

Web Mining für die Personalisierung von e-Portalen

Dr. Peter Gentsch

Dr. Peter Gentsch.: Web Mining für die Personalisierung von e-Portalen

1	Einleitung	3
2	Das betrachtete e-Portal	3
3	Preprocessing	5
3.1	Datenquellen	5
3.2	Bereinigung der Registrierungsdaten	5
3.3	Transformation der Registrierungsdaten	6
3.4	Aufbereitung der Logfiles	6
3.5	Abgeleiteter Handlungsbedarf aus dem Preprocessing	7
4	Business Cases	7
4.1	Segmentierung von Besuchertypologien	7
4.2	Prognose des Geschlechts eines Users	10
4.3	Personalisierte Navigation	11
4.3.1	Identifikation zusammengehöriger Interessensgebiete	11
4.3.2	Navigationsverhalten	14
5	Abgeleitete Massnahmen	15
6	Fazit und Ausblick	17

1 Einleitung

Die systematische Sammlung, Aufbereitung, Analyse und Interpretation von Daten über Märkte und deren Beeinflussungsmöglichkeiten ist angesichts der nachhaltigen Markt-Veränderung ein wichtiges Thema im Marketing. Jeden einzelnen Kunden mit seinem individuellen Kundenwert an das Unternehmen zu binden und persönlich individuell zu betreuen, erscheint dabei sowohl in der traditionellen wie auch in der „eWelt“ der Königsweg zu sein, um Wettbewerbsvorteile zu erreichen. Vor dem Hintergrund der Informationsflut in internet-basierten Commerce- und Business-Anwendungen ist es offensichtlich, dass das Thema Kundenbindung im Internet eng mit der Fähigkeit grosse Datenmengen zu managen und zu analysieren und damit mit dem Thema Data Mining respektive Web Mining verbunden ist.¹

Für ein Internetunternehmen ist es essentiell, einen Kunden längere Zeit an sich zu binden, da die Gewinnung eines Kunden gerade im Internet mit enormen Kosten verbunden ist, ein Onlinekunde aber auch einen vergleichsweise viel höheren Customer Lifetime Value hat. Dies liegt daran, dass sich höhere Cross- und Up- Selling Potentiale ergeben, da sich das Sortiment eines Online-Stores schnell und leicht erweitern lässt. Web Kunden neigen auch dazu, ihre Einkäufe auf einen Haupt-Anbieter zu konsolidieren, dessen virtueller Besuch Teil der täglichen Routine wird. Auch die Wirkung von Weiterempfehlungen ist im Internet viel größer. Eine Empfehlung, die z.B. in ein Diskussionsforum oder eine Newsgroup eingestellt wird, wird von Tausenden potentieller Kunden gelesen; dadurch könnte ihre Kaufentscheidung beeinflusst werden.²

Das elektronische Zeitalter hat zu einem Überfluss an Informationen und Reizen geführt, die nur noch zu einem Bruchteil wahrgenommen werden können. Laut der Unternehmensberatung McKinsey sind 98 Prozent der Massenmarketing-Aktionen für den einzelnen Nutzer uninteressant. Genau hier können Personalisierungsansätze Verbesserungen in der Angebotswahrnehmung erreichen. Durch die Personalisierung von Informationen kann die „Awareness“ für diese erhöht und damit auch das Bedürfnis nach Individualität und nutzergerechter Information befriedigt werden.³

Die vorliegende Ausführungen zeigen am Beispiel eines grossen Internet-Portals die Möglichkeiten von Web Mining auf, systematisch Kundenwissen zur Personalisierung von Angeboten zu entwickeln.

2 Das betrachtete e-Portal

Das im folgende beschriebene Internetportal hat sich auf die Bereitstellung von Informationen im weitesten Sinn spezialisiert. Im Sinne eines Informationsportals bietet es, vergleichbar mit einer Online-Enzyklopädie, zahlreiche, kostenlose Informations- und Wissensbestände im Bereich Allgemein- und Spezialwissen an.

Neben digitalen Lexika, Wörterbüchern und Chroniken besteht eine Online Lern-Funktion und ein umfassender Online-Ratgeber, der praktische Tips zu Alltagsfragen, von Geldanlagen

¹ Vgl. Grothe/ Gentsch, (2000), S. 177 ff.

² Vgl. Gentsch, (2001).

³ Vgl. Gentsch et al. (2001), S. 3.

und Bewerbungen bis zu Trendsportarten, Gesundheitstips und Kochrezepten bereithält. Im Community-Bereich können sich die User mit Experten über verschiedene, themenbezogene Chats und Diskussionsforen austauschen.

Die Plattform finanziert sich zur Zeit primär aus der Vermarktung von Werbebannern. Langfristig besteht allerdings das Ziel, neben dem bereits bestehenden Online-Store, vor allem über die Vermarktung der Informationsinhalte über Contentsyndication, zusätzliche Einnahmequellen zu erschließen.

Mit Hilfe von Web Mining sollen die versteckten Potentiale des e-Portals erkennbar gemacht werden, indem verborgene Zusammenhänge in den automatisch anfallenden Daten transparent gemacht und die relevanten Informationen über die Nutzung der Inhalte, aber auch über die Interessen und Bedürfnisse der Kunden extrahiert werden, um diese dann gezielt zur Personalisierung des e-Portals einsetzen zu können.

Das folgende Praxisbeispiel gibt darüber Auskunft, welche Informationen Web Mining generieren kann, welche Vorbereitungen für eine erfolgreiche Analyse zu treffen sind und welche Vorgehensweise letztendlich zu verfolgen ist, um die gewünschten Informationen aus denen Datenbergen zu erschließen.

Zu diesem Zweck werden mehrere Business Cases dargestellt, die so auch in der Praxis umgesetzt wurden.

Ein vorrangiges Ziel des e-Portals ist es, die User langfristig an die Seite zu binden und den Lifetime Value seiner Kunden zu maximieren. Die Seite besticht in erster Linie durch die Qualität der Inhalte und Informationen. Um die gesteckten Ziele zu verwirklichen, ist es allerdings unumgänglich, seinen Kunden echte Mehrwerte gegenüber anderen Anbietern zu bieten. Dies ist zur Zeit noch nicht der Fall. Wichtige Fragestellungen sind in diesem Zusammenhang noch ungeklärt, deren Beantwortung jedoch essentiell wäre, um den Kunden Mehrwerte liefern zu können:

- Wer sind die User und woher kommen sie?
- Wie verwenden die User das e-Portal?
- Woran sind die User interessiert und welche Bedürfnisse haben sie?
- Wie verhalten sich die User auf der Seite?
- Welche Bereiche werden besucht und wie verteilen sich die Besucherströme?
- Welchen Wert haben die User für das e-Portal (Customer Lifetime Value)?
- Wie hält man die User und macht sie zu treuen Kunden?
- Wie lassen sich Werbemaßnahmen effizient gestalten?

