hooolp.net

Exposé zur Diplomarbeit

Identifizieren und Extrahieren von Musikveranstaltungen aus dem Web

Hung Le
hle@informatik.hu-berlin.de

29. Juni 2009

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 detaillierte Informationen wie Bandname, den Ort der Veranstaltung, den Zeitpunkt oder das Genre spezifizieren. Durch diese Informationen können Benutzer schnell und überall nach favorisierten Veranstaltungen suchen.

Dadurch, dass die Veranstaltungen noch manuell vom Veranstalter oder den Bands eingetragen werden müssen, ist es schwierig eine umfassende und aktuelle Datenbank für Musikveranstaltungen aufzubauen. Es ergeben sich dabei folgende Schwierigkeiten:

- Ein vollständiges Erfassen der Veranstalter oder Bands kann nicht gewährleistet werden, wodurch Veranstaltungen fehlen.
- Am Anfang ist es schwierig, die Veranstalter/Bands zu motivieren, ihre Veranstaltungen regelmäßig einzutragen. Es muss daher ausreichender Traffic für die Seite geschaffen werden, damit sie einen Vorteil darin sehen ihre Veranstaltungen einzutragen.

Aufgrund dieser Schwierigkeiten ist es sinnvoll, zusätzlich zur manuellen Eingabe, eine Möglichkeit zu finden, Veranstaltungen automatisch im Internet zu erfassen.Um die automatische Extraktion und Evaluation der Veranstaltungsinformationen von Webseiten im Internet zu ermöglichen, sind zwei Schritte erforderlich.

1. Zuerst müssen Webseiten identifiziert 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.

Diese Diplomarbeit baut auf die Studienarbeit [7] zum gleichen Thema „Identifizieren und Extrahieren von Musikveranstaltungen aus dem Web“ auf.

2 Zielsetzung

In diesem Zusammenhang verfolgt die Diplomarbeit folgende Ziele:

- Verbesserung und Verallgemeinerung des Extraktionsalgorithmus auf Veranstaltungsdarstellungen auf Webseiten (genaue Fehleranalyse der Ergebnisse der Studienarbeit, Einführung von visuell basierte Techniken [3], Extraktion der Daten mittels Partial Tree Alignment [2])
- Extraktion von Bandnamen (Vergleich mit Datenbanken, Bewerten und Gewichten von möglichen Bandnamen via MySpace/Google)
- Verbesserung der Identifikationsmethode von Veranstaltungsseiten (Verbesserung des Crawlers, Verbesserung der Score-Funktion [7])
- Entwicklung eines webbasierten Tools, basierend auf der Grundlage der oben genannten Punkte.

Ziel dieser Diplomarbeit ist es zu zeigen, dass es mit Hilfe von modernen Ansätzen der Informationsextraktion, wie z.B. in [1, 2, 3, 4, 5, 6] beschrieben, möglich ist, Veranstaltungsseiten aus dem Web zu identifizieren und die dazugehörige Veranstaltungsinformationen mit hoher Genauigkeit und Abdeckung zu extrahieren.

3 Herangehensweise

Veranstaltungen, die von Webseiten bereitgestellt werden, sind meist strukturierte Daten, die in Datenbanken gespeichert sind. Sie werden erst verarbeitet und dann dem Benutzer als Listen von ähnlichen Einträgen in Form von HTML präsentiert [7]. Die Listenelemente werden auch Datensätze genannt. Da die Datenbanken aus Benutzersicht verborgen sind, ist die Umkehrung dieses Schrittes notwendig.

3.1 Identifizieren von Datenregionen von Datensätzen

Wie in der Studienarbeit, auf der diese Diplomarbeit aufbaut, wird zuerst versucht, Datenregionen von Datensätzen aus vorgegebenen Veranstaltungsseiten zu extrahieren. Datenregionen sind Bereiche auf der Seite, welche ähnliche Datensätze enthalten. Statt sich wie in der Studienarbeit nur auf Tabellen (Abbildung 1) zu konzentrieren, bei dem jeder Datensatz in einer Tabellenreihe befindet, wird in diese Arbeit versucht, auch auf anderen Darstellungsformen von Datensätzen (Abbildung 2, Abbildung 3) zu einzugehen.

