Programmierprojekt zu „Einführung in die Informatik 2“

– Automatische Datenextraktion aus HTML-Dokumenten –

Sven Kosub

AG Algorithmik/Theorie komplexer Systeme
Universität Konstanz

E 202 | Sven.Kosub@uni-konstanz.de | Sprechstunde: Mittwoch, 14:00-15:00 Uhr, o.n.V.

Sommersemester 2008

Inhalt

Ziel des Projektes: Java-Implementierung eines Algorithmus zur automatischen Datenextraktion aus HTML-Dokumenten mittels baumartig srukturierten Wrappern

Hintergrund:

- viele Web-Seiten werden automatisch aus Datenbank heraus als HTML-Dokumente generiert
- HTML-Dokumente sind durch Tags semi-strukturierte Texte
- Tags schließen die relevante Daten ein (z.B. `<tr><td>...</td></tr>`)
- bei Suche nach relevanten Daten sind lokale Umgebungen bedeutsam
- Wrapper sind möglichst eindeutige Beschreibungen lokaler Umgebungen um Textbausteine

Beispiel

HOCLRT-Wrapper:

- Head = `Firma</th>`
- Open = `user-name`
- Close = `</tr`
- Left_1 = `>`
- Right_1 =`</a>`
- Left_2 = `br>`
- Right_2 = `</td`
- Tail = `</tbody`

1. Aufgabenblock

- Lesen Sie den Aufsatz:

  Matthias Hanitzsch:
  Automatische Datengenerierung aus HTML-Dokumenten.
  http://www.inf.uni-konstanz.de/algo/lehre/ss08/info2/

- Suchen Sie sich 5 HTML-Dokumente von verschiedenen Domänen, die für HOCLRT-Wrapper geeignet sind
- Überlegen Sie sich eine einfache Beschreibungssprache für HOCLRT-Wrapper
- Beschreiben Sie für Ihre HTML-Dokumente Wrapper und legen Sie diese in WRP-Dateien ab
- Implementieren Sie eine Klasse, die WRP-Dateien für HOCLRT-Wrapper versteht
- Implementieren Sie eine Klasse, die HOCLRT-Wrapper auf HTML-Dokumente anwendet und die Daten in einer CSV-Datei ausgibt
- Testen Sie Ihre Klasse auf Ihren Beispiel-Dokumenten

Organisatorisches

Ablauf:

- zwei Aufgabenblöcke á 3 Wochen (jeweils im Wert von einem Übungsblatt)
- erster Block bis 26.06.2008 (Termin für Projekttreffen?)
- zweiter Block bis 17.07.2008 (Termin für Projekttreffen?)
- ZIP-Pools G 228-230
- Anmeldung über Account-Tool (Passwörter?)

Hilfsmittel:

- http://java.sun.com/j2se/1.5.0/docs/api/
- http://www.eclipse.org/