Grundsätzlich lässt sich nicht immer trennscharf festlegen, welche Fragestellung und da Web Mining Analyse welchem der folgenden, Business Cases zuzuordnen ist. Der Grund dafür liegt in der multiplen Verwendbarkeit der Ergebnisse. So werden z.B. die Ergebnisse der Analyse des Surf-Verhaltens der User dazu verwendet dem Kunden individuelle Inhalte zu liefern, aber auch, um ihm die Navigation zu erleichtern.

Zuerst ist es wichtig, verschiedene Besucherklassen zu identifizieren und herauszufinden, welche unterschiedlichen Bedürfnisse und Interessen sie haben. Mit diesen Informationen lassen sich individuelle Angebote erstellen und für den Kunden interessante Inhalte selektieren. Auch Werbemaßnahmen lassen sich dann gezielt an den einzelnen User anpassen. Auch die Performanz einer Seite, also die technische und strukturelle Qualität (Usability), ist ein entscheidendes Merkmal einer Web-Site, das einen Mehrwert darstellt. Denn die Internet-Nutzer bevorzugen Seiten, die als angenehm und leicht zu bedienen wahrgenommen werden und gleichzeitig ein hohes Maß an Intuitivität beinhalten. Daher sollte die Navigation durch die Seiten an den individuellen User angepasst werden.

3 Preprocessing

3.1 Datenquellen

Für die Web Mining Analysen standen zwei verschiedenen Datenquellen zur Verfügung. Zum einen wurde auf die Benutzerdatenbank zurückgegriffen, die sich im Wesentlichen aus den Stammdatendaten der User, die im Registrierungsprozess erfasst worden sind, zusammensetzt. Zusätzlich wurden noch zwei weitere Merkmale (LOGINCONTER und LAST_LOGIN) aufgenommen, die das Anmelden der Users auf der Seite beschreiben. Insgesamt besteht die Benutzerdatenbank aus 400964 verschiedenen Datensätzen, deren einzelne Felder in Tabelle 1 aufgeführt sind.

USER_ID	Ganze Zahl zur eindeutigen Identifikation eines einzelnen Users. Wird in der User-Datenbank als Primärschlüssel verwendet.
AGE_ID	Alter der User, aufgeschlüsselt in 9 Altersklassen (unter 16,16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, über 50 Jahre).
SEX	Geschlecht des Users (männlich/weiblich)
POST_CODE	Postleitzahl
CITY	Wohnort
TITLE	akademischer Titel des Users
ANREDE	Herr/Frau
INTEREST_ID	11 Interessengebiete (Kultur [1], Geschichte [2], Gesellschaft [3], Geographie [4], Natur [5], Technik [6], Wirtschaft [7], Politik [8], Sport [9], Musik [10], Reisen [11])
NEWS	Bestellung des Newsletters (ja/nein)
CREATED	Zeitpunkt, zu dem der Registrierungsdatensatz erzeugt wurde.
LAST_LOGIN	Zeitpunkt zu dem sich der User das letzte Mal auf der Seite mit seinem Passwort angemeldet bzw. eingeloggt hat.
LOGINCOUNTER	Anzahl aller Logins eines Users.

Tabelle 1: Bestandteile der Benutzerdatenbank

Die zweite Datenquelle bilden die gesammelten Logfiles (erweitertes CLF) des Internet-Portals, in denen im Rahmen eines Jahres insgesamt 9.454.814 einzelne Sessions identifiziert werden konnten.

3.2 Bereinigung der Registrierungsdaten

Um die Registrierungsdaten zum Web Mining verwenden zu können, müssen diese zunächst auf ihre Qualität hin überprüft und gegebenenfalls bereinigt werden.

Zunächst wurde entschieden, welche Variablen für eine Analyse nicht sinnvoll verwendet werden können. Solche Variablen wurden dann nicht weiter betrachtet. Das Merkmal „ANREDE“ fällt dabei als erstes auf. Da nur zwischen „Herr“ und „Frau“ gewählt werden kann, gibt es letztendlich nur das Geschlecht eines Users wieder. Dieses wird jedoch schon in der Variablen „SEX“ abgefragt. Das Merkmal „ANREDE“ enthält daher keinen zusätzlichen Informationsgehalt und wird ausgeschlossen.

Die Variable „LOGINCOUNTER“ gibt an, wie oft sich der User auf der Seite eingeloggt hat. Da jedoch, wie bereits oben angesprochen, eine Anmeldung auf der Seite überflüssig ist, enthält die Variable „LOGINCOUNTER“ keinen Hinweis auf die tatsächliche Nutzungshäufigkeit. Sie beinhaltet daher keine verwertbare Information und wird nicht weiter verwendet. Gleiches gilt auch für die Variable „LAST-LOGIN“, die den Zeitpunkt festhält, zu dem sich der jeweilige User das letzte Mal auf der Seite eingeloggt hat. Auch sie wurde von der weiteren Analyse ausgeschlossen.

Bei der Betrachtung des Anteils der missing values der einzelnen Variablen, fiel der hohe Anteil an fehlenden Merkmalswerten bei die Variable „TITLE“ auf. Dies ist zu gravierend, als dass diese Variable hätte weiter verwendet werden kann. Da der akademische Titel zudem als Freitext eingegeben wurde, existieren sehr viele verschiedene Ausprägungen dieser Variable, die z.T. nur auf unterschiedliche Schreibweisen zurückzuführen sind. Die Variable „TITLE“ wurde daher nicht weiter berücksichtigt.

Anschliessend wurde mit einem entsprechenden Programm eine Plausibilitätsprüfung durchgeführt. So müssen z.B. die Eintragungen im Feld „POST_CODE“ dem Format der Postleitzahl des jeweiligen Landes entsprechen. Für Deutschland ist dies eine fünfstellige, ganze Zahl. Ebenso lassen sich durch die Plausibilitätsprüfung auch Datensätze eliminieren, die sinnlose Zeichenketten aufweisen, oder die Angabe von „Fantasienamen“ wie z.B. User mit dem Namen „Donald Duck“. Datensätze, die wenig plausible Werte enthalten haben, wurden aus der Datenmenge entfernt.

3.3 Transformation der Registrierungsdaten

Einige Variablen waren auch nach der Bereinigung nicht besonders zum Aufspüren von Mustern und Zusammenhänge in den Daten geeignet und mussten vor der Analyse transformiert werden.

