Webarchiv

Mehrere Forschungsprojekte in dem Institut für Informationssysteme beschäftigen sich mit Algorithmen für die Verwaltung und Nutzung von Texten. Für diese Forschungsprojekte wird in der Regel ein umfangreicher Datensatz für den Test der Algorithmen benötigt. Im Rahmen dieses Projekts soll daher ein Webarchiv aufgebaut werden, das eine große Anzahl an Texten aufnehmen, verwalten und durch neue Dokumente aus dem Internet ergänzen kann. Dabei soll eine Größenordnung von mehreren Petabyte an Daten problemlos unterstützt werden können.

Der Fokus dieses Projekts ist eine robuste und dauerhafte Speicherung der Webseiten und der beschreibenden Daten (Metadaten). Als Metadaten sind mindestens die URI, der Zeitpunkt der Akquisition sowie die Titel der Dokumente zu speichern. Der Zugriff auf diese Daten soll durch ein Datenbankmanagementsystem unterstützt werden.

Das Webarchiv soll folgende Funktionen unterstützen:

- Ausgehend von einigen Start-URIs soll ein *Crawler* regelmäßig das Internet nach neuen Dokumenten durchsuchen. Dazu soll das System der Linkstruktur der gefunden Seiten folgen. Der Crawler soll nur HTML-Seiten laden, die einer vorgegebenen Liste von Sprachen entsprechen. Dazu soll ein geeignetes Verfahren zur Bestimmung von Sprachen implementiert werden. Außerdem soll die Möglichkeit für ein Plug-In geschaffen werden, das weitere Filterfunktionen integrieren kann. Die Frequenz der Suche sowie die allgemeine Suchtiefe sollen einstellbar sein. Die gefundenen Seiten werden in einem „gepackten Format“ gespeichert. Das Crawler soll als Metadaten bereits den Zeitpunkt der Akquisition sowie den Titel des Dokuments speichern (HTML-Element „Title“). Es soll außerdem ein Mechanismus implementiert werden, mit dem Duplikate in dem Textarchiv vermieden werden. Das Format soll mit der Gruppe der räumlichen Suchmaschine abgestimmt werden.
- Das System soll die parallele Akquisition durch mehrere Instanzen des Crawler unterstützen. Dabei soll ein Ansatz gewählt werden, so dass die unterschiedlichen Crawler-Instanzen möglichst wenig miteinander interagieren müssen. Auch dieser Ansatz soll Duplikate vermeiden.
- Das System soll die Integration verschiedener Analysemethoden unterstützen, von denen ebenfalls jeweils mehrere parallele Instanzen möglichst unabhängig ablaufen können. Die Ergebnisse dieser Methoden sollen als Metadaten in Dateien gespeichert werden. Die Analysemethoden sollen hier mit Hilfe der Extraktion von normalisierten Begriffen aus Webseiten demonstriert werden (Normalisierung Zeichenketten einschließlich Konfertierung in Kleinbuchstaben, Elimination von Stoppwörtern, Ableitung von Wortstämmen).
- Für den Zugriff auf das Webarchiv soll eine *Datenbank* realisiert werden, welche die gesamten Metadaten aufnehmen kann. Für das Einlesen der Daten solle eine eigene Komponente mit der Bezeichnung „*Ingestion*“ realisiert werden, die Metadaten aus ausgewählte Dateien oder alle Metadaten aus allen Dateien liest und in die Datenbank schreibt. Die Komponente Ingestion soll konfigurierbar sein, so dass auch Metadaten lesen kann, die von zukünftigen Analysetools erzeugt werden.