Einsatz ausgewählter Data-Mining-Methoden zur Web-Nutzungsanalyse für die PC-Ware Information Technologies AG, Leipzig

Dipl-Kfm. (FH) Sebastian Reichmann, Prof. Dr. Klaus Kruczynski
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Zusammenfassung

In der vorliegenden Studie wird ein Referenzmodell für die Anwendung von Data-Mining-Methoden zur Analyse der Nutzung von Webseiten vorgestellt. Nach der Diskussion wesentlicher Merkmale des Web Usage Mining wird der Methodeneinsatz anhand eines konkreten Projekts bei einem IT-Dienstleister demonstriert. Im ersten Teil der Arbeit wird nach einer kurzen Einführung in die Bereiche Data Mining und Web Mining ein Referenzmodell für Web Usage Mining abgeleitet. Dem schließt sich in Teil zwei die Beschreibung der praktischen Umsetzung des entwickelten Vorgehensmodells an. In Zusammenarbeit mit der PC-Ware Information Technologies AG Leipzig wurde die Website des Unternehmens einer Untersuchung des Nutzungsverhaltens unterzogen. Dabei fanden sowohl verschiedene künstliche neuronale Netze als auch das Klassifikationsbaumverfahren C5.0 Anwendung. Zur Realisierung dieser Analysen wurde die Data-Mining-Software SPSS Clementine eingesetzt.

Primäres Ziel der Untersuchung war die Steigerung des Nutzungsverständnisses für die ausgewählten Webseiten. Besonders die speziellen Aufgaben bei der Aufbereitung webgenerierter Daten, aber auch die Anforderungen der ausgewählten Data-Mining-Methoden an Form und Struktur der Ausgangsdaten machten die Datenaufbereitungsphase zu einem besonderen Arbeitsschwerpunkt. Darauf aufbauend erfolgte die Anwendung der verschiedenen Analysemethden.

Im Ergebnis der Untersuchung konnten wesentliche Einflussfaktoren auf das Verhalten der Webnutzer aufgedeckt werden. Daraus wurden Handlungsempfehlungen für die Optimierung der vorhandenen Online-Services und die Ausgestaltung spezieller Marketingaktivitäten abgeleitet.

1 Einführung

Durch seine Konzeption verfügt das Internet bereits über die wesentlichen Voraussetzungen für die Aufzeichnung und Speicherung aller darin stattfindenden Aktivitäten.

Die Speicherung dieser Daten stellt jedoch nur eine notwendige Voraussetzung für die Nutzung des darin enthaltenen Wissens und damit für das Durchdenken des Geschehenen dar. Insbesondere das Volumen der generierten Datenbestände erweist sich heute zunehmend als limitierender Faktor für ihre sinnvolle Nutzbarkeit. Konventionelle Analyseverfahren sind diesen Datenmengen meist nicht mehr gewachsen und können das „Mehr“ an vorhandenen Daten nicht in ein „Mehr“ an Information und Wissen umsetzen.

Die Anwendung von Data Mining bietet einerseits die Möglichkeit, großvolumige und komplexe Datenbestände sinnvoll nutzbar zu machen und andererseits für den Anwender potenziell neues Wissen daraus zu gewinnen. Im Sinne von Knowledge Discovery in Databases (KDD) wird Data Mining häufig als Prozess definiert, der automatisiert bzw. halb automatisiert interessante Muster in großen Datenbeständen identifiziert. Aufgrund dieser Konstellation stellt das Internet eines der viel versprechendsten Anwendungsfelder für Data Mining dar. Besonders die von einem Webserver generierten Log-Files bieten eine hervorragende Ausgangsbasis für Marketinganalysen, da in ihnen das Verhalten der Webnutzer fast lückenlos nachvollziehbar ist.

Vielfach wird für die Anwendung von Data-Mining-Methoden auf webgenerierte Daten der Begriff Web Mining verwendet. In Abhängigkeit vom jeweiligen Untersuchungsziel lässt sich Web Mining in die Teilbereiche Web Content Mining und Web Usage Mining unterteilen. Während im Fokus des Web Content Mining die Analyse des Inhaltes von Webseiten und der semantischen Beziehungen zwischen ihnen steht, beschäftigt sich Web Usage Mining mit der Untersuchung des Nutzungsverhaltens von Webseiten, die im Falle des personalisierten Web-Usage-Mining mit weiteren Informationen über den Nutzer angereichert werden [Spiliopoulou, 2001].

Nach der formalen Einordnung des Web-Mining-Begriffes beschreiben die folgenden Darstellungen ein konkretes Projekt. Ronny Kohavi, Director of Data Mining, Blue Martini Software [Kohavi, 2001], charakterisiert das Wesen von Web-Mining-Projekten wie folgt: *„Mining E-Commerce Data: The Good, the Bad, and the Ugly.*

Electronic Commerce provides all the right ingredients for successful data mining (the Good). Web logs, however, are at a very low granularity level, and attempts to mine e-commerce data only using web logs often result in little interesting insight (the Bad). Getting the Data into minable formats requires significant pre-processing and data transformations (the Ugly).”

