Universität Leipzig

Fortgeschrittene Methoden des Information Retrievals

Praktikum

Endbericht: ProxySwitcher

Quan Nguyen, Sergej Sintchilin

Inhaltsverzeichnis

1 Projektbeschreibung

Der *ProxySwitcher* ist ein Java Tool zur Ausführung von Webanfragen an die Suchmaschinen *eTools.ch* und *Google*. Dabei ist es in der Lage automatisch die öffentliche IP-Adresse zu verändern, um IP-Blockaden Seitens der Suchmaschine zu umgehen und ein flüssiges Suchen zu ermöglichen. Das Tool extrahiert zur Laufzeit selbstständig IP:Port Adressen von Proxyservern aus der Webseite `http://www.hidemyass.com/proxy-list`, welche einer der größten real-time Datenbanken für kostenfreie öffentliche Proxies besitzt. Die Ergebnisse von ausgeführten Suchanfragen speichert das Tool in einer CSV Datei. Dem Nutzer werden dabei Einstellungsparameter zur Verfügung gestellt um die Formatierung der Suchergebnisse zu beeinflussen.

Dieses Tool kann als Kommandozeilenprogramm ausgeführt werden, dabei werden sieben Parameter benötigt, welche die Verarbeitungsmethoden, den Input und Output steuern. Genaue Informationen über die Eingabeparameter und die Bedienung des Tools finden Sie im Abschnitt drei.

Der Quellcode des Projekts ist unter folgender URL erreichbar:
`https://github.com/Q1I/ProxySwitcher`

2 Softwarearchitektur

Die abstrakte Klasse *MultiRequest* deklariert Methoden zur Ausführung von HTTP Anfragen an Webressourcen und zur Verarbeitung der zurückgelieferten Ergebnisse. Dabei kann eine Anfrage direkt an einen Webservice gestellt werden oder es werden mit Hilfe der *ProxySwitcher* Klasse verschiedene Proxyserver dazwischen geschaltet. Als Anfrageergebnis erwartet das Tool ein HTML Dokument, welches mit Hilfe der *jsoup*[1] Bibliothek geparsed wird.

Für das Tool sind genau drei Webressourcen von Bedeutung: Die Suchmaschinen *Google* und *eTools.ch* und die Proxy-Liste (von *hidemyass.com* oder eine selbstdefinierte Proxyliste). Für diese drei Ressourcen existieren jeweils eine Spezialisierung der *MultiRequest* Klasse: *GoogleSearchRequest*, *EtoolsSearchRequest* und *HideMyAssProxyRequest*. Diese Klassen mit der Endung „Request“ sind auf die individuellen Webressourcen zugeschnitten und ermöglichen ein autonomes Bearbeiten der Anfragen und Ergebnisse.

Es wurde bei dem Entwurf und der Implementierung des *ProxySwitchers* darauf geachtet, dass das Tool Möglichkeiten zur Erweiterung auf andere Webressourcen bietet. Dank der abstrakten *MultiRequest* Klasse ist die Ein-

[1] Der Jsoup - Java HTML Parser - ist eine Java Bibliothek für das Arbeiten mit HTML Dokumenten. Es bietet eine umfangreiche API für das Extrahieren und Manipulieren von Daten, mit Hilfe von DOM, CSS Methoden. `http://jsoup.org`