Aus der Variablen „POST_CODE“ kann man die Information gewinnen, aus welcher Gegend ein User stammt. Die regionale Unterteilung eines Landes auf Basis der Postleitzahlen ist jedoch meist zu detailliert, um die geographischen Gemeinsamkeiten von Benutzern zu beschreiben. Es ist ein höherer Aggregationsgrad erforderlich. Daher wurde aus der Variablen „POST_CODE“ die neu Variable „Region“ gebildet. Sie enthält jeweils nur die erste Ziffer der Postleitzahl, so dass die Herkunft der User in 10 große Gebiete eingeteilt wird.

Auch die ursprüngliche Aufteilung der Altersklassen wurde verändert, um größere Klassen zu erhalten, die eine höhere Aussagekraft besitzen. Es existieren nun 5 verschiedene Klassen in der Abstufung: „unter 21 Jahre“, „21-30“, „31-40“, „41-50“ und „über 50 Jahre“.

Auch die Variable „CREATED“ kann in ihrer ursprünglichen Form nicht sinnvoll zum Web Mining eingesetzt werden. Sie gibt den Zeitpunkt der Registrierung auf die Sekunde genau an. Gemeinsamkeiten zwischen mehreren Datensätzen bezüglich dieses Merkmals sind somit nicht zu erwarten. Anstatt des genauen Zeitpunkts wurde daher der Monat verwendet, in dem die Registrierung erfolgt ist.

Weitere Korrelationen bestanden zwischen den verschiedenen Interessensgebieten wie z.B. zwischen Politik und Wirtschaft, aber auch zwischen einzelnen Interessengebieten und dem Geschlecht der User, wie z.B. zwischen Technik und dem Geschlecht „männlich“. Diese Korrelationen sind aber gewollt und geben bereits erste Hinweise auf wichtige Zusammenhänge und Muster in den Daten.

3.4 Aufbereitung der Logfiles

Bei der Aufbereitung des Logfiles wird die Protokolldatei zunächst um alle Einträge bereinigt, die nicht für die weitere Untersuchung relevant sind. Enthält eine HTML-Seite

eingebettete Elemente, wie Grafiken, so wird bei einem Aufruf des Dokuments sowohl für die HTML-Datei selbst, als auch für jede darin enthaltene Grafik, ein eigener Eintrag im Logfile erzeugt. Einträge, die sich auf eingebettete Elemente beziehen, gilt es herauszufiltern. Weitere irrelevante Daten sind Einträge über fehlerhafte Anforderungen. Sie sind lediglich für administrative Belange von Bedeutung. Zusätzlich werden Protokolleinträge, die Zugriffe von Suchmaschinen oder administrativen Zugriffen dokumentieren, identifiziert und eliminiert. Für die angestrebte Web Mining Analyse sind nicht alle Informationen, die im Logfile erfasst werden von Interesse. Wichtig sind der Zeitstempel, die URL der aufgerufenen Webseite und die Session-ID, mit deren Hilfe einzelne Logfiles zu einer Session zusammengefügt werden können.

Eine weitere Codierung der Session-ID ist hier nicht nötig, da die Sequenzanalyse, die hier zum Einsatz kommt, die Session-Daten direkt verarbeiten kann. Würden man die Logfile-Daten für eine Clusteranalyse verwenden wollen, müssten die Daten strukturiert in Form einer Datenmatrix vorliegen.

3.5 Abgeleiteter Handlungsbedarf aus dem Preprocessing

Bereits aus dem Preprocessing konnte eine Anzahl von Handlungsempfehlungen gewonnen werden, die nicht nur den Aufwand in der Preprocessing-Phase deutlich reduzieren, sondern insbesondere auch die späteren Analysepotentiale signifikant erhöhen können.

Besonders wichtig erscheint es zunächst, die Voraussetzungen dafür zu schaffen, dass die Registrierungsdaten mit den Stammdaten verbunden werden können, so dass ein umfassenderes Bild der User erzeugen werden kann. Dies könnte z.B. mit Hilfe von Cookies oder einer Ausweitung des Login-Bereichs geschehen.

Aber auch die Registrierung bedarf einiger Veränderungen. Es sollten schon bei der Abfrage der einzelnen Merkmale strengere Plausibilitätskontrollen implementieren werden, um die Anzahl der fehlerhaften Datensätze von vornherein zu minimieren. Dazu gehören auch konsistente Methoden zur Erfassung der Interessensgebiete (z.B. durch Rotation der Anordnung), sowie die Möglichkeit, auch keines der Interessensgebiete anzugeben, um für den User nicht zutreffende Angaben von missing values unterscheiden zu können. Denkbar wäre hierfür auch ein Freitext Eingabefeld zur Angabe von weiteren, aus Usersicht nicht zuzuordnenden Interessen.

Für die Abfrage des Merkmals „TITEL“ sollte eine Auswahlliste angeboten werden, um die Anzahl der fehlenden Datensätze zu verringern und die vielen unterschiedlichen Schreibweisen, die bei diesem Merkmal auftreten zu vermeiden.

Letztendlich sollte auch die Möglichkeit evaluiert werden, die vorhandenen Daten durch externe Daten anzureichern.

4 Business Cases

4.1 Segmentierung von Besuchertypologien

Das e-Portal hat es unter anderem durch den massiven Einsatz von Werbung geschafft, viele Besucher auf die Seite aufmerksam zu machen und einen hohen Traffic zu erzeugen. Eine genauere Kenntnis der Besucher und ihrer Interessen ist jedoch nicht vorhanden. Die Besucher bleiben daher weitestgehend anonym.

Aufgrund der hohen Userzahlen und des breiten Angebotes wird jedoch unterstellt, dass innerhalb Gesamtheit der User des Informationsportals unterschiedliche Zielgruppen existieren.

Um seinen Besuchern Mehrwerte liefern zu können, ist es jedoch wichtig, zu wissen, welche unterschiedlichen, homogenen Gruppen innerhalb der heterogenen Userschaft existieren.

Über eine Segmentierung auf Basis der Registrierungsdaten wird versucht, die verschiedenen Besuchergruppen, die innerhalb der Gesamtheit der Besucher bestehen, zu identifizieren.

Auf diese Weise erhält man tiefreichende Kenntnisse über die Struktur der Nutzer und deren unterschiedliche Eigenschaften und Präferenzen. Diese Kenntnis wurde dazu genutzt, sowohl den Aufbau als auch den Inhalt der Seiten auf die individuellen Bedürfnisse und Interessen eines jeden Besuchers abzustimmen.

Werden bei der Segmentierung User-Gruppen entdeckt, die man nicht auf den eigenen Seiten vermutet hätte, können diese entweder gezielt bearbeitet oder aber bewusst ausgegrenzt werden, um das Marketingbudget effizient aufzuteilen.

