HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

Kurzfassung der Semesterarbeit

Abteilung	Informatik
Name der Diplomandin / des Diplomanden	Thomas Schwyn Adrian Höhn
Diplomjahr	2006
Titel der Semesterarbeit	WebCrawler für Geoinformationen und Geowebdienste
Examinator	Prof. S. F. Keller, Abt. Informatik HSR

Die Studienarbeit befasste sich damit, einen WebCrawler (GeometaBot) zu programmieren, welcher das Internet nach georelevanten Inhalten durchsucht und die Urls dazu in einer Datenbank ablegt. Der GeometaBot sollte in Java geschrieben werden. Um eine gute Grundlage für den GeometaBot zu haben, mussten wir eine Evaluation von bestehenden Open Source – Java - WebCrawlern machen. Nach Abschluss der Evaluation war Heritrix als Grundlage gegeben. In einer ersten Iteration ging es darum Heritrix mit einigen Binär-Erkennungen auszustatten. Binär-Erkennungen sind Erkennungsmechanismen, welche mit Ja / Nein klassifizieren. Diese dienen dazu verschiedene Geoformate und -dienste zu finden. Einige Beispiele für Geoformate wären geotiff, Interlis – Formate oder GML. Beispiele für Geodienste wären WMS und WFS so genannte Kartenserver, welche bei Anfrage eine (Land-) Karte zurückliefern. In einer zweiten Iteration wurde der unscharfe Erkennungsmechanismus hinzugefügt. Dieser ist für die Text-Klassifikation zuständig und gibt pro Webseiten-Inhalt eine Relevanz von 0 (gar nicht relevant) bis 100 (maximal relevant) an. Die Text-Klassifikation wurde mit Hilfe des NaiveBayes - Klassifizierers implementiert. Des Weiteren wurde in dieser Iteration die Erkennung von parametrisiert aufrufbaren Kartenseiten wie map24 implementiert. Ebenfalls in dieser Iteration wurde der Scheduler realisiert, welcher einen GeometaBot – Job zu gegebener Zeit pausiert bzw. wieder startet. Um den GeometaBot möglichst userfreundlich zu gestalten, wurde eine webbasierte Verwaltung integriert. Diese ist in Java Server Pages geschrieben um Interaktivität mit dem GeometaBot zu ermöglichen, HTML wird für die Darstellung verwendet und JavaScript um clientseitige Fehleingaben abzufangen.

Das Projekt wurde für http://www.geometa.info entwickelt, eine themenspezifische Suchmaschine, welche seit dem Jahre 2003 am Kompetenzzentrum für integrierte Geo-Informationssysteme an der Hochschule für Technik, Rapperswil (CC integis HSR) betrieben wird.