Data Mining im e-commerce am Beispiel der Deutschen Bahn AG

Katja Steuernagel
Universität Kaiserslautern
Momentan: GIP AG
Göttelmannstraße 17
55130 Mainz
katja@katja-steuernagel.de

Zusammenfassung

Die Deutsche Bahn AG bietet einen Großteil ihrer Fahrkarten bereits im Internet zum Kauf an. Die vorliegende Diplomarbeit versucht, diese Käufer zu klassifizieren und weitergehende Aussagen über den Verlauf der Verkaufstransaktion zu treffen. Dabei kommen Methoden der Sequenz- und Clusteranalyse sowie Neuronale Netze und Entscheidungsbaumverfahren zum Einsatz.

Keywords: Web Mining, Clusteranalyse, Sequenzanalyse, SAS WebHound, SAS Enterprise Miner.

1 Problemstellung

Die Deutsche Bahn AG bietet einen Großteil ihrer Fahrkarten bereits im Internet zum Kauf an. Diese Möglichkeit wird jedoch nur von einem geringen Prozentsatz der Besucher der Webseiten genutzt. Diese Diplomarbeit beschäftigt sich mit der Frage, wie man die Käufer bzw. Nicht-Käufer charakterisieren kann und welche Aussagen über den Verlauf der Verkaufstransaktion zu treffen sind.

Es wird untersucht, welche Pfade die Benutzer auf den Webseiten genommen haben und welche davon besonders häufig zu einem erfolgreichen Ende oder auch einem Abbruch führten.

Mit Hilfe dieser gewonnenen Erkenntnisse und weiterer Daten zu den gekauften Fahrkarten wird dann versucht, eine Einteilung der Besucher in Segmente vorzunehmen.

2 Vorhandene Daten

Zur Analyse werden aus dem Webangebot der Deutschen Bahn AG zwei Transaktionen herausgegriffen und miteinander verglichen.

Dies ist zum Einen das Angebot „Surf&Rail“. Dabei handelt es sich um ausschließlich im Internet buchbare Fahrkarte zu speziellen Konditionen.

Zum Anderen werden Daten aus dem Fahrkartenshop zur Analyse herangezogen. Hier kann man beliebige Fahrkarten zum Normalpreis (keine Sonder- und Pauschalangebote) der Deutschen Bahn AG erwerben.

Bei beiden Transaktionen sind Daten zu folgenden Bereichen aus dem Januar und Februar 2002 vorhanden:

- Start- und Zielpunkt der Reise, Datum
- Gebuchte Zugnummer, Klasse, Sitzplatz (Großraum, Abteil usw.), Besonderheiten (Handyzone, Ruhezone), Nichtraucher- oder Raucherbereich
- Datum und Uhrzeit der Buchung, IP-Adresse und User Agent des Surfers
- (nicht) erfolgreiches Ende der Buchung
- Anzahl der gemeinsam buchenden Personen, bei „Surf&Rail“ auch Vorhandensein einer Bahncard
- Verweildauer auf den einzelnen Webseiten, Ein- und Ausstiegsseiten

Diese Angaben reichen für ein erfolgreiches Data Mining wohl noch nicht aus und werden deshalb durch weitere Daten aufbereitet. Dazu gehören beispielsweise:

- Bundesland und Einwohnerzahl des Start- bzw. Zielortes
- Zur benutzten Zugnummer gehöriger Zugtyp (ICE, IC usw.)
- Wochentage der benutzten Daten
- Per DNS-Reverse-Lookup aus der IP-Adresse ermittelte Angaben (Surfer surft über den Zugang eines Providers, einer Universität o.ä.)

Problematisch ist vor allem die große Anzahl der möglichen Ausprägungen eines Attributes. So können bei der Deutschen Bahn AG Fahrkarten zu etwa 6000 verschiedenen Bahnhöfen innerhalb Deutschlands gekauft werden. Diese

Angabe des Start- und Zielbahnhofs in eine für das Data Mining verwendbare Größe zu transformieren ist beispielsweise eine Aufgabe, die bewältigt werden muss.

Leider kann aufgrund des Fehlens von eindeutigen Identifikationsmerkmalen (Cookies, CGI-Parameter etc.) der Benutzer nur während der Session verfolgt werden. Eine sessionübergreifende Wiedererkennung bereits „bekannter“ Benutzer ist damit nicht möglich.

3 Datenanalyse

Zur Einlesen der Logdaten im Extended Log File Format wird der SAS WebHound 4.0 verwendet. Die Logs im serverspezifischen Format werden mittels eigener Perl-Skripten eingelesen und danach in Base SAS mit den bereits eingelesenen Logdaten verknüpft.

Mit dem SAS WebHound 4.0 wird außerdem noch die Untersuchung der Benutzerpfade durchgeführt. Für das eigentliche Data Mining kommt der SAS Enterprise Miner zum Einsatz.