Durch die Identifikation unterschiedlicher Besuchergruppen, wird es aber auch möglich, gezielt verkaufsfördernde Maßnahmen durchzuführen, indem für bestimmte Kampagnen die Zielgruppenprofile ausgewählt werden, die für die anstehende Kampagne besonders empfänglich sind. Diese Zielgruppen können dann individuell angesprochen werden. Auf diese Weise wurden sowohl die direct-mail Aktionen als auch die Bannerauslieferungen optimiert.

Auch im Hinblick auf die geplante kommerzielle Contentsyndication ist es wichtig, die Kernzielgruppe des jeweiligen Contentpartners innerhalb der Usern identifizieren zu können und festzustellen, welche Contentseiten diese Gruppen vorzugsweise verwendet haben.

Zur Clusterbildung wurde ein neuronales Kohonen Netz eingesetzt, das aufgrund der meist höheren Präzision einem hierarchischer Verfahren vorgezogen wurde.

Nach mehreren Durchläufen mit unterschiedlichen Clusterzahlen wurde aus Plausibilitätsgründen die Lösung mit vier Clustern als Endergebnis gewählt. Diese Lösung hebt die in den Daten enthaltenen Strukturen am deutlichsten hervor und ermöglicht eine aussagekräftige Interpretation. Auffällig war dabei, dass die Interessensgebiete die Hauptkristallisationspunkte boten; also von besonderem Gewicht bei der Bildung der Usergruppen waren, während, abgesehen vom Geschlecht, die demographischen Merkmale wie Alter und Wohnort nur eine untergeordnete Rolle bei der Einteilung der Cluster spielten.

Die User des, mit 28,7% der Grundgesamtheit, größten Cluster haben annähernd alle Interessensgebiete angegeben. Dies ist jedoch nicht als besonders weites Interesse an allen Wissensgebieten zu werten, sondern deutet eher auf ein stark indifferentes Interesse hin. Die User dieses Clusters haben bei ihrer Registrierung alle Interessensgebiete angegeben, um „nichts zu verpassen“. Ein wirklich zielgerichtetes Interesse an bestimmten Themengebieten besteht allerdings nicht.

Im zweiten Cluster (27,7% der Grundgesamtheit) findet sich das entgegengesetzte Bild. Hier wurden so gut wie keine Interessen angegeben. Das stärkste Interessen besteht hier noch in den Bereichen Technik, Musik und Natur. Aber selbst in diesen Gebieten bleibt das Interesse zum Teil stark hinter dem Durchschnitt zurück. Die User dieser Gruppe haben zwar durch ihre Registration ein grundsätzliches Interesse an der Seite zum Ausdruck gebracht, dass sich jedoch nicht in der Angabe von Interessengebieten niederschlägt. In diesem Cluster befinden sich auffällig viele User unter 20 Jahren und relativ wenige, die Älter als 40 Jahre sind.

Im dritten Cluster sind dagegen eindeutige Präferenzen auszumachen (siehe Abbildung 1). Dieses Cluster stellt 22,7% der Grundgesamtheit. Besonderes Interesse besteht für Wirtschaft und Politik. Aber auch an Technik und Geschichte ist ein überdurchschnittliches Interesse zu verzeichnen. Die User dieses Clusters sind dementsprechend fast ausschließlich Männer.

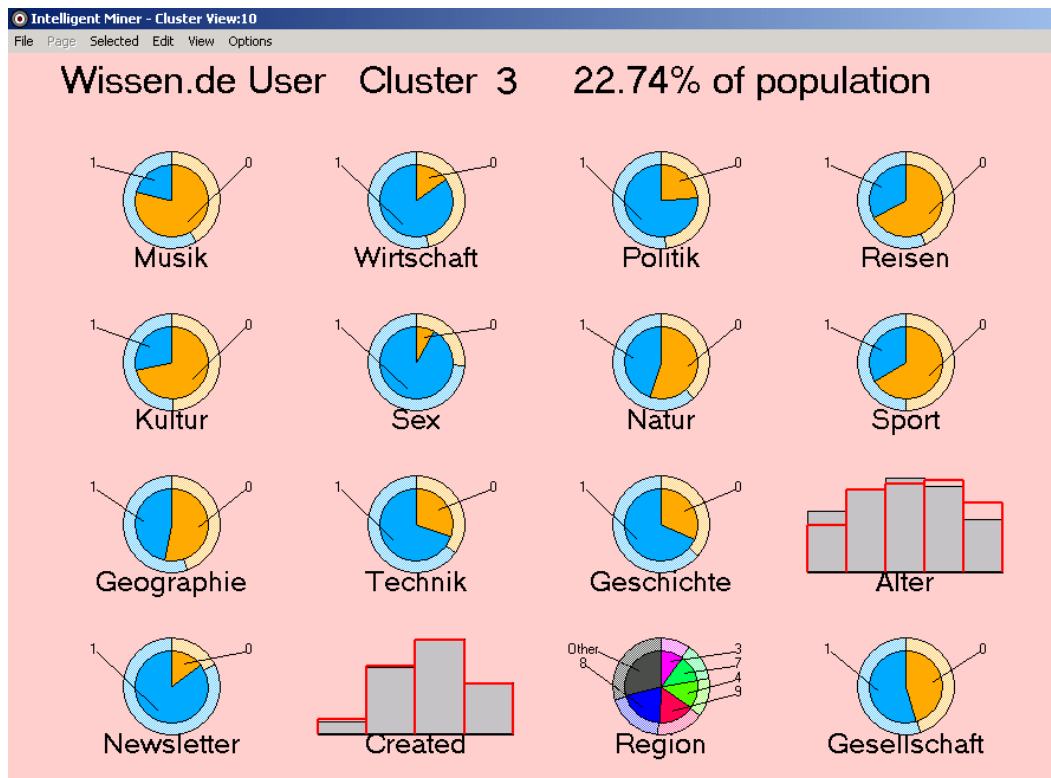


Abbildung 1: Cluster 3 – „Politisch-wirtschaftlich interessierte Männer“⁴

Im letzten Cluster (22,00% der Grundgesamtheit) sind die Interessen für Kultur, Reisen, Musik, aber auch für Gesellschaft, Geschichte und Natur besonders ausgeprägt. In diesem Cluster befinden sich erwartungsgemäß ausgesprochen viele Frauen.

Cluster 1 und 2 umfassen zusammen 56,5% der registrierten User. Über die Hälfte aller registrierten User hat demnach kein tiefergehendes bzw. zielgerichtetes Interesse an der Seite. Nur der kleinere Teil der User besitzt Interessenschwerpunkte, die bei den Männern eher im Bereich Wirtschaft, Politik und Technik liegen und für die Frauen in Kultur, Reisen und Musik zu finden sind.