2 Referenzmodell CRISP-DM für Web Usage Mining

Als Grundlage für die Realisierung des Projektes diente das CRISP-DM-Vorgehensmodell (Cross-Industry Standard Process for Data Mining). Dieses Modell wurde im Rahmen eines gleichnamigen Gemeinschaftsprojektes von Industrieunternehmen, Herstellern von Data-Mining-Produkten und Anbietern von Datenbanklösungen unter Schirmherrschaft der Europäischen Kommission entwickelt und inzwischen bereits vielfach erfolgreich eingesetzt. Es bildete unter anderem auch eine Grundlage bei der Konzeption der innerhalb des vorliegenden Projektes verwendeten Data-Mining-Software Clementine von SPSS. Ein Data-Mining-Projekt wird nach dieser Methodik in sechs Phasen unterteilt, die konzeptionell unabhängig vom jeweiligen Anwendungsproblem gestaltet sind, um so eine breite Verwendbarkeit des Modells zu gewährleisten. Das Modell besteht aus den Phasen: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment. Die Stärken dieses Ansatzes liegen besonders im umfassenden Charakter sowie in seiner Flexibilität. So obliegt es dem Anwender, ob er nach Abschluss einer Phase zur nächsten übergeht, sie wiederholt durchläuft oder wieder zu einer vorgelagerten Phase zurückkehrt. Der Analyseprozess gestaltet sich auf diese Weise dynamisch aus sich selbst heraus, was den allgemeinen Charakter von Data-Mining-Projekten sehr gut widerspiegelt.

Für die speziellen Aufgaben des Web Usage Mining erwies es sich als zweckmäßig, das CRISP-DM-Modell anzupassen und zu konkretisieren [Reichmann, 2003]. Der Gesamtprozess des CRISP-DM-Modells wurde zunächst in die zwei Hauptprozesse Analysevorbereitung und Analyse unterteilt. Dadurch wurde der besonderen Bedeutung der Datenvorbereitung innerhalb von Web-Usage-Mining-Projekten Rechnung getragen, die auch innerhalb des durchgeführten Projektes ca. zwei Drittel des gesamten Arbeitsaufwands ausmachte. Zudem wurde ein Arbeitsschritt „Konzepthierarchie“ in die Datenvorbereitungsphase eingeführt. Dieser Teilschritt diente der Aggegration der Vielzahl von Einzelseiten zu übergeordneten Kategorien, wodurch vor allem die Komplexität der Analyse für den Anwender reduziert werden sollte. Jedoch wurde dadurch auch die Anzahl der zu betrachtenden Dimensionen der Datenbasis verringert, was sich positiv auf die Performance der verwendeten Algorithmen und die benötigte Menge an Trainingsdaten auswirkte. Die Transaktionsableitung bzw. die Zusammenfassung einzelner Seitenabrufe eines Nutzers zu Sessions erfolgte innerhalb des Schrittes „Daten konstruieren“. Die Analyse der gefundenen Muster und gebildeten Modelle erfolgte innerhalb eines zweistufigen Prozesses: zunächst unter Data-Mining-Aspekten in der Prozessphase „Modellierung“ und darauf aufbauend auf betriebswirtschaftlicher Ebene innerhalb der Prozessphase „Bewertung“.