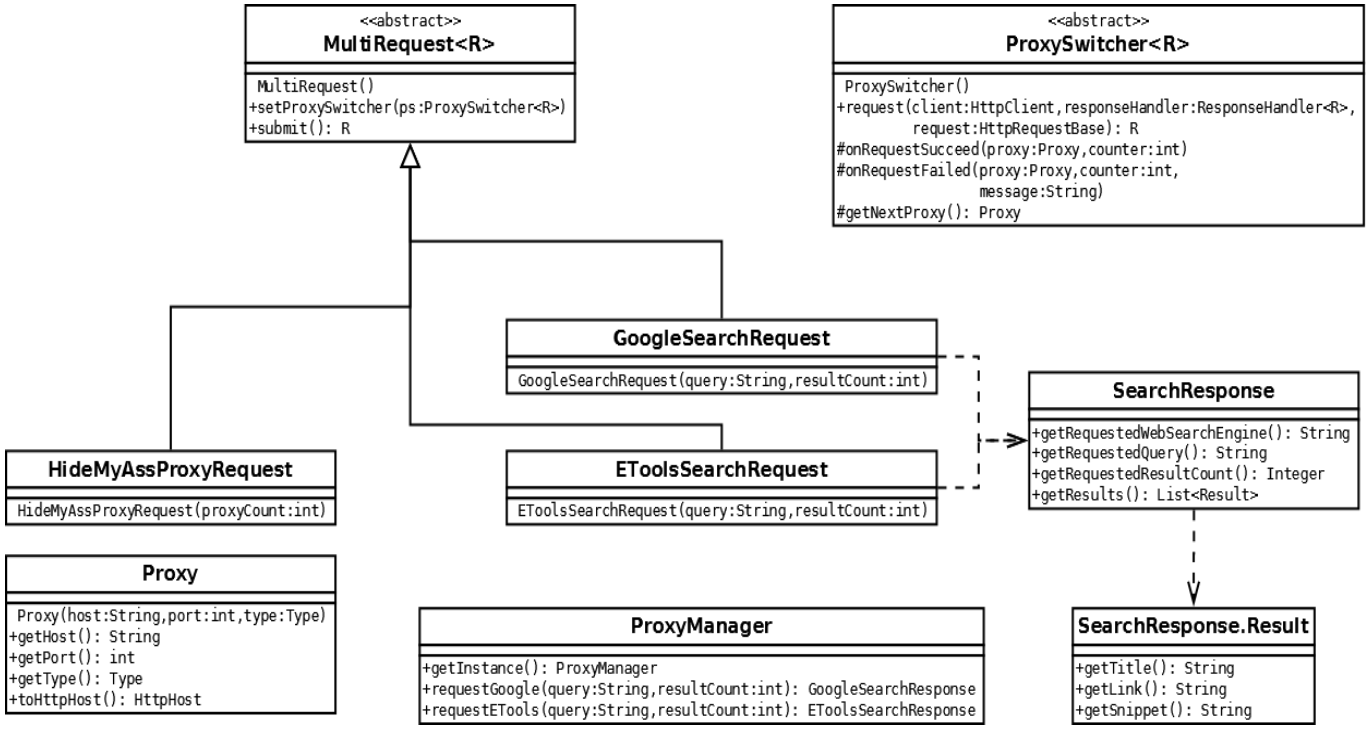

Abbildung 1: UML Klassendiagramm

bindung einer neuen Suchmaschine mit relativ wenig Aufwand verbunden. Zwei wichtige Methoden die bei einer Spezialisierung dieser Klasse implementiert werden müssen sind die *buildUri* und *handleResponse* Methode. Diese spezifizieren für die individuellen Webressourcen die Anfrage URI und die Methodik zur Bearbeitung der Ergebnisse. In der Regel können die Ergebnisse einer Anfrage in zwei Typen unterteilt werden: Der Erste Ergebnistyp wird als *SearchResponse* Objekt modelliert und repräsentiert die Suchergebnisse, die von den Suchmaschinen *Google* bzw. *eTools.ch* zurückgeliefert werden. Bei dem zweiten Ergebnistyp handelt es sich um eine Liste von Proxies, welche dynamisch von der *hidemyass.com/proxy-list* Seite extrahiert werden.

Die *ProxyManager* Klasse verwaltet alle Anfragen und Ergebnisse. Darüber hinaus kümmert es sich um den Konsolen Output und informiert den Nutzer über den Ablauf, Fortschritt und Zustand des Programms und seiner Prozesse.

3 Bedienungsanleitung

Der *Proxyswitcher* kann als Kommandozeilenprogramm ausgeführt werden. Die Klasse *WebSearchLauncher* beinhaltet die Main-Methode, mit dessen Hilfe der *ProxySwitcher* gestartet werden kann. Der Main-Methode können bis zu acht Parameter übergeben werden.