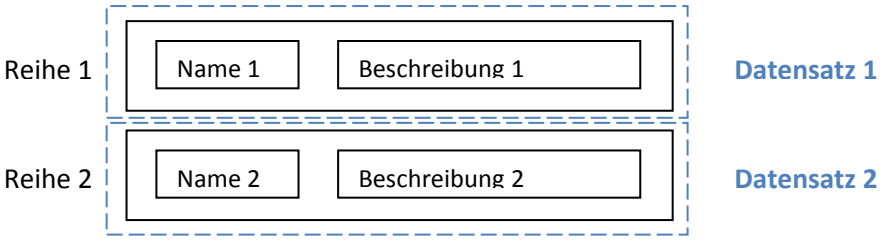

Abbildung 1: Tabellenstruktur, jeder Datensatz in einer Reihe

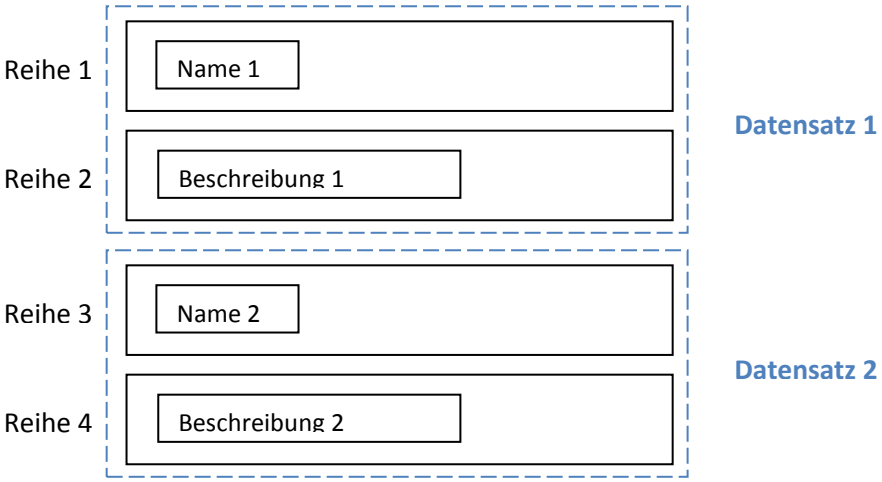

Abbildung 2: ein Datensatz aus mehreren Reihen (Blöcken)

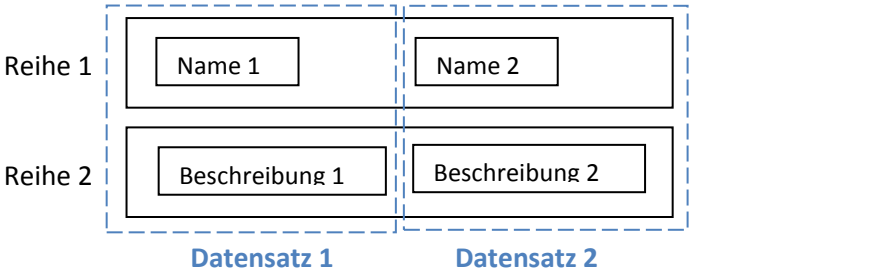

Abbildung 3: Mehrere Datensätze in einer Reihe

Webseiten sind Hypertext-Dokumente, die in HTML geschrieben werden. Als Erstes muss von der Webseite den Quellcode heruntergeladen und daraus einen HTML Tag Tree aufgebaut werden. Ein HTML Tag Tree ist ein Baum, der aus Knoten von HTML Tags besteht. Die Kinder eines Knoten sind diejenigen Tags, die von dem Tag des Knoten eingeschlossen sind. Entweder man erstellt so einen HTML Tag Tree mit Hilfe von XPATH [1]oder mit einer visuell basierten Methode [3]. Die visuell basierte Methode hat den Vorteil, dass sie nicht stark von HTML abhängig ist. Bei fehlerhaften HTML Codes kann der HTML Tag Tree trotzdem noch aufgebaut werden und der Baum ist so aufgebaut wie der Benutzer ihn im Browser sieht.

[1] http://www.w3.org/TR/xpath

Nachdem der HTML Tag Tree aufgebaut ist, wird versucht Datenregionen auf der Webseite zu lokalisieren. Um Datenregionen zu ermitteln, vergleicht man die Tagspfade von benachbarten Knotenkombinationen miteinander [1]. Ein Tagspfad eines Knoten beinhaltet auch Tags, die unterhalb des Knoten auftreten. Voraussetzung dafür ist, dass mehr als ein Datensatz in einer Datenregion existiert; ein Umstand, der für Veranstaltungen meist zutreffend ist. Für den Vergleich wird der Levenshtein-Distanz-Algorithmus [2] verwendet. Ein Datensatz besteht normalerweise aus höchstens 10 Tags (Abbildung 4), so ergeben sich folgende Vergleiche:

- (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8), (8, 9), (9, 10)
- (1-2, 3-4), (3-4, 5-6), (5-6, 7-8), (7-8, 9-10)
- (1-2-3, 4-5-6), (4-5-6, 7-8-9)

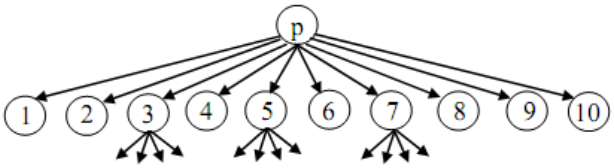

Abbildung 4: Knotenkombinationen

Ähnliche benachbarte Kombinationen werden zu einer Datenregion zusammengefasst. Wenn beispielweise der Vergleich (1,2) ähnlich ist, dann sind die Knoten 1 und 2 Datensätze und gehören zu einer Datenregion. Falls der Vergleich (2,3) auch ähnlich ist, dann ist der Knoten 3 ein Datensatz in der Datenregion, in der Knoten 1 und 2 auch gehören.