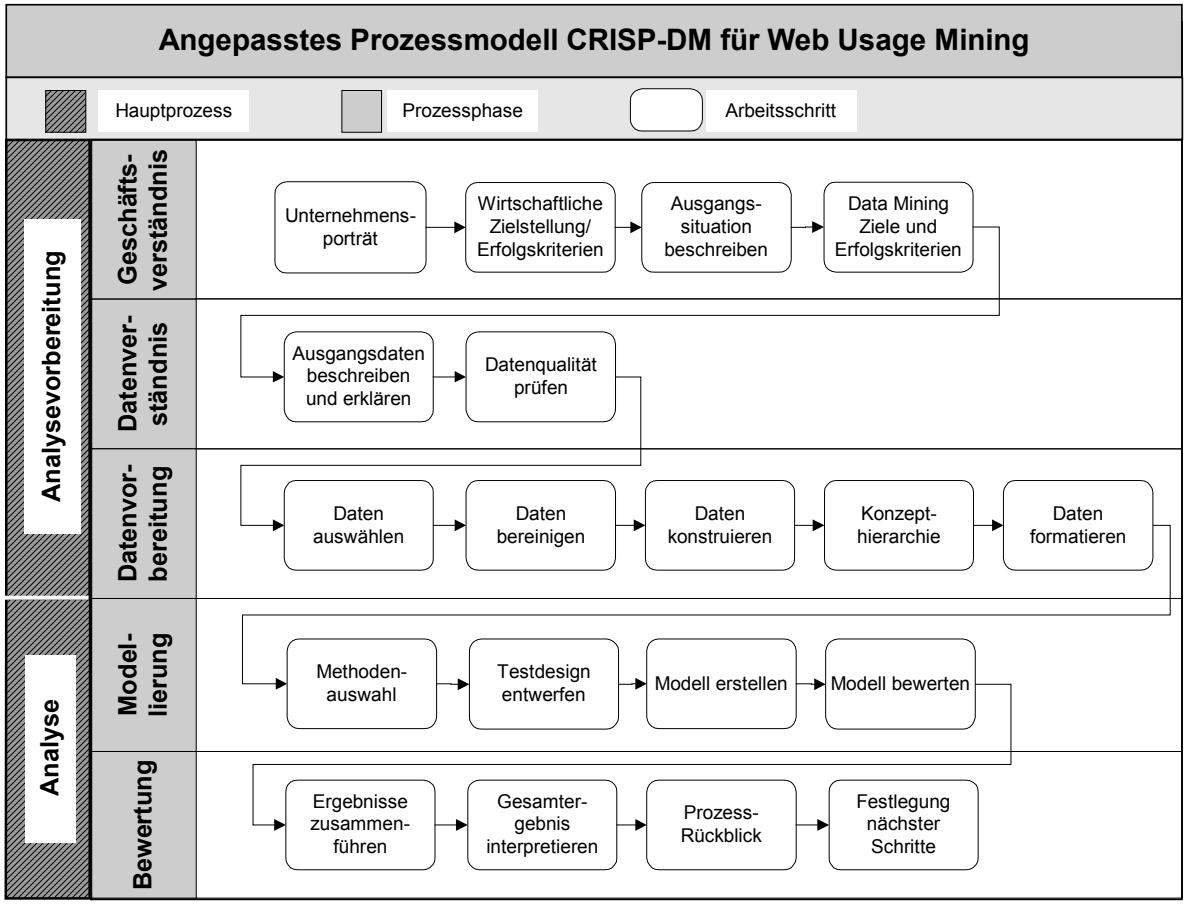

Bild 1: Referenzmodell CRISP-DM für Web Usage Mining

3 Analysevorbereitung

Grundlage der Untersuchung bildeten die Log-Dateien des Webservers der PC-Ware Information Technologies AG mit Stammsitz in Leipzig. Dieser IT-Dienstleister ist seit seiner Gründung im Jahre 1990 zu einem internationalen Konzern mit zahlreichen deutschen Niederlassungen und Tochtergesellschaften in mehreren Staaten Europas gewachsen. Auf dem Gebiet der Softwarelizenzierung gehört das Unternehmen heute zu den Marktführern in Europa. Den mehr als 48.000 vertraglich gebundenen Kunden bietet PC-Ware ein breites Spektrum an speziellen Dienstleistungen und Komplettlösungen aus verschiedenen Bereichen des IT-Sektors. Ihr Kundenstamm umfasst industrielle Großkunden, mittelständische Unternehmen und Einrichtungen des öffentlichen Dienstes. Die Realisierung dieser Leistungen erfolgt in den vier Geschäftsbereichen Software Sales & Licensing (SSL), Software Support & System Management (S^3M bzw. SSS), Integrated System Solutions & Support (ISS) und Enterprise Solution (ES).

In den betreffenden Dateien waren alle Zugriffe aufgezeichnet, die in der Zeit vom 1. bis 31. Oktober 2002 auf die Hauptseiten des Unternehmens erfolgten. Jeder Zugriff wurde als gesonderter Eintrag in den Log-Dateien aufgezeichnet. Im Untersuchungszeitraum kam ein Microsoft Internet Information Server 5.0, Version 1.0 zur Anwendung. Die Einträge der Server-Transfer-Log-Datei wurden von diesem Server im Common-Log-Format gespeichert und enthielten im Untersuchungszeitraum insgesamt mehr als 1,4 Millionen Datensätze.

3.1 Untersuchungsziele und Methoden

Das erste Hauptziel der Untersuchung bestand in der Abbildung der bei der Seitennutzung entstandenen Nutzungsmuster, da diese Rückschlüsse auf spezielle Ziele und Informationsbedürfnisse bestimmter Nutzergruppen zulassen und damit eine wertvolle Informationsbasis für die Ausgestaltung individueller Marketingmaßnahmen darstellen. Ein Muster im Sinne der Untersuchung stellt dabei einen Ausdruck dar, der eine Teilmenge von Daten aus der Ausgangsmenge beschreibt. In diesem speziellen Fall ergibt sich ein solches Muster durch die Kombination aus unterschiedlichen Seitenabrufen, die einem Besucher während eines Besuches auf den Webseiten zuzuordnen sind. Ähnliche Nutzungsmuster können zu Mustergruppen zusammengefasst werden. Diese erste Teilaufgabe stellte eine typische Clusterungsaufgabe dar, da ohne Vorgabe von bestimmten Kriterien eine Gruppeneinteilung erreicht werden sollte.

