HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

Kurzfassung der Semesterarbeit

Abteilung	**Informatik**
Name der Diplomandin / des Diplomanden	**Maurus Gmünder** **Boris Burgstaller**
Semesterarbeit	**Sommer Semester 2003**
Titel der Semesterarbeit	**WebEx -Web Daten Extractor**
Betreuer	**Bruno Feurer**

Kurzfassung der Semesterarbeit

Das Ziel der Semesterarbeit WebEx ist eine Applikation zu entwickeln, mit deren Hilfe man Daten aus Webseiten extrahieren kann.
Unser Ansatz zur Lösung dieses Problems ist, zuerst die Daten aus dem Web zu bereinigen und in eine XML Struktur zu bringen und dann zu filtern.
Möglichkeiten Webcontent herunterzuladen:
Um Webcontent herunterzuladen haben wir zwei Möglichkeiten. Wir laden die Seite so wie sie ist mit einem http get request herunter und erhalten so den original Sourcecode. Die zweite Möglichkeit ist sie mit dem im NET Framework enthaltenen axBrowser, einem nahen Verwandten des Internet Explorers, herunterzuladen. Der Sourcecode wird dann mit Hilfe einer Internet Explorer Instanz korrigiert und abgespeichert.
Mit verschiedenen Filtern können danach die Daten weiterverarbeitet werden. Wir bieten folgende zur Auswahl:

- HTML2XML: konvertiert HTML nach XML
- XSLT Converter: führt XSL Transformationen durch
- Regexp: wendet "Regular Expressions" auf die Daten an
- StringMod: Bietet String Funktionen wie toUpper, toLower, removeLineBrakes, Find&Replace an.
- AddString und AddFile: fügen einen Pre- oder Postfix an die Daten an

Die Filter können in beliebiger Reihenfolge hinzugefügt werden. Ihre Aufgabe ist, die Daten, die wir vom Internet herunterladen in die gewünschte Form zu bringen. Mit diesen Filtern können Konfigurationen erstellt werden, die es erlauben Informationen immer noch zu finden, auch wenn sich die Struktur der Inputdaten bis zu einem gewissen Grad ändert. Das Resultat wird dann in einem integrierten Browser dargestellt, und in eine Datei abgespeichert.
Ein Beispiel wäre das tägliche Zusammentragen von Wetterdaten verschiedener Meteoseiten. Die Wetterseiten könnten sich in ihrer Struktur ändern, ohne dass sich dies auf den Output auswirken würde.