3.2 Identifizieren von Veranstaltungsdatensätzen

Die gewonnenen Datenregionen enthalten Datensätzen. Man weiß aber nicht, wo sich mögliche Veranstaltungsdatensätzen befinden. Deshalb werden die einzelnen Datenregionen mittels einer Score-Funktion bewertet [6, 7] um die Datenregionen mit Veranstaltungen zu identifizieren. Die Funktionsweise der Score-Funktion ist jener aus der Studienarbeit ähnlich, wird jedoch nicht mehr auf Tabellen sondern auf Datensätzen angewendet. Die Datenregion mit der höchsten Bewertung beinhaltet mit großer Wahrscheinlichkeit Veranstaltungen.

3.3 Extrahieren von Informationen aus den Veranstaltungsdatensätzen

Die Extraktion von Veranstaltungsinformationen aus den Datensätzen erfolgt zum Einen über reguläre Ausdrücke. Veranstaltungsattribute wie Datum, Zeit, Genre oder Preise können als regulärer Ausdruck beschrieben und in den Texten der einzelnen Datensätzen gesucht werden. Etwas schwieriger ist die Bestimmung von Ort und Bandnamen. In der Diplomarbeit beschränken wir uns auf Lokationswebseiten, daher bekommen wir den Ort von der Webseite mitgeliefert.

Für die Bestimmung von Bandnamen in einem Datensatz werden folgende Ansätze verwendet:

1. Hooolp hat bereits eine Band-Datenbank mit mehreren Tausend Datensätzen aufgebaut. Für die Bestimmung von Bandnamen wird zuerst überprüfen, ob aus den

[2] http://www-igm.univ-mlv.fr/~lecroq/seqcomp/node2.html

restlichen Informationen, die noch nicht extrahiert wurden, einzelne Wörter oder Wortkombinationen in der Datenbank zu finden sind.
2. Außerdem kann versuchen werden, Wörter oder Wortkombinationen aus den restlichen Informationen in MySpace zu suchen. Falls es Treffer existiert, ist es ein Indiz für eine Band.
3. Weiterhin werden die Wörter oder Wortkombinationen auf ihre visuelle Darstellung hin analysiert. Bandnamen können auf der Webseite fettmarkiert sein oder hinter einem Link stehen.

Wörter oder Wortkombinationen aus den restlichen Informationen werden unter den oben genannten Gesichtspunkten untersucht und gewichtet. Bei bestimmten Grenzwerten können Wörter als Bandname markiert werden. Andere Wörter unter dem Grenzwert werden als Vorschlag angezeigt.

Wörter, die zweifelsfrei als Bandname identifiziert wurden, können in die Banddatenbank eingefügt werden, sodass für eine weitere Überprüfung bessere Ergebnisse zu erwarten sind.

3.4 Indizieren von Veranstaltungsseiten

Bei diesem Schritt wird versucht Veranstaltungsseiten, die vorher noch manuell ermittelt wurden, automatisch aus einer Lokationseite zu identifizieren. Dieser Schritt ist notwendig, da die Links zu Veranstaltungsseiten nicht immer statisch sind. Damit wird festgelegt von welcher Seite die Veranstaltungen geholt werden sollen. Für das Indizieren werden folgende Schritte unternommen:

1. Unterseiten von einer Site, die möglicherweise Veranstaltungen enthalten können, werden gecrawlt.
2. Diese Unterseiten werden bewertet, sodass die Seite mit der höchsten Bewertung als die Veranstaltungsseite markiert wird. Für die Bewertung einer Unterseite müssen dessen Datenregionen bewertet werden (siehe 3.2). Den höchstbewertete Datenregionen entspricht dann die Bewertung der Unterseite.
3. Zum Schluss muss noch überprüft werden, ob aus die als „Veranstaltungsseite“ markierte Seite Veranstaltungsinformationen zu extrahieren sind.

3.5 Evaluation der Ergebnisse

Die Ergebnisse aus 3.3 und 3.4 werden anhand von Exemplaren aus der Lokationdatenbank manuell überprüft. Vergleiche werden zu den Ergebnissen der Studienarbeit gezogen und eine genaue Fehleranalyse wird erstellt.

4 Literatur

[1] Liu Bing, Robert Grossman & Yanhong Zhai, “Mining data records from Web pages”, KDD-03, 2003

[2] Yanhong Zhai & Liu Bing, “Web data extraction based on partial tree alignment”, WWW’05, 2005

[3] Wei Liu, Xiaofeng Meng & Weiyi Meng, „Vision-based Web Data Records Extraction“, 9th SIGMOD Int’l Workshop on Web and Databases (WebDB 2006), ACM Press, 2006

[4] D. Buttler, L. Lie & C. Pu, „A fully automated extraction system for the world wide web“, in IEEE ICDCS-21, April 2001

[5] Hongkun Zhao, Weiyi Meng, Zonghuan, Vijay Raghavan & Clement Yu, „Fully automatic wrapper generation for search engines“, WWW’05, 2005

[6] Theodore W. Hong & Keith L. Clark, “Towards a Universial Web Wrappers”, AAAI Press, FLAIRS Conference, 2004

[7] Hung Le, “Identifizieren und Extrahieren von Musikveranstaltungen aus dem Web”, HU-Berlin, Studienarbeit, 2009