Besonders interessante Muster bzw. Mustergruppen wurden daran anschließend detaillierter untersucht. Ziel war es, die inhaltlichen Merkmale aufzudecken, die für die Zuordnung der enthaltenen Datensätze bzw. Sessions zu einer bestimmten Musterklasse verantwortlich waren. Dieser Schritt entspricht einer Klassifizierung, wobei jedoch nicht die Klassifikation neuer Daten angestrebt wurde, sondern die Analyse des Klassifikationsmodells.

Im zweiten Teil der Untersuchung wurde eine Klassifikation von Seitenbesuchern in Abhängigkeit von ihrer Kaufaffinität angestrebt. Es sollten die wesentlichen Unterschiede im Nutzungsverhalten in Abhängigkeit von der Kaufaffinität eines Besuchers innerhalb eines Modells abgebildet werden. Im Speziellen sollte damit ein Klassifikationsmodell erstellt werden, welches Aussagen über die Wahrscheinlichkeit eines Besuches des Shop-Bereiches (SSL-Shop) der Webseite aufgrund der zuvor besuchten Seiten ermöglicht.

3.2 Datenvorbereitung

Ziel dieser Projektphase war es, die vorliegenden Ausgangsdaten derart zu bearbeiten, dass sie eine geeignete Grundlage für die Anwendung der ausgewählten Data-Mining-Methoden darstellen. Dazu zählten die Teilschritte:

- Datenauswahl,
- Datenbereinigung,
- Datenkonstruktion und
- Datenformatierung.

Nach der Auswahl der für die Untersuchung benötigten Log-Dateien wurde innerhalb der Datenbereinigung versucht, alle nicht benötigten oder ungeeigneten Seitenabrufe aus den Log-Dateien zu entfernen. Mit dem Ziel der Performancesteigerung erfolgte innerhalb dieses Bereinigungsprozesses die sukzessive Eliminierung nicht mehr benötigter Log-Datei-Felder.

Vor allem waren dabei die Einträge zu entfernen, die Fehler in der Datenübertragung enthielten, zusätzlich zu einem normalen Seitenabruf automatisch durch den Server generiert wurden (z.B. Bilddateien), von unternehmenseigenen Rechneradressen stammten oder von Suchrobotern erzeugt wurden. Als Resultat der Datenbereinigungsphase ergab sich eine Datei mit 73.033 Datensätzen und damit lediglich noch ca. fünf Prozent der Ausgangsdaten. Dass dieser relativ geringe Anteil an verwertbaren Datensätzen durchaus nicht unüblich ist, zeigen die Ergebnisse verschiedener vergleichbarer Analysen [Rahm und Stöhr, 2003].

Für die weitere Arbeit mit den Ausgangsdaten war im nächsten Schritt die Konstruktion verschiedener Werte nötig. Zum einen erfolgte die Aggegration vorhandener Attribute zur Steigerung ihrer Aussagekraft, wie beispielsweise die Aggegration eines eindeutigen Zeitstempels aus den Datum- und Uhrzeit-Feldern der Log-Einträge, zum anderen die Ableitung neuer Felder. Einen zentralen Schritt innerhalb jedes Web-Usage-Mining-Projektes stellt die Transaktionsableitung bzw. das Sessioning dar. Über einen eindeutigen Schlüssel werden dabei alle Seitenabrufe eines Nutzers, die dieser während eines Besuches auf den Web Seiten tätigt, zusammengefasst. Bei der Mehrzahl der heute verwendeten Webserver geschieht dies bereits automatisch während eines Besuches. Im vorliegenden Fall war es jedoch notwendig, dieses Schlüsselfeld (Session-ID) nachträglich zu erzeugen. Hierzu musste zum einen überprüft werden, ob es sich jeweils um den gleichen Nutzer handelte und zum anderen, ob die Zeitdifferenz zwischen den Seitenabrufen einen Grenzwert nicht überstieg, der als Indikator für den inhaltlichen Zusammenhang der Abrufe verwendet wurde.

Zur Reduzierung der Komplexität und zur Steigerung der Untersuchungstransparenz erfolgte im Schritt Konzepthierarchie die Zusammenfassung der im Webauftritt enthaltenen Einzelseiten. Allein in den Aufzeichnungen des Untersuchungszeitraumes wurden mehr als

1200 abgerufene Einzelseiten verzeichnet. Gemäß der Grundstruktur der Website wurde eine Unterteilung in vier Hauptbereiche (SSL, SSS, ISS und allgemeine Seiten) vorgenommen. Diese Hauptbereiche wurden danach wiederum in verschiedene Seitenkategorien untergliedert, wobei ebenfalls die vorhandene Hierarchie der Site als Vorlage diente. Die entstandene Konzepthierarchie enthielt noch 37 einzelne Seitenkategorien. Durch diese Dimensionsreduktion konnten die Komplexität für die weitere Untersuchung und die Menge an benötigten Trainingsdaten erheblich vermindert werden [Pyle, 1999], wobei die semantische Grundstruktur der Site nahezu erhalten blieb.

