

Ortsbezug für Web-Inhalte

Diplomarbeit

Mihály Sütö

Institut für Parallele und Verteilte Höchstleistungsrechner (IPVR)
Abteilung Anwendersoftware
Universität Stuttgart
Breitwiesenstraße 20-22
D-70565 Stuttgart

1. Einleitung

Die Forschergruppe NEXUS entwickelt eine offene und globale Infrastruktur für ortsbezogene Anwendungen für mobile Benutzer. Im Zentrum der Infrastruktur steht das Augmented World Model, das eine durch virtuelle Objekte erweiterte Abbildung der realen Welt ist. Diese Objekte können der Einbindung externer Informationsquellen, wie zum Beispiel des World Wide Web, dienen. Das Ziel dieser Diplomarbeit ist die Suche und die Bereitstellung von Webseiten mit Ortsinformation für die NEXUS Infrastruktur.

In der Studienarbeit [Süt01] des Autors wurde bereits die Möglichkeit untersucht, einen Ortsbezug von Webseiten über die Information in Meta-Tags herzustellen. Durch Metadatenstandards, wie das *Dublin Core Metadata Element Set*, kann in einem Meta-Tag einer Webseite geographische Information gespeichert werden. Leider werden Metadatenstandards und vor allem Meta-Tags mit genauer geographischer Information sehr selten verwendet.

Hauptaufgaben dieser Diplomarbeit sind Zusammenfassung, Analyse und Bewertung von weiteren Verfahren, die einen bestehenden Ortsbezug aus Webseiten extrahieren können.

2. Ortsinformation in Webseiten

Für die Extrahierung von geographischer Ortsinformation aus Webseiten bieten sich unter anderem folgende Möglichkeiten:

- Die Ermittlung der Adresse des administrativen Ansprechpartners für die entsprechende Domain aus der DENIC-Whois Datenbank.
- Die Ermittlung der Adresse des technischen Ansprechpartners für den entsprechenden IP-Adressenbereich aus der RIPE-Whois Datenbank.
- Die Analyse von Webseiten unter Einbeziehung aller Webseitenkomponenten und die Extrahierung von Kandidaten, die möglicherweise Ortsinformationen enthalten. Solche Kandidaten können Ortsnamen, wie zum Beispiel „Stuttgart“, sein. Anschließend müssen diese durch einen Vergleich mit einem Katalog verifiziert werden.
- Ermittlung einer zuverlässigen Ortsinformation aus den Webseitenkomponenten. Eine solche Ortsinformation könnte zum Beispiel eine komplette Adresse sein, die aus dem Webseiteninhalt extrahiert wird und die Webseite mit einer eindeutigen, geographischen Position verbindet.

Anschließend werden die ermittelten Ortsnamen in genaue Positionen überführt. Zu diesem Zweck wird eine Datenbank, die Ortsnamen auf Positionen abbildet, verwendet.

3. Erweiterung des Webroboters DCbot

Die erfolgversprechendsten Verfahren, einen bestehenden Ortsbezug aus einer Webseite zu extrahieren, werden im Rahmen dieser Arbeit implementiert. Zu diesem Zweck wird der Webroboter DCbot, der während der Studienarbeit des Autors entstand, zweckmäßig modifiziert.

Der Webroboter ist zur Zeit in der Lage, einen Ortsbezug zu einer Webseite durch Auswertung von, auf der Seite vorhandenen, Meta-Tags herzustellen.

4. Ergebnisse

Die effektive Arbeitsweise des Webroboters wird mit mehreren Testläufen überprüft. Die verschiedenen Verfahren zur Extrahierung von Ortsinformation aus Webseiten werden auf Basis der Testergebnisse miteinander verglichen und bewertet. Abschließend werden aus den Untersuchungen gewonnene Erkenntnisse vorgestellt.

Literaturverweise

[Süt01] Sütö, M.: *Ortsbasierter Web-Zugriff*. Studienarbeit Nr. 1834, Institut für Parallele und Verteilte Höchstleistungsrechner, Universität Stuttgart, 2001