Es wäre nun wünschenswert, feststellen zu können, welcher Kundentyp sich vorzugsweise in welchen Bereichen der Seite aufhält. Besonders bei der geplanten Content-Syndication wäre es für den Verkauf von Inhalten förderlich, dem jeweiligen Kunden Informationen über die Besucher liefern zu können, die den betreffenden Inhalt verwenden. Aufgrund der fehlenden Verbindungsmöglichkeiten zwischen Registrierungsdaten und Logfiles ist dies allerdings nicht möglich.

⁴ Die Reihenfolge der Merkmale innerhalb des Cluster ergibt sich aus der Bedeutung der einzelnen Merkmale für die Aufteilung der Cluster, beginnend mit dem Merkmal, das für die Bildung des Clusters am wichtigsten ist.
Bei den Interessegebieten bedeutet eine „1“, dass das Interessensgebiet bei der Registrierung ausgewählt wurde, eine „0“ bedeutet hingegen, dass den User das Gebiet nicht interessiert. Bei dem Merkmal „SEX“, steht die „1“ für die Merkmalsausprägung männlich und die „0“ für die Merkmalsausprägung weiblich.
Bei den Kreisdiagramme gibt der innerer Kreis die Verteilung eines Merkmals innerhalb des jeweiligen Clusters wieder. Zum Vergleich stellt der äußerer Ring die Verteilung des Merkmals in der Grundgesamtheit dar. Bei den Säulendiagramme nimmt die rote Säule Bezug auf das jeweilige Cluster und die graue Säule auf die Grundgesamtheit.

4.2 Prognose des Geschlechts eines Users

Die Verweildauer eines Users auf den Seiten des e-Portals und die Besuchshäufigkeit wären interessante Prognoseobjekte für unser Informationsportal. Ohne eine Identifikation einzelner Kunden können solche Modelle aber nicht erstellt werden. Deshalb wurde, vorrangig, um die Potentiale eines Web Mining Prognosemodells aufzuzeigen, ein Modell zur Vorhersage des Geschlechts eines Users aufgebaut.

Der Aufbau des Modells erfolgte über einem CART-Algorithmus, der den Gini-Index verwendet.

- weiblich
- männlich

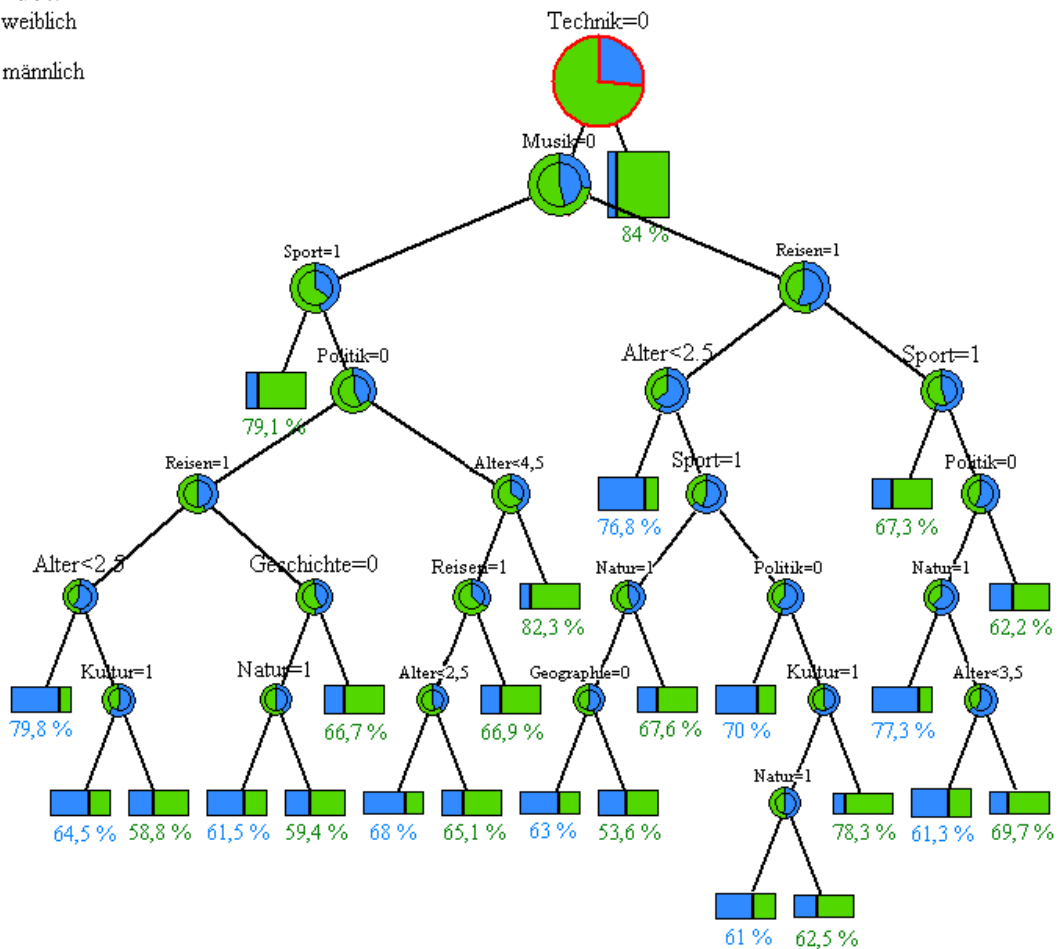


Abbildung 2: Entscheidungsbaum zur Prognose des Geschlechts

Das resultierende Modell ist in Abbildung 2 dargestellt. Es besitzt eine Prognosegenauigkeit von 79,1 %, d.h. bei 79,1% der User wird das Geschlecht richtig vorhergesagt. Dies entspricht einer Steigerung von 5,6% gegenüber einer zufälligen Auswahl: Wenn man das Geschlecht eines unbekanntem User raten müsste, würde man vermuten, dass er männlich sei, da 73,5% aller bekannten User männlich sind. Somit ist es wahrscheinlich, dass auch ein unbekannter User männlich ist. 26,5% des User sind aber weiblich. Die Fehlerrate einer zufälligen Auswahl beträgt also 26,5%. Prognostiziert man das Geschlecht eines Users dagegen mit dem zuvor modellierten Entscheidungsbaum beträgt die Fehlerrate nur 20,9%.

Wie bereits die vorhergehende Segmentierung angedeutet hat, zeigt sich auch beim Entscheidungsbaum, dass ein deutlicher Zusammenhang zwischen einem ausgeprägten technischen Interesse und dem männlichen Geschlecht besteht.