In der abschließenden Datenformatierung mussten die aufgezeichneten Seitenbesuche in binäre Werte umgewandelt werden, da viele Data-Mining-Methoden entweder ausschließlich solche Werte verarbeiten oder damit bessere Ergebnisse liefern [Witten und Frank, 2000]. Um dies realisieren zu können, wurden alle enthaltenen Seitenkategorien in beschreibende Merkmale umgewandelt. Ob die Seite abgerufen wurde oder nicht, wurde durch binäre Schlüssel angezeigt. Nach der Aggegration der Seitenabrufe innerhalb einer Session zu einem Datensatz stellte dieser einen Vektor dar, dessen Komponenten die Seitenkategorien widerspiegelten.

Eine weitere Steigerung der Aussagekraft wäre durch das Hinzufügen der Besuchsanzahl einer Seite während einer Session erreichbar gewesen. Da jedoch keine genauen Erkenntnisse darüber vorlagen, wie oft der wiederholte Abruf einer Seite aus dem Cache des Browsers oder eines vorgeschalteten Proxy-Servers erfolgte und damit vom Server nicht registriert wurde, kam diese Möglichkeit nicht in Betracht. Im Interesse der Erweiterung der Anwendungsmöglichkeiten und der Erhöhung der Aussagekraft der Analyse könnte auch die Abfolge der Seitenabrufe in Form von Sequenzen in die Analyse einbezogen werden. Da dies aber ebenfalls durch die beschriebene Caching-Problematik stark beeinträchtigt wird, wurde auch auf diese Erweiterung verzichtet.

4 Analyse

Im zweiten Hauptprozess des Referenzmodells erfolgten die Auswahl und die Anwendung geeigneter Data-Mining-Methoden sowie die Bewertung und Interpretation der erzielten Ergebnisse.

4.1 Methodenauswahl und Modell erstellung

Die im ersten Untersuchungsziel angestrebte Clusterung wurde mit einem künstlichen neuronalen Netz vom Typ Kohonen gelöst. Die wesentlichen Merkmale dieses Ansatzes stellen die Verwendung des unsupervised learning und die Einbeziehung der geografischen Anordnung der Neuronen in die Analyse dar [Kohonen, 2001]. Das Grundkonzept des Ansatzes liegt in der niederdimensionalen Abbildung von Strukturen aus höherdimensionalen Datenräumen [Lämmel und Cleve, 2001]. Die speziellen Eigenschaften dieses Ansatzes erschienen für die geplante Untersuchung besonders wertvoll. Hervorzuheben sind hierbei vor allem die Fähigkeit zur

- niederdimensionalen Abbildung höherdimensionaler Datenstrukturen,
- Robustheit bzw. Fehlertoleranz und die
- Hypothesenfreiheit dieser Methode.

Besonders unter Berücksichtigung von Volumen und Komplexität der Ausgangsdaten sollte die angestrebte Clusterung mit dieser Methode in geeigneter Weise realisiert werden können.

Für die Klassifizierung der in den Clustern enthaltenen Datensätze wurde ein C5.0-Klassifikationsbaumalgorithmus nach Quinlan verwendet, da damit die relative Bedeutung der einzelnen Merkmale bzw. Seitenabrufe für die Zuordnung eines Datensatzes zu einer Mustergruppe ersichtlich wird. Die Darstellung des Ergebnisses solcher Algorithmen kann in einer Baumstruktur erfolgen, wodurch sich ihre Interpretierbarkeit deutlich steigern lässt.

Das angestrebte Klassifikationsmodell wurde mit einem Backpropagation-Netz erstellt, da dieses aufgrund seiner Eigenschaften besonders gut geeignet erschien, die in den Daten vorliegende Grundstruktur aufzudecken und wesentliche Unterschiede im Verhalten unterschiedlicher Nutzergruppen abzubilden. Durch die Robustheit und die relativ große Fehlertoleranz künstlicher neuronaler Netze sollten eventuell noch vorhandene Inkonsistenzen in den Ausgangsdaten aufgefangen werden können. Backpropagation-Netze zeichnen sich zudem besonders durch eine relativ einfache Netzkonfiguration aus, was sich positiv auf ihre Handhabung bei ihrer Erstellung und ihre Verständlichkeit auswirkt [Björn, 1997].

