Prof. Franz Schweiggert Abteilung Angewandte Informationsverarbeitung 11. Dezember 2003
Christian Ehrhardt Blatt 8

Unix-basierte Implementierung kleiner Datenbanken (WS 03/04)

Abgabetermin 18.12.2003

WWW-Seiten (10 Punkte)

Sie haben sich eine schöne neue WWW-Seite gebaut, allerdings hat diese Seite wie so manche ein leidiges Problem: Viele Links verweisen auf Seiten, die nicht existieren. Deshalb schreiben Sie ein Perl-Skript, das eine Internetseite herunterlädt und die URLs aller Links in dieser Seite auflistet. Um den Link auch zuordnen zu können soll außerdem der Text, der zu dem Link gehört (das was sich im Webbrowser anklicken läßt) mit ausgegeben werden. Keine Bange, alles was Sie dazu wissen müssen steht unten.

Wie komme ich an den Inhalt einer Internetseite?

Es gibt in Perl verschiedene Möglichkeiten, für unsere Zwecke hilft das Kommando `wget`. Mit `wget -q -Otmp http://www.mathematik.uni-ulm.de` wird zum Beispiel die Homepage der Fakultät in der Datei `tmp` gespeichert. Der Exitstatus von wget ist 0, wenn alles geklappt hat und verschieden von 0, wenn ein Fehler auftrat.

Was steht in der heruntergeladenen Datei

Die Datei ist ganz normaler Text, in dem sogenannte Tags eingestreut sind. Tags beginnen mit einem “<”-Zeichen und Enden mit einem “>”-Zeichen. Dazwischen steht eine Folge von Zeichen, die kein “<”und kein “>” enthält. Ein kurzes Beispiel gibt es unter `http://www.mathematik.uni-ulm.de/sai/ws03/udb/blatt8/bsp.html`

Wie erkenne ich einen Link?

Ein Link besteht aus 3 Teilen (wenn hier von Worten die Rede ist spielt Groß-/Kleinschreibung keine Rolle):

- Ein Tag, dessen Inhalt mit dem Wort “a” beginnt und keinen Zeilentrenner enthält. Der restliche Inhalt des Tags kann eine beliebige Folge von Worten und Zeichen sein, aber er muß irgendwann das Wort “href” gefolgt von einem Gleichheitszeichen enthalten. Auf das Gleichheitszeichen folgt dann in Anführungszeichen das Ziel des Links (siehe unten). Die Anführungszeichen können entfallen, wenn das Ziel keine Whitespacezeichen enthält.
- Der zweite Teil ist beliebiger Text, der selbst auch wieder Tags enthalten kann. Dieser Text ist das, was im Browser angeklickt werden kann.
- Der Link wird durch ein spezielles Tag beendet, das nur aus einem Schrägstrich (“/”) und dem Wort “a” besteht.

Wie sieht eine URL aus?

Eine URL bezeichnet eindeutig eine bestimmte WWW-Seite. Sie besteht aus mehrere Teilen. Vorne steht in der Regel “http://”, es kann aber gelegentlich statt “http” auch eine andere Buchstabenfolge vorkommen. Dann folgt der Name des Servers, auf dem die Seite zu finden ist. Der Name erstreckt sich bis zum nächsten Schrägstrich (oder bis zum Ende der URL). Alles ab diesem Schrägstrich ist der Pfad der Seite auf dem Server.

Wie genau wird das Ziel eines Links angegeben?

Um eine vollständige URL aus dem Text hinter HREF zu bekommen benötigt man zusätzlich auch noch die URL der Seite, auf er sich der Link befindet. Für den Text hinter HREF gibt es folgende Möglichkeiten:

- Es handelt sich um eine vollständige URL. Dann ist einfach diese URL das Ziel des Links.
- Wenn keine vollständige URL angegeben wurde, dann ist eine Datei auf dem selben Server wie die Seite selbst gemeint. In diesem Fall wird nur der Pfad der Datei angegeben und zwar
  - Absolut, wenn der Pfad mit einem Schrägstrich (“/”) beginnt.
  - Relativ zum Pfad der Seite, in allen anderen Fällen.

Bei der Ausgabe sollen alle Links mit vollständigen URLs angegeben werden, die auch kein “..” mehr enthalten.

Beispiel

Das oben bereits erwähnt Beispiel sieht so aus:

```
<html>
<head>
<title>
Unix-basierte Implementierung kleiner Datenbanken
(WS 2003/2004)
</title>
</head>
<body background="/sai/melzer/gifs/wall.jpg">
<address><A HREF="http://www.uni-ulm.de/"><B>Universit&auml;t</B>
Ulm</A>, <A HREF="/sai/">SAI</A>,
```

```
<!-- Kommentar ueber
mehr als eine Zeile -->
</FONT></FORM></CENTER>
<HR><A HREF="index.html">hier</A><a href=../blatt1/>Aufgabe</a>,
</TD></TR><TR><TD></TD><TD>
</ADDRESS>
</body>
</html>
```