4 Erste Ergebnisse

Ein Vergleich der angebotenen und nachgefragten „Surf&Rail“-Verbindungen ergab folgendes Bild:

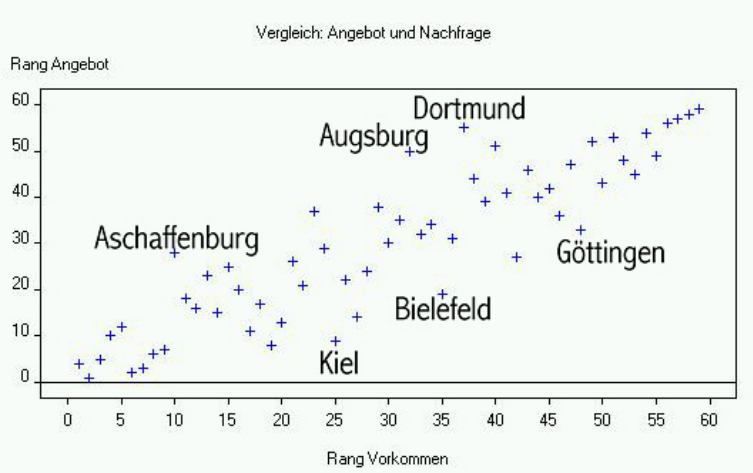

Abbildung 1: Vergleich von Angebot und Nachfrage der Verbindungen bei „Surf&Rail“

Man kann leicht erkennen, daß es einen relativ linearen Zusammenhang zwischen den angebotenen und benutzten Zugverbindungen gibt. Nur einige Ausreißer wie beispielsweise Dortmund und Augsburg mit einer relativ guten Nachfrage bei einem schlechten Angebot sowie Kiel und Bielefeld mit einer relativ schlechten Nachfrage bei einem guten Angebot sind vorhanden.

Weiterhin konnten mit dem SAS WebHound die Benutzerpfade visualisiert werden, wie dies beispielhaft in der untenstehenden Grafik gezeigt wird:

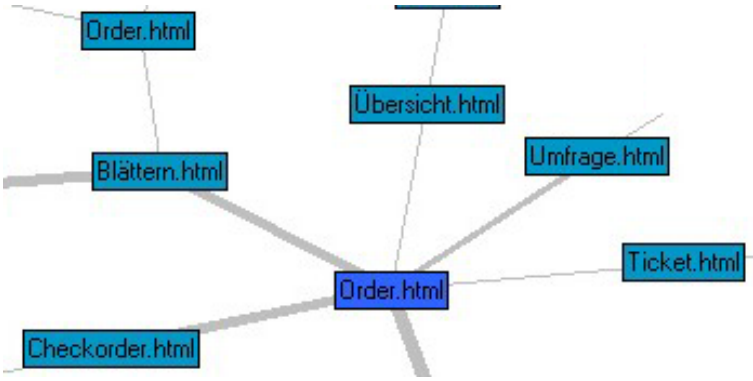

Abbildung 2: Visualisierung der Benutzerpfade

Dabei ist deutlich zu sehen, daß ein Großteil der Besucher des Webangebotes sehr lange in den Angeboten blättern muß, um die passende Verbindung zu finden. Allerdings erreicht ein sehr großer Anteil der Surfer die Buchung eines Tickets.

Weiterhin wurde eine erste Einteilung der Käufer der „Surf&Rail“-Fahrkarten in Cluster durchgeführt. Dies führte zu folgendem Ergebnis:

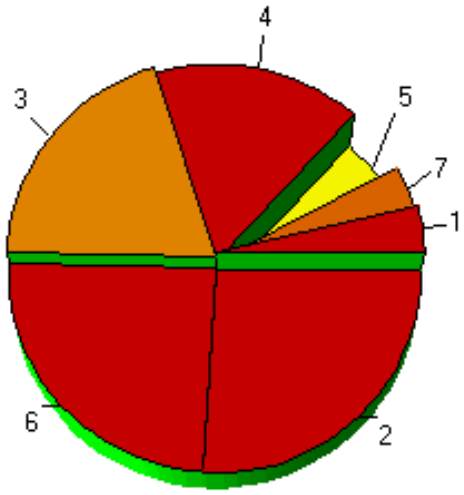

Abbildung 3: Erstes Ergebnis einer Clusteranalyse der Käufer

Bei dieser Analyse wurden 7 Cluster gefunden, die eine Einteilung der Käufer ermöglichen. Dabei steht beispielsweise ein Cluster für die allein reisenden Kunden, die nur wenige Tage unterwegs sind und vorzugsweise den ICE benutzen. In einem anderen Cluster finden sich dann Gruppenreisende, die meist übers Wochenende wegfahren und auf keinen bestimmten Zugtyp festgelegt sind. Diese Clusterung ist aber noch zu verbessern, da nicht alle möglichen Attribute der Käufer berücksichtig wurden. Eine Einbeziehung weiterer Daten und eigener Algorithmen sollte obiges Ergebnis noch verbessern.

5 Ausblick

Die Diplomarbeit befindet sich momentan noch in der Startphase. Deshalb werden in nächster Zukunft noch weitere Ergebnisse dazukommen. Spätestens ab Juli 2002 werden diese auch in der schriftlichen Ausarbeitung auf meiner Webseite (http://www.katja-steuernagel.de) verfügbar sein. Aufgrund einiger Probleme mit dem Webdesign der Deutschen Bahn AG könnte eine leichte Veränderung des Aufbaus der Buchungsdialoge ein deutlich besseres Data Mining ermöglichen. Eine bessere Verfolgbarkeit der Benutzerpfade sollte dann eine bessere Segmentierung des Kundenverhaltens ermöglichen.

Literatur

[1] Pitkow, J. (1997). In search of reliable usage data on the WWW. In: Sixth International World Wide Web Conference. pages 451-463. Santa Clara, CA

[2] Hippner, H., Küsters, U., Meyer, M. und Wilde, K. (2001). Handbuch Data Mining im Marketing. Vieweg Verlag. Braunschweig/Wiesbaden

[3] Enterprise Miner:
`www.sas.com/offices/europe/germany/solutions/customer.html`

[4] Webhound:
`www.sas.com/offices/europe/germany/solutions/`
`e-intelligence.html`

[5] Jäger, B. , Wodny, M. und Rudolph, P.E. (2001). Clusteranalyse mit Binärdaten. Vortrag KSFE 2001. Universität Hohenheim.