Nach der Auswahl der zu verwendenden Modelle musste ein Testdesign für das Modelltraining erstellt werden. In einem ersten Schritt wurde dazu aus der Ausgangsdatenmenge eine Evaluierungsmenge aus zufällig ausgewählten Datensätzen gebildet. Die verbliebenen Daten dienten als Trainings- und Testdaten. Begründet durch die unterschiedlichen Anforderungen an das Training der einzelnen Modelle, mussten jeweils unterschiedliche Konfigurationen der ausgewählten Trainings- bzw. Testmengen vorgenommen werden. Beispielsweise wurde für das Training des Backpropagation-Modells eine Balancierung der Trainingsdaten in Bezug auf die Ausprägung der Zielvariable vorgenommen. Das war erforderlich, um das Netz nicht zu stark auf das Profil negativer Beispiele (SSL-Shop wurde nicht besucht) auszurichten, die erheblich häufiger auftraten. Netze, die mit einer ausgewogeneren Datenstruktur trainiert werden, liefern zudem häufig auch bessere Klassifizierungsergebnisse [Björn, 1997].

In das Modelltraining wurden lediglich die erstellten Seitenkategorien einbezogen. Weitere Werte, wie beispielsweise die Session-ID oder die IP-Adresse, wurden nicht verwendet, da diese keine verhaltensrelevanten Größen darstellen. Zusätzlich musste für die Klassifikationsbäume und Backpropagation-Netze eine Zielgröße für das Training bestimmt werden, die ebenfalls nicht als beschreibendes Merkmal (Attribut) in die Modellerstellung einfloss.

Bei der Modellerstellung wurden jeweils mehrere Modelle mit variierenden Parametereinstellungen generiert und anschließend verglichen. Beispielsweise ergaben sich beim Training des Kohonen-Netzes sinnvolle Variationsmöglichkeiten durch unterschiedliche Festlegungen für die Anzahl der Netzneuronen, die Lernrate und das Abbruchkriterium des Trainings. Zusätzlich

dazu wurden die in Clementine verfügbaren automatischen Trainingsmethoden genutzt, die dem Nutzer programmgesteuert optimierte Netzvarianten vorschlagen.

4.2 Ergebnisse der Untersuchung

Nach Abschluss des Modelltrainings für das Kohonen-Netz erwies sich eine Netzvariante mit 12*8 Netzneuronen am geeignetsten, da die darin abgebildeten Mustergruppen bereits deutlich unterscheidbar waren und zugleich noch über eine ausreichende Größe verfügten.

Im nächsten Schritt wurden besonders auffällige Mustergruppen (Cluster) ausgewählt, für die eine weiterführende Untersuchung sinnvoll erschien. In der Ergebnisdarstellung (Bild 2) war die isolierte Belegung eines Neurons mit den Koordinaten $KX-Kohonen = 2 und $KY-Kohonen = 5 besonders auffällig. Der betreffende Bereich deutete darauf hin, dass sich die enthaltene Mustergruppe sehr stark von den übrigen Mustern unterscheidet. Dieses Cluster wurde unter der Bezeichnung Bereich 1 für die weitere Analyse ausgewählt. Der zweite deutlich abgrenzbare Bereich (Bereich 2) ergab sich ebenfalls eindeutig aus der Darstellung.

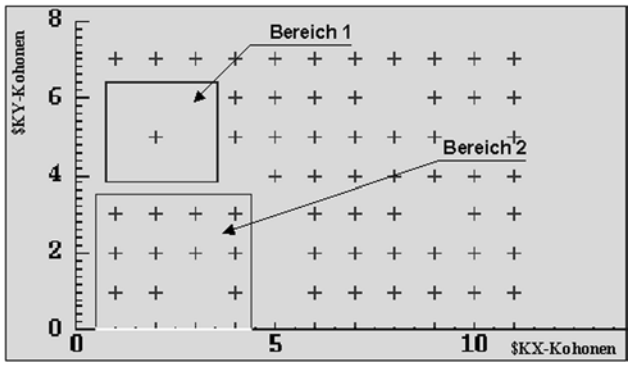

Bild 2: Ergebnis der Clusterung mit dem Kohonen Netz

Auf die Auswahl weiterer Cluster wurde verzichtet, da in der Ergebnisdarstellung der verwendeten Netzvariante keine weiteren deutlich abgrenzbaren Bereiche ersichtlich waren. Bei der Verwendung größerer Netze wäre eine detailliertere Auswahl zwar realisierbar, jedoch wurde diese Möglichkeit im Hinblick auf eine angemessene Größe der Cluster nicht aufgegriffen.

Für die weitere Analyse der ausgewählten Bereiche erwies sich ein C5.0-Modell mit einer Pruning-Rate von 85 % und einer minimalen Datensatzanzahl pro Verzweigung von 20 als am besten geeignet. Die Pruning-Rate bezeichnet das Maß für die Beschneidung eines Klassifikationsbaumes. Entfernt werden dabei solche Äste, deren Erklärungsanteil ein bestimmtes Maß unterschreitet. Die in den entsprechenden Blättern enthaltenen Datensätze werden dem nächsthöheren Blatt zugewiesen. Mit diesem Modell konnte für Bereich 1 eine allgemeine Zuordnungsgenauigkeit von 99,69 % und für Bereich 2 von 99,42 % erreicht werden.

Durch das erstellte Kohonen-Netz in Kombination mit dem C5.0-Klassifikationsverfahren konnten aus der Gesamtmenge der Sessions zwei deutlich abgrenzbare Bereiche von Nutzungsmustern inclusive ihrer wichtigsten Merkmale identifiziert werden.

Zur Klassifikation von SSL-Shop-Besuchern wurde ein Backpropagation-Netz mit zwei verdeckten Schichten ausgewählt, auf die 20 bzw. 5 Netzneuronen entfielen. Die allgemeine Zuordnungsgenauigkeit dieses Modells von 81,92 % wurde zwar von anderen Varianten übertroffen, jedoch konnte damit das Profil von positiven Fällen für einen Shop-Besuch mit 60,44 % durchschnittlich um 10 % besser klassifiziert werden als mit allen anderen Modellen. Auch die Klassifikation negativer Fälle wurde mit 83,12 % sehr gut realisiert. Diese Modellvariante erschien dadurch am geeignetsten, da sie die Struktur der Ausgangsdaten am ausgewogensten abbildete und über eine ausreichende Genauigkeit verfügte. Zudem war auch die Generalisierungsfähigkeit des Modells hinreichend gegeben, da die Klassifikationsergebnisse, bezogen auf die Evaluierungsmenge, gleich bleibend gut waren.

4.3 Ergebnisinterpretation und -umsetzung

Die Untersuchungen zeigten, dass die Besuche auf der PC-Ware-Website unterschiedliche Muster aufweisen. Es wurden die Faktoren aufgedeckt, die wesentlichen Einfluss auf die Art und Weise der Seitennutzung besitzen. Insbesondere die jeweiligen Ziele eines Seitenbesuches haben großen Einfluss auf die situationsbezogenen Informationsbedürfnisse eines Nutzers. Die durchgeführte Untersuchung konnte zudem den Beweis liefern, dass die Ziele und Informationsbedürfnisse der Nutzer sich im Nutzungsverhalten auf den Webseiten von PC-Ware widerspiegeln.

Es konnten zwei Typen von Sessions identifiziert werden, die ganz spezielle Informationen auf der Website suchten bzw. ganz spezielle Servicefunktionen, wie die Suchfunktion für Softwarehersteller, nutzten. Es wurde dabei deutlich, dass sich diese Gruppen in ihrem Verhalten deutlich von anderen Nutzungsmustern abhoben. Diese Kenntnisse um die speziellen Informationsbedürfnisse der Nutzer bilden eine wertvolle Basis für die Ausgestaltung individueller Serviceangebote und Marketingbotschaften für diese Nutzer bzw. für solche Zugriffe.

Im Gegensatz zu anderen Marketingansätzen ermöglicht diese Vorgehensweise eine differenzierte Marketingbearbeitung, die den situations- und besuchszielbedingten Unterschieden im Informationsverhalten und den damit verbundenen Bedürfnissen von Nutzern Rechnung trägt. Die Marketingmaßnahmen können an die jeweiligen Ziele und die Informationsbedürfnisse der Nutzer angepasst werden, wodurch sich die Effizienz dieser Maßnahmen deutlich erhöhen lässt. Zudem konnten die wichtigsten Seitenabrufe identifiziert werden, in denen sich das Verhalten in diesen Sessions deutlich vom durchschnittlichen Verhalten unterscheidet. Gerade diese Seiten bieten ein hohes Potenzial für Modifikationen am Webauftritt und damit zur Optimierung der Online-Services.

Eine sehr gute Möglichkeit, die stark auf die Angebote des Softwarebereiches fokussierten Kunden auch über andere Leistungen wie Consulting oder Schulungsangebote zu informieren, würde sich aus einer Modifikation der Seiten zur SSL-Herstellersuche ergeben. Neben den eigentlichen Suchergebnissen könnten zusätzlich passende Angebote aus anderen Geschäftsbereichen von PC-Ware dargestellt werden. Solche si-

tuativen Offerten sind erheblich chancenreicher als Angebotshinweise während einer möglicherweise später erfolgenden Bestellung per Telefon.

Der zweite identifizierte Bereich war deutlich größer als der erste, jedoch unterschieden sich die enthaltenen Sessions etwas mehr voneinander. Es konnten aber auch für diesen Bereich gemeinsame inhaltliche Merkmale identifiziert werden, die eine einheitliche Marketingbearbeitung ermöglichen. Die enthaltenen Sessions wiesen eine starke Fokussierung auf den SSL-Bereich (Software Sales & Licensing) und den SSS-Bereich (Software Support & System Management) der Site auf. Die betreffenden Seitenabrufe waren mehr als doppelt so häufig zu verzeichnen wie im Durchschnitt der gesamten Ausgangsdaten. Besonders interessant war die überdurchschnittlich häufige SSL-Shop-Frequentierung in diesem Bereich. Andere Seitenbereiche, insbesondere der ISS-Bereich, wurden dagegen unterdurchschnittlich besucht.

Dieses einheitliche Informationsverhalten bietet ein großes Potenzial für individualisierte Marketingaktivitäten. Die betreffenden Nutzer stellen für PC-Ware eine wertvolle Zielgruppe dar. Die überdurchschnittliche Anzahl an Besuchen der Seitenkategorien PCW_Suche und Unternehmen_PC_Ware_Group in den betreffenden Sessions kann zu möglichen Seitenmodifikationen genutzt werden. Zudem ist die Platzierung von gezielten Marketingbotschaften auf diesen Seiten denkbar. Offen bleibt jedoch die generelle Frage, ob diese Botschaften auf die bekannten Interessenschwerpunkte oder auf bislang weniger genutzte Angebote ausgerichtet werden sollten. Durch die gegebenen Ausgangsdaten war es jedoch nicht möglich, Nutzer über mehrere Sessions hinweg identifizieren zu können. Dies wäre beispielsweise durch die Verwendung von Registrierungsmechanismen auf der Website oder persistente Cookies möglich gewesen.

Aus den Untersuchungen mit dem verwendeten Backpropagation-Modell wurde ersichtlich, dass sich Unterschiede im generellen Verhalten von Nutzersitzungen in Abhängigkeit von ihrer Kaufaffinität identifizieren lassen. Diese Erkenntnisse wären durchaus in Modellen zur Berechnung eines Online-Kundenwertes verwendbar. Zudem ist es mit dem generierten Modell bereits möglich, potenzielle Kunden zu identifizieren, die zwar bislang noch nichts gekauft haben und damit dem Unternehmen auch noch nicht bekannt sind, bei denen aber aufgrund ihres Verhaltens eine hohe Kaufaffinität vermutet werden kann.

5 Resümee

Mit dem durchgeführten Projekt konnte ein Ausschnitt aus den vielfältigen Nutzenpotenzialen dargestellt werden, die sich durch die Anwendung von Data-Mining-Methoden auf das Internet erschließen lassen. Sowohl zur Steigerung des allgemeinen Web-Nutzungsverständnisses als auch für die Implementierung spezieller Servicefunktionalitäten, wie etwa zur Personalisierung des Webangebotes, können dem Site-Betreiber mit den verschiedenen Methoden des Data-Mining adäquate Problemlösungen zur Seite gestellt werden. Es wurde während der Untersuchung jedoch auch deutlich, dass dieses spezielle Anwendungsgebiet zugleich sehr hohe Ansprüche an die Auswahl und die Vorbereitung der Ausgangsdaten stellt.

Literatur:

[Björn, 1997] Axel Björn. Künstliche neuronale Netze in Management-Informationssystemen; Grundlagen und Einsatzmöglichkeiten. Wiesbaden: Gabler 1998.
Zugl.: Diss., Hamburg, Univ., 1997.

[Kohavi, 2001] Ronny Kohavi. Mining E-Commerce Data. In: Advances in knowledge discovery and data mining: 5th Pacific Asia Conference; proceedings/PAKDD 2001, Hong Kong, China, April 16--18, 2001. David Cheung (ed.). Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Mailand, Paris, Singapur, Tokyo: Springer 2001.

[Kohonen, 2001] Teuvo Kohonen. Self Organizing Maps. 3. Aufl., Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Mailand, Paris, Singapur, Tokyo: Springer 2001.

[Lämmel und Cleve, 2001] U. Lämmel, J. Cleve. Künstliche Intelligenz. Leipzig: Fachbuchverlag 2001.

[Moe und Fader, 2001] Wendy W. Moe, P. S. Fader. Uncovering Patterns in Cybershopping. California Management Review, 43(4), 2001, S. 106--117. *[www.wendy.moe.com]*

[Pyle, 1999] Dorian, Pyle. Data Preparation for Data Mining. Morgan Kaufmann Publishers, San Francisco, CA 1999.

[Rahm und Stöhr, 2003] Erhard Rahm, Thomas Stöhr. Data-Warehouse-Einsatz zur Web-Zugriffsanalyse. In: Rahm, Erhard; Vossen, Gottfried (Hrsg.): Web & Datenbanken - Konzepte, Architekturen, Anwendungen. Dpunkt Verlag, Heidelberg 2003.

[Reichmann, 2003] Sebastian Reichmann. Einsatz ausgewählter Data Mining Methoden zur Web-Nutzungsanalyse für die PC-Ware Technologies AG Leipzig. Diplomarbeit an der Hochschule für Technik Wirtschaft und Kultur (HTWK), Leipzig 2003.

[Spiliopoulou, 2001] Myra Spiliopoulou. Web Usage Mining - Data Mining über die Nutzung des Web. In: Hippner H. J. et. al.: Handbuch Data Mining im Marketing; Knowledge Discovery in Marketing Databases. 1. Aufl.; Vieweg, Braunschweig 2001.

[Witten und Frank] Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann Publishers, San Francisco, CA 2000.