Benutzung:

```
java WebSearchLauncher [Suchmaschine] [Input] [Output]
[zeige Suchstring] [zeige Link] [zeige Titel] [zeige Snippet]
[Proxyliste]
```

Erklärung der Parameter:

1. *Suchmaschine:*
   Wert = google/etools.
2. *Input:*
   Pfad zur Eingabedatei, welche die Suchstrings beinhaltet. Suchstrings werden durch Newline Zeichen getrennt.
3. *Output:*
   Pfad zur Ausgabedatei. Die Ergebnisse der Suche werden in die Ausgabedatei geschrieben.
4. *zeige Suchstring:*
   Wert = true/false. Suchstrings werden in der Ausgabedatei geschrieben.
5. *zeige Link:*
   Wert = true/false. Links werden in der Ausgabedatei geschrieben.
6. *zeige Titel:*
   Wert = true/false. Titel werden in der Ausgabedatei geschrieben.
7. *zeige Snippet:*
   Wert = true/false. Snippets werden in der Ausgabedatei geschrieben.
8. *[optional] Proxyliste:*
   Pfad zur Proxyliste. Es werden nur die Proxies aus dieser Datei verwendet.

Ausgabe:

Die Ergebnisse der Suche werden am Ende in eine Datei im CSV-Format geschrieben, mit der Form:

```
[Suchstring];[Link];[Titel];[Snippet]
...
```

4 Ergebnisse

Im Rahmen des Praktikums haben wir ein Java-Tool entwickelt, dass automatisch die öffentliche IP-Adresse verändern kann und mit dessen Hilfe es möglich ist, mehrere Webanfragen an die Suchmaschinen *Google* und *eTools.ch* zu stellen.

4.1 Einschränkungen

Ein wichtiger Bestandteil für den Betrieb des Tools sind funktionierende Proxies. Einer der größten real-time Listen von frei verfügbaren öffentlichen Proxies befindet sich auf der *hidemyass.com* Seite. Das Tool extrahiert die IP:Port Adressen dieser Proxies und benutzt sie um Anfragen an die Suchmaschinen zu stellen. Leider sind die Proxies dieser Listen nicht immer zuverlässig und performant, da sie sehr beliebt sind und von vielen in Anspruch genommen werden. Oft lassen Suchmaschinen diese Proxies auch sperren. Dadurch kann es sein, dass das Tool erst mehrere Proxies durchprobieren muss bis es eine zuverlässige Proxy gefunden hat.

Das Tool extrahiert Informationen, wie Suchergebnisse oder Proxy-Listen, von HTML Seiten, indem es bestimmte strukturelle Gegebenheiten des HTML Dokumentes ausnutzt. Sollte es in Zukunft aber zu strukturellen Umgestaltungen der HTML Elemente einer Webressource kommen, müssten die Regeln für die Informationsextraktion angepasst werden.

4.2 Erweiterung

Der *ProxySwitcher* unterstützt im Moment zwei Suchmaschinen. Er bietet jedoch die Möglichkeit mit relativ wenig Aufwand neue Suchmaschinen anzubinden. Soll zum Beispiel eine neue Suchmaschine ***Test*** hinzugefügt werden, kann das durch folgende Schritte realisiert werden (Bitte orientieren Sie sich an den Code der existierenden Suchmaschinen):

1. Im Paket `de.uni_leipzig.asv.web.search` eine Klasse ***Test****SearchRequest* anlegen, die von der *MultiRequest* Klasse erbt

2. In der Klasse ***Test****SearchRequest*, die abstrakten Methoden implementieren. Wobei in die *handleResponse* Methode der Code für das Parsen

des HTML Dokumentes geschrieben werden soll.

3. In der Klasse ***Test****SearchRequest*, eine *buildURI* Methode anlegen für die Generierung der Anfrage URI an die Suchmaschine.

4. In der Klasse ***Test****SearchRequest*, eine ***Test****SearchResult* Klasse anlegen, die von *SearchResponse* erbt.

5. Den Code der Klasse *ProxyManager* anpassen. Eine *request****Test*** Methode anlegen.

6. Den Code der Klasse *WebSearchLauncher* anpassen.