Denn das wichtigste Trennungskriterium des Baumes ist das Interesse an Technik für den Vorhersagewert „männlich“. Weitere wichtige Kriterien für die Aufteilung der Geschlechter

sind Interesse an Musik, Reisen, Sport und Politik. Von den demographischen Merkmalen leistet nur das Alter der User einen Erklärungsbeitrag.

Obwohl dieser Baum zur Vorhersage des Geschlechts eher der Illustration der Potentiale eines Prognosemodells dienen soll, kann ein solcher Baum aber durchaus für eine Online-User-Klassifizierung verwendet werden, sofern auf der Site mit geschlechtsspezifisch zugeschnittenen Inhalten gearbeitet werden soll.

Aber auch um bei zukünftigen Web Mining Analysen die missing values der Variablen „Geschlecht“ fundierter behandeln zu können, ist dieses Modell durchaus hilfreich.

4.3 Personalisierte Navigation

In Folge des ständigen Wachstums an Informationsangeboten im Internet ist es oftmals schwierig, genau die Informationen zu erhalten, die man gerade benötigt. Wird jedoch das Profil eines Users und sein bisheriges Surf-Verhalten bei der Auswahl der Inhalte berücksichtigt, können dem Nutzer aus der Fülle der vorhandenen Informationen gezielt diejenigen präsentiert werden, die seinen Interessen am besten entsprechen.

Auch die Reihenfolge bzw. die Verlinkung der Seiten bis hin zur Zusammenstellung einzelner Menüs kann an den einzelnen User angepaßt werden. Seiten, die häufig zusammen abgerufen werden können enger miteinander verlinkt werden, so dass man von einer Seite direkt auf die nächste gelangen kann ohne Umwege über weitere Seiten nehmen zu müssen. Dadurch wird die Navigation erleichtert und beschleunigt.

Sowohl Assoziations- als auch Sequenzanalysen können dafür eingesetzt werden, die Navigation das Informationsangebot des e-Portals für die Kunden zu optimieren. Bei den technischen Möglichkeiten der situativen, personalisierten Anpassung von Web-Seiten dürfen jedoch auf keinen Fall die Usability-Gesetze verletzt werden. So sollte eine Verwirrung bei der Seitennavigation durch einen blinden „Anpassungsaktionismus“ vermieden werden. Einen echten Mehrwert im Sinne des „Guided Browsing“ erreicht man durch systematische, kontinuierliche Benutzerführung, die den Bedürfnissen der User entspricht. Die adaptive Benutzermodellierung sollte dem Grundsatz der Erwartungskonformität folgen, d.h. personalisierte Inhalte und Verlinkungen sollten in möglichst fest definierten Bereichen der Webseite angezeigt werden.

4.3.1 Identifikation zusammengehöriger Interessensgebiete

Das Ergebnis der Segmentierung der Kundentypologien ließ bereits vermuten, dass bestimmte Interessensgebiete, wie z.B. „Wirtschaft“ und „Politik“ häufiger zusammen auftreten als andere. Mit Hilfe einer Assoziationsanalyse soll nun der Frage genauer nachgegangen werden, ob unter den elf verschiedenen Interessengebieten häufige und gerichtete Verknüpfungen existieren. Auch auffällig selten vorkommende Interessenkombinationen und unerwartete Interessenverknüpfungen sollen identifiziert werden.

Solche Informationen können dann dabei helfen, zusammengehörige Interessengebieten auch bei der Seitengestaltung zu berücksichtigen und es zu vermeiden, den Usern irrelevante Inhalte aber auch Werbungen zu präsentieren.

Nach mehreren Testläufen mit jeweils unterschiedlichen Einstellungen für Support und Confidence, wurde der Support auf 30% und die Confidence auf 55% festgelegt, da diese Einstellungen hier den besten Kompromiss zwischen Anzahl und Stärke der gefundenen Regeln darstellen, so dass eine überschaubare Anzahl aussagekräftiger Regeln gefunden wurde.

Der Support misst dabei die statistische Signifikanz einer gefundenen Regel, also den Anteil der Datensätze, die die betreffenden Interessensgebiete aufweisen. Für die erste Zeile der Tabelle 3 wären das z.B. der Anteil alle Datensätze, die die Interessensgebiete Technik und gleichzeitig Natur enthalten.

Die Confidence gibt dagegen Auskunft über die Stärke einer Regel. Sie sagt aus, bei welchem Anteil der Datensätze, die das Interessensgebiet Technik enthalten auch der Bereich Natur von Interesse ist.

Neben dem Support und der Confidence einer Regel wird zusätzlich deren Lift angegeben. Der Lift einer Regel gibt die Abweichung der tatsächlichen von der statistisch zu erwarteten Häufigkeit einer Regel an, die nach der Verteilung der Interessensgebiete in der Grundgesamtheit für den unabhängigen Fall berechnet wird.

Unter den generierten Regeln wurde eine Gruppe häufig zusammen auftretender Interessensgebieten gefunden, die die Bereiche Natur, Geschichte, Geographie und Technik umfasst (siehe Tabelle 2).

Association Rules - "Untitled"						
File	View	Sort	Filter	Format		
Support	Confidence	Type	Lift		Rule	
49.253	73.6	+	1.2	[Technik]	==>	[Natur]
49.253	78.8	+	1.2	[Natur]	==>	[Technik]
46.320	69.3	+	1.1	[Technik]	==>	[Geschichte]
46.320	71.5	+	1.1	[Geschichte]	==>	[Technik]
46.025	73.7	+	1.1	[Natur]	==>	[Geschichte]
46.025	71.1	+	1.1	[Geschichte]	==>	[Natur]
45.720	70.6	+	1.3	[Geschichte]	==>	[Geographie]
45.720	82.0	+	1.3	[Geographie]	==>	[Geschichte]
44.892	71.8	+	1.3	[Natur]	==>	[Geographie]
44.892	80.5	+	1.3	[Geographie]	==>	[Natur]
43.502	78.0	+	1.2	[Geographie]	==>	[Technik]
43.502	65.0	+	1.2	[Technik]	==>	[Geographie]
43.454	67.1	+	1.2	[Geschichte]	==>	[Gesellschaft]
43.454	77.8	+	1.2	[Gesellschaft]	==>	[Geschichte]
38.341	85.4	+	1.3	[Geographie] AND [Natur]	==>	[Geschichte]
38.341	83.9	+	1.3	[Geschichte] AND [Geographie]	==>	[Natur]
38.341	83.3	+	1.5	[Geschichte] AND [Natur]	==>	[Geographie]
38.199	82.5	+	1.3	[Geschichte] AND [Technik]	==>	[Natur]
38.199	83.0	+	1.2	[Geschichte] AND [Natur]	==>	[Technik]
38.199	77.6	+	1.2	[Natur] AND [Technik]	==>	[Geschichte]
37.989	84.6	+	1.3	[Geographie] AND [Natur]	==>	[Technik]
37.989	77.1	+	1.4	[Natur] AND [Technik]	==>	[Geographie]
37.989	87.3	+	1.4	[Geographie] AND [Technik]	==>	[Natur]
36.942	79.8	+	1.4	[Geschichte] AND [Technik]	==>	[Geographie]
36.942	80.8	+	1.2	[Geschichte] AND [Geographie]	==>	[Technik]
36.942	84.9	+	1.3	[Geographie] AND [Technik]	==>	[Geschichte]
33.314	87.2	+	1.6	[Geschichte] AND [Natur] AND [Technik]	==>	[Geographie]
33.314	90.2	+	1.4	[Geschichte] AND [Geographie] AND [Technik]	==>	[Natur]
33.314	87.7	+	1.4	[Geographie] AND [Natur] AND [Technik]	==>	[Geschichte]
33.314	86.9	+	1.3	[Geschichte] AND [Geographie] AND [Natur]	==>	[Technik]

