hooolp.net

Exposé zur Studienarbeit

Identifizieren und Extrahieren von Musikveranstaltungen aus dem Web

Hung Le
hle@informatik.hu-berlin.de

19. Oktober 2008

Betreuer: Prof. Dr. Ulf Leser, Manfred Pokrandt

1 Hintergrund

Hooolp - hooolp.net ist ein Startup-Unternehmen, welches sich die Aufgabe stellt, eine internationale Live-Music-Event-Online-Datenbank aufzubauen. Die Idee hierbei ist, dass sich Musikveranstalter sowie Bands auf der Seite anmelden und dort ihre Veranstaltungen selbständig eintragen können. Dabei sollen sie die Veranstaltungen durch spezielle Informationen wie Bandname, Ort der Veranstaltung, Zeitpunkt oder Genre spezifizieren. Durch diese Informationen sollen Benutzer schnell und überall nach favorisierten Veranstaltungen suchen können.

Dadurch, dass die Veranstaltungen noch manuell vom Veranstalter oder den Bands eingetragen werden müssen, ist es schwierig eine umfassende und aktuelle Musikveranstaltungsdatenbank aufzubauen. Dabei gibt es folgende Schwierigkeiten:

- Veranstalter oder Bands können nicht komplett erfasst werden. Dadurch fehlen Veranstaltungen.
- Für den Start ist es schwierig, die Veranstalter/Bands zu motivieren, ihre Veranstaltungen dort regelmäßig eintragen zu lassen. Es muss ausreichender Traffic für die Seite geschaffen werden, damit sie einen Vorteil darin sehen ihre Veranstaltungen dort einzutragen.

Aufgrund dieser Schwierigkeiten ist es sinnvoll, zusätzlich zur manuellen Eingabe, eine Möglichkeit zu finden, Veranstaltungen automatisch im Internet zu erfassen.

2 Zielsetzung

Um die automatische Extraktion und Evaluation der Veranstaltungsinformationen von Webseiten im Internet zu ermöglichen, sind zwei Schritte erforderlich.

1. Zuerst müssen Webseiten gesucht werden, die Veranstaltungsinformationen enthalten. Bei diesem Schritt geht man ähnlich wie bei der Indizierung von Webseiten durch Suchmaschinen vor. Die indizierten Seiten werden dann nach „Veranstaltung enthalten“ und „Keine Veranstaltung enthalten“ klassifiziert.
2. Aus den Webseiten, die als „Veranstaltung enthalten“ klassifiziert wurden, wird versucht, die Veranstaltungsinformationen zu extrahieren. Hierbei werden die Webseiten nach strukturellen und semantischen Merkmalen untersucht.

Dieser Studienarbeit befasst sich mit dem zweiten Schritt, der Extraktion von Veranstaltungsinformationen. Hierbei werden gängige Ansätze und Tools für die

Informationsextraktion aus Webseiten untersucht. Im Anschluss soll versucht werden ein Tool zu entwickeln, dass die Informationsextraktion aus Webseiten ermöglicht.

3 Herangehensweise

Viele Informationen, die von Webseiten bereitgestellt werden, sind strukturierte Daten, die sich in Datenbanken befinden. Sie werden erst verarbeitet und dann dem Benutzer in einer nicht strukturierten Form (HTML für Webseiten) angezeigt [1]. Da die Datenbanken aus Benutzersicht verborgen sind, ist die Umkehrung dieses Schrittes erwünscht.

Web-Wrapper sind spezialisierte Programmroutinen, die automatisch Daten von Webseiten extrahieren und in strukturiertem Format ausgeben. Dabei erfolgt der Prozess in drei Phasen [2]. Diese Phasen können sich wiederholen, wenn sich die Informationen auf einer anderen Seite sich befinden.

1. Download der HTML-Seite
2. Erkennen und Auslesen der Daten
3. Speicherung der Daten in strukturiertem Format

Jeder Web-Wrapper kann von Grund aus mittels regulärer Ausdrücke für jeden Anwendungsbereich entwickelt werden. Für Webseiten ist es gebräuchlich, dass die Seiten in einem DOM-Baum geparst und die Daten dann über Traversierung oder mittels XPATH ausgelesen werden. Es gibt aber auch Tools die Web-Wrapper mit vordefinierten Parametern generieren. Die Ausgabe kann dann in ein strukturiertes Format exportiert werden.

Für die Studienarbeit sind folgende Aufgaben zu erfüllen:

1. **Auswahl eines geeigneten Web-Wrapper-Tools**
   Zunächst muss untersucht werden, welche Tools verfügbar sind und ob sie für diese Aufgabe geeignet sind. In [2], [3] und [4] werden Web-Wrapper-Tools unter verschiedenen Aspekten untersucht und miteinander verglichen. Die Aspekte sind u.a. der Automatisierungsgrad, Aufgabenbereiche und verwendete Techniken. Es gibt sowohl kommerzielle als auch nicht-kommerzielle Lösungen. Diese unterscheiden sich von der Bedienung, zum anderen nutzen sie unterschiedliche Verfahren.
2. **Datenaufbereitung und Anwendung der Tools**
   In diesem Schritt werden Webseiten gesammelt, die Veranstaltungsinformation enthalten. Hooolp bietet bereits eine Liste von Seiten mit Veranstaltungen. Hier ist es auch sinnvoll auf Webseiten von Bands-, Veranstaltern oder auf lokalen Seiten zu schauen. Die Tools werden dann konfiguriert und mit den gesammelten Seiten trainiert.

3. Evaluation der Ergebnisse

Aus den Ergebnissen wird die bestmögliche Lösung erstellt und von Hooolp im Hinblick auf die Korrektheit überprüft. Dabei wird untersucht, ob die ausgebebenen mit den gewünschten Formaten übereinstimmen. Sind es überhaupt Musikveranstaltungen? Stimmen Bandname, Ort und Zeitpunkte überein? Das Ergebnis wird dem Aufwand für das Entwickeln der Lösung gegenübergestellt und über die Machbarkeit der Lösung entschieden. Ein Nebenprodukt dieses Schrittes ist der systematische Aufbau einer Sammlung von Musterlösungen der Art (Webseite, Veranstaltungsliste). Mit dieser Liste können spätere Weiterentwicklungen automatisch getestet werden.

4. Entwicklung eines Tools

Lösungsansätze und Integrationsmöglichkeiten für den Einsatz auf Hooolp.net werden herausgearbeitet. Das daraus resultierende Ergebnis wird für die Entwicklung eines Tools zum Extrahieren von Veranstaltungsinformationen aus Webseiten genutzt.

4 Literatur

[1] Kevin Arnoult (Sep 2007), Extend Data Sources Available in Alibaba, Masterarbeit, Heriot-Watt University, Edinburgh, Scotland

[2] Stefan Kuhlins und Ross Tredwell (2003), Toolkits for Generating Wrappers - A Survey of Software Toolkits for Automated Data Extraction from Web Sites, University of Mannheim - Department of Information Systems III

[3] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled Shaalan (2006). A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering, 18(10):1411 - 1428

[4] Shin Wee Chuang (June 2004), A Taxonomy And Analysis Of Web Wrapping Technologies, MIT Massachusetts, Working Paper CISL# 2004-08