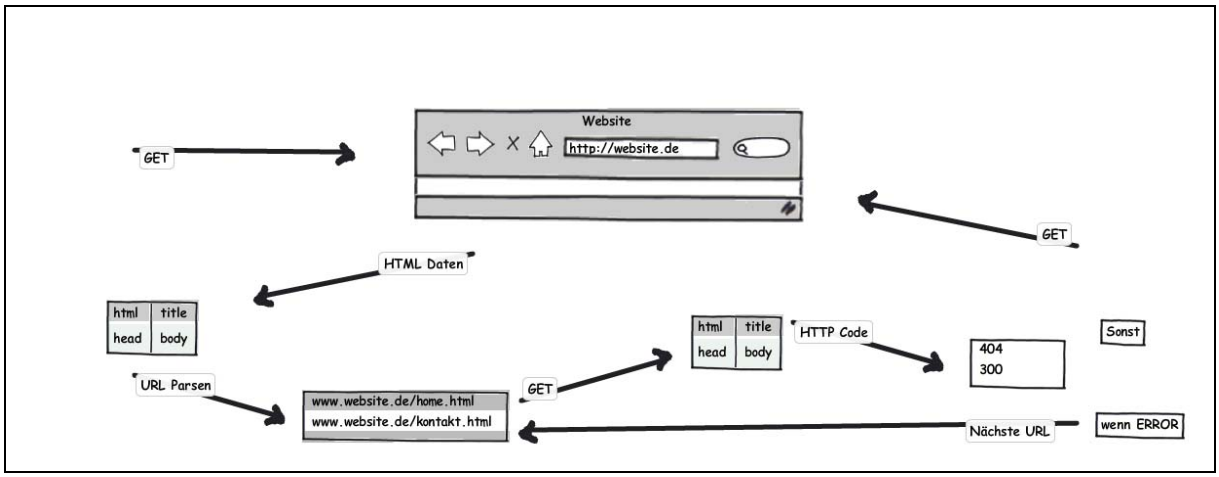

Abb. 1: WebCrawler - grober Aufbau

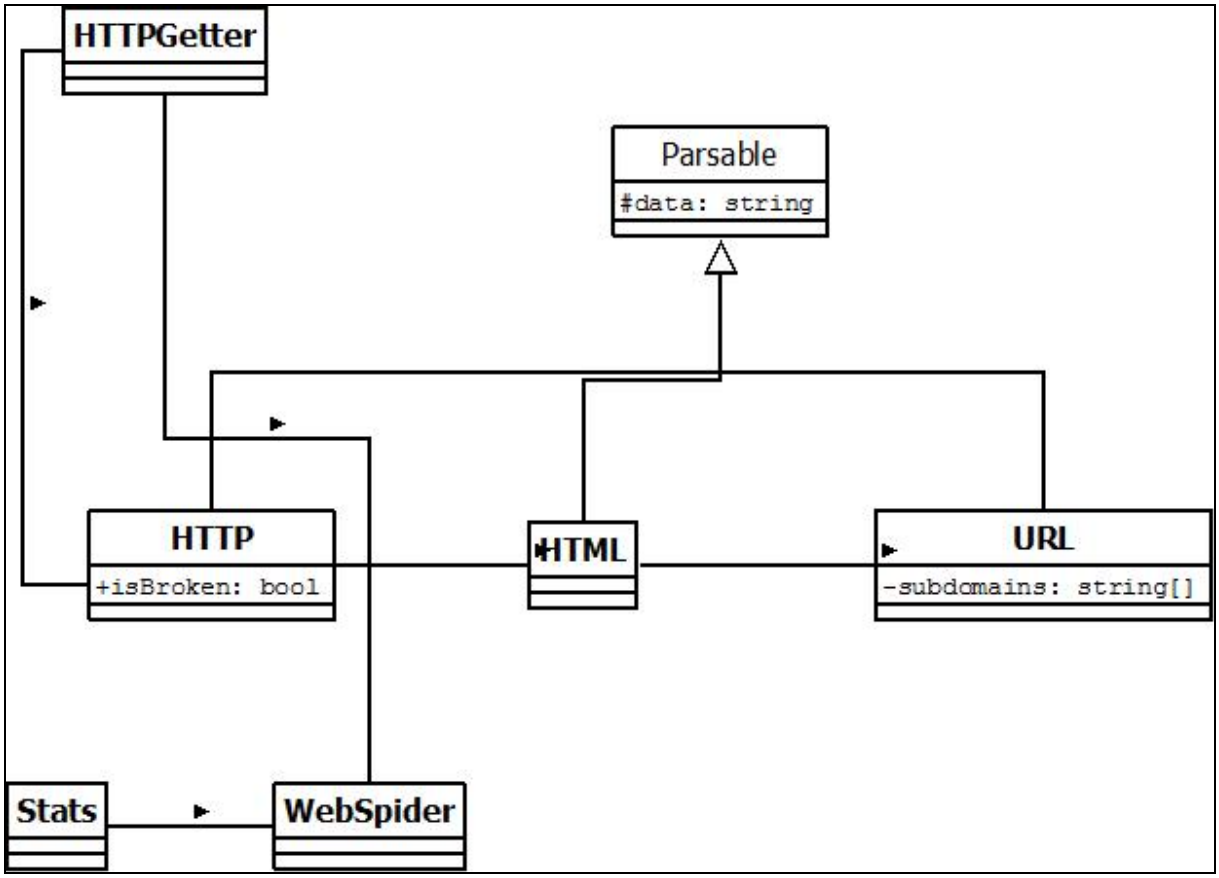

Abb. 2: UML Diagramm

(Pseudo-)code Webspider:

Die Klasse HTMLGetter beinhaltet eine Klasse WebSpider. Diese wiederum enthält die Funktion `crawlURL`:

```
void crawlURL (url) {
    crawledURLs.add (url); // using vec to save data

    if (status_code != 200) {
        brokenURLs.add (url);
    }
    else {
        URLs = HTMLParser.getURLs (url);
        for (int i = 0; i < crawledURLs.size (); i++) {
            for (int j = 0; j < URLs.size (); j++) {
                if (crawledURLs[i] != URLs[j]) {
                    crawlURL (URLs[j]);
                }
            }
        }
    }
}
```

Abb. 3: Code URL Crawling

Die Funktion stellt zunächst noch einen Pseudocode dar, der die URLs verarbeiten kann.

3. Automatisierter Programmaufbau

4. Threading

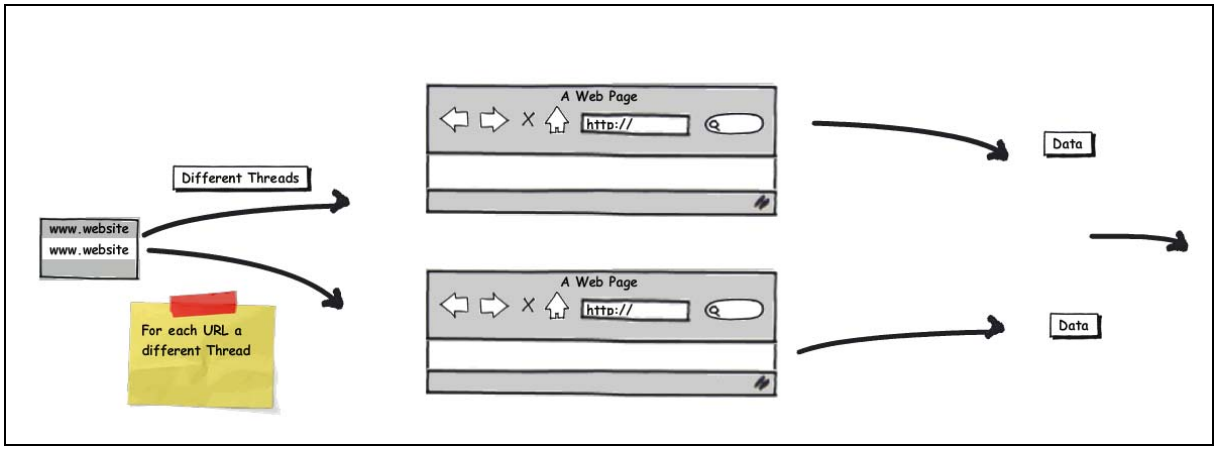

Abb. 4: Threading

5. Multithreading Statistik

6. Anhang