Tabelle 2: Assoziationen I

Die Zusammensetzung des Clusters 3 der Besuchertypologien weist auf eine Verbindung der Interessen Wirtschaft und Politik hin. Diese Verbindung konnte mit der Assoziationsanalyse bestätigt werden (siehe Tabelle 3). Werden Politik oder auch Wirtschaft angegeben, so interessiert wahrscheinlich auch Geschichte. Ein geschichtliches Interesse lässt dagegen nicht unbedingt auf die Interessensgebiete Wirtschaft und Politik schließen.

Association Rules - "Untitled"						
File	View	Sort	Filter	Format		
Support	Confidence	Type	Lift		Rule	
41.341	78.7	+	1.4	[Politik]	==>	[Wirtschaft]
41.341	75.1	+	1.4	[Wirtschaft]	==>	[Politik]
43.169	82.2	+	1.3	[Politik]	==>	[Geschichte]
43.169	66.7	+	1.3	[Geschichte]	==>	[Politik]
41.712	64.4	+	1.2	[Geschichte]	==>	[Wirtschaft]
41.712	75.8	+	1.2	[Wirtschaft]	==>	[Geschichte]
35.123	85.0	+	1.3	[Wirtschaft] AND [Politik]	==>	[Geschichte]

Tabelle 3: Assoziationen II

Interessiert sich ein User für Wirtschaft, Geographie, Gesellschaft, Musik, Politik, Reisen, oder Kultur, ist meist auch technisches Interesse vorhanden. Andersherum gilt diese Regel allerdings nur eingeschränkt (siehe Tabelle 4).

Association Rules - "Untitled"						
Support	Confidence	Type	Lift	Rule		
43.626	65.2	+	1.2	[Technik]	==>	[Wirtschaft]
43.626	79.3	+	1.2	[Wirtschaft]	==>	[Technik]
43.502	65.0	+	1.2	[Technik]	==>	[Geographie]
43.502	78.0	+	1.2	[Geographie]	==>	[Technik]
40.274	60.2	+	1.1	[Technik]	==>	[Gesellschaft]
40.274	72.1	+	1.1	[Gesellschaft]	==>	[Technik]
39.760	70.6	+	1.1	[Musik]	==>	[Technik]
39.760	59.4	+	1.1	[Technik]	==>	[Musik]
39.160	58.5	+	1.1	[Technik]	==>	[Politik]
39.160	74.6	+	1.1	[Politik]	==>	[Technik]
38.798	70.1	+	1.1	[Reisen]	==>	[Technik]
38.798	58.0	+	1.1	[Technik]	==>	[Reisen]
37.523	71.6	+	1.1	[Kultur]	==>	[Technik]
37.523	56.1	+	1.1	[Technik]	==>	[Kultur]

Tabelle 4: Assoziationen III

Darüber hinaus wurden mehrere Einzelregeln mit sehr hoher Eintrittswahrscheinlichkeit gefunden (Confidence um 90%), die in Tabelle 5 aufgeführt sind.

Association Rules - "Untitled"						
Support	Confidence	Type	Lift	Rule		
34.143	92.7	+	1.4	[Kultur] AND [Geographie]	==>	[Geschichte]
30.334	92.2	+	1.4	[Gesellschaft] AND [Geographie] AND [Natur]	==>	[Geschichte]
33.324	91.5	+	1.4	[Geographie] AND [Politik]	==>	[Geschichte]
31.201	91.4	+	1.5	[Kultur] AND [Geschichte] AND [Geographie]	==>	[Natur]
32.914	91.4	+	1.4	[Kultur] AND [Politik]	==>	[Geschichte]
30.039	91.0	+	1.4	[Kultur] AND [Natur] AND [Technik]	==>	[Geschichte]
30.258	90.7	+	1.7	[Geschichte] AND [Gesellschaft] AND [Wirtschaft]	==>	[Politik]
30.039	90.5	+	1.5	[Kultur] AND [Geschichte] AND [Technik]	==>	[Natur]
33.314	90.2	+	1.4	[Geschichte] AND [Geographie] AND [Technik]	==>	[Natur]
33.152	90.0	+	1.4	[Kultur] AND [Geographie]	==>	[Natur]
30.334	89.8	+	1.4	[Geschichte] AND [Gesellschaft] AND [Geographie]	==>	[Natur]
33.762	89.8	+	1.4	[Gesellschaft] AND [Geographie]	==>	[Geschichte]
30.258	89.3	+	1.4	[Gesellschaft] AND [Wirtschaft] AND [Politik]	==>	[Geschichte]
31.381	89.3	+	1.4	[Kultur] AND [Wirtschaft]	==>	[Geschichte]
30.334	89.2	+	1.6	[Geschichte] AND [Gesellschaft] AND [Natur]	==>	[Geographie]
34.714	89.0	+	1.3	[Natur] AND [Wirtschaft]	==>	[Technik]
31.201	88.5	+	1.6	[Kultur] AND [Geschichte] AND [Natur]	==>	[Geographie]
30.077	88.5	+	1.4	[Politik] AND [Musik]	==>	[Geschichte]
33.114	88.5	+	1.4	[Geographie] AND [Wirtschaft]	==>	[Geschichte]
33.181	88.4	+	1.4	[Kultur] AND [Technik]	==>	[Geschichte]
32.391	88.3	+	1.4	[Natur] AND [Politik]	==>	[Geschichte]

Tabelle 5: Assoziationen IV

Die Suche nach aussagekräftigen und relevanten Regeln kann durch die Visualisierung der Ergebnisse unterstützt werden, indem die gefundenen Assoziationen graphisch darstellen werden. Abbildung 3 zeigt beispielhaft die Darstellungsweise von zwei-elementige Regeln für die Interessensgebiete Wirtschaft und Politik.

