HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

Kurzfassung der 2. Studienarbeit

Abteilung	**Informatik**
Name der Studentin / des Studenten	**Sebastian de Castelberg** **Markus Kinzler**
Studienjahr	**SS04**
Titel der 2. Studienarbeit	**Projekt Geometabot – Ein fachspezifischer WebCrawler**
Examinatorin / Examinator	**Stefan F. Keller**

Kurzfassung der 2. Studienarbeit

Im Projekt GeometaBot0.1 geht es darum einen fokussierten WebCrawler auf Basis von Java zu Implementieren. Der Crawler sammelt automatisch URL's von georelevanten Seiten und bereitet diese auf, so dass sie in den Index vom Suchportal www.geometa.info aufgenommen werden können.

Als Basis des Crawlers dient der WebCrawler Heritrix. Heritrix bietet sehr gute Schnittstellen um dem Crawler zu konfigurieren und zu jedem Zeitpunkt des Crawls seine eigenen Module einzubinden.

Um GeometaBot auf den Geobereich zu fokussieren wurden zwei Komponenten eingefügt:

- LinkDistiller: Priorisiert das Besuchen von Links die georelevante Wörter beinhalten
- Classifier: Der Inhalt der heruntergeladenen Seiten wird analysiert und mittels Mathematischer Methoden auf die Georelevanz geprüft.

Der Crawler ist zurzeit noch etwas langsam, was aber in zukünftigen Heritix Releases verbessert wird.

Der Classifier, der ganze Texte klassifizieren soll, ist nicht so gut wie erwartet. Dies funktionierte mit den WEKA Implementationen von J48 und NaiveBayes nicht wie gewünscht. Zuviele (auch grobe) Fehler machen den Classifier in diesem Entwicklungsstadium nicht brauchbar .