Support	Itemsets
14.007	[Lernen.] [Lernen.]
12.575	[Wissen.] [Wissen.]
11.670	[Lernen.] [Lernen.] [Lernen.]
9.476	[Woerterbu] [Woerterbu]
9.347	[Lernen.] [Lernen.] [Lernen.] [Lernen.]

Webseiten bestand, wurden die Seiten nur in die groben Kategorien [Lernen], [Wissen], [Wörterbuch], [Ratgeber] und [Aktuell] aufgeteilt.

Bei den gefundenen sequentiellen Mustern fällt auf, dass sich die User innerhalb einer Sitzung typischerweise nicht aus der einmal gewählten Hauptcontentgruppe entfernen (siehe Abbildung 4). Die Bereiche Lernen, Wissen und auch Wörterbücher stehen dabei im Fokus der User.

Sitzungen, in denen mehrere Bereiche besucht wurde, treten nur relativ selten auf, umfassen maximal zwei Bereiche und sind eher kurz (siehe Abbildung 5).

Dieses Verhalten bedeutet entweder, dass die Besucher sehr themen- bzw. problemorientiert sind, oder aber, dass der Anreiz für einen Bereichswechsel zu gering ist.

Abbildung 4: Sequenzielle Muster I

Support	Itemsets
3.998	[Wissen.] [Aktuell.]
3.656	[Aktuell.] [Wissen.]
3.399	[Wissen.] [Lernen.]
3.071	[Wissen.] [Woerterbu]
3.043	[Woerterbu] [Wissen.]

Ein Interessantes Navigationsmuster ist der Pfad [Wissen], [Woerterbuch]. Der User der diesen Pfad wählt stößt wahrscheinlich in einem Artikel im Wissen Bereich auf ein Wort oder eine Bezeichnung, die ihm nicht bekannt ist. Um die Bedeutung des Ausdrucks zu klären begibt er sich in den Bereich Wörterbuch, wo er das entsprechende Wort „nachsschlägt“. Dieses Wissen wurde zur Platzierung entsprechender Wörterbuch-Links genutzt.

Interessant wäre die Frage, ob sich die Vertreter der bereits ermittelten unterschiedlichen Besuchertypologien auch in ihren Navigationsmustern unterscheiden. Das würde z.B. die Entscheidung über die Art und Platzierung dynamischer Inhalte erleichtern. Diese Fragestellung kann aber aufgrund der dieser Analyse zugrunde liegenden Datenbasis zur Zeit noch nicht beantwortet werden.

Abbildung 5: Sequenzielle Muster II

5 Abgeleitete Massnahmen

Um das Potential der Web Mining-Analyse erhöhen zu können, müssen in erster Linie die Erfassungsmöglichkeiten der Logfiles betrachtet werden. Auf Basis der Logfiles, die im vorliegenden Beispiel Datengrundlage für die Analyse waren, ist nur im begrenzten Masse möglich, differenzierte Gruppen-spezifische Aussagen über Kundenverhalten zu machen.

Der genauen Kundenverhaltensspezifischen Datenerfassung im Internet und damit auch der darauf aufbauenden Analyse sind jedoch aufgrund der Besonderheiten des Internet-Protokolls zunächst bestimmte Grenzen gesetzt. Das Internet-Protokoll „http“ ist ein sog. zustandsloses Protokoll, das von sich aus nicht in der Lage ist, einzelne Informationsanforderungen (Requests) einer bestimmten „Session“ zuzuordnen. Eine Zuordnung von Logfile Einträgen

zu einzelnen Sessions über die IP-Adresse des Users ist nicht möglich, da die meisten Providern IP-Adressen dynamisch vergeben. Selbst statische IP-Adressen können oftmals nicht eindeutig einzelnen Personen zugeordnet werden. Bei großen Organisationen und Bildungseinrichtungen können z.B. mehrere Nutzer hinter einem Hostnamen (Router) stehen. Die Rechner können sich auch in PC-Pools befinden und von verschiedenen Personen benutzt werden.⁵

Um trotz dieser Internet-immanenten Limitationen tiefergehende Aussagen über das Kundenverhalten entwickeln zu können, wurde im vorliegenden Beispiel ein User-Tracking-System eingeführt. Dieses System erfasst auf einer genauen Session-Basis über die Logfiles hinaus CGI-Parameter (Benutzereingaben) und Keywords. Dies eröffnet die Möglichkeit, Aktivitäten der User (Angaben bei Postings, E-Mails, SMS, etc.) zu erkennen sowie semantischen Inhalte einer Seite zu erfassen. Semantische Schlagworte sind für die Kundensegmentierung mittels Data-Mining-Verfahren sowie die Vergleichbarkeit der Informationsplattform über lange Zeiträume hinweg ausgesprochen wichtig, da sich Struktur und URLs dieser Informationsplattform im Lauf der Zeit verändern können, wodurch eine Vergleichbarkeit auf dieser Ebene unmöglich wird.

Das aufgrund der Web Analyse eingesetzte Tracking-Verfahren basiert auf einem sog. Reverse-Proxy-Filter, der in den Kommunikationsstrom zwischen User und Webserver geschaltet wird. Dieser Filter fungiert als „Durchlauf-Filter“ der die gesamte Kommunikation zwischen Web-Clients und Webserver beobachtet und ggf. modifiziert. Dazu verhält sich der Reverse-Proxy gegenüber den Web-Clients wie der echte Webserver, und gegenüber dem Webserver wie ein Web-Client.⁶ (siehe Abbildung 6)



Abbildung 6: Funktionsweise des Reverse-Proxy-Filtersystems

Einerseits baut der Filter in allen URLs der ausgelieferten HTML-Seiten eindeutige Session-IDs ein und fügt an relevanten Stellen der Informationsströme (an wichtigen URLs) eindeutige Marker ein, die dann eine leichte Session-Zusammenführung ermöglichen. Andererseits wird der gesamte Inhalt von User-Anfragen und Server-Antworten gefiltert. Somit werden nicht nur inhaltliche Schlagworte und verschiedene andere kommunikationsspezifische Daten erfasst, sondern auch CGI-Parameter, mit deren Hilfe sowohl der E-Mail-Versand als auch andere Kennzahlen wie z.B. über Aktivitäten im Postingbereich nachvollzogen werden können.

Durch die Fähigkeit Inhalte zu erfassen ist es darüber hinaus möglich genau auszuwählen, welche URLs von Interesse sind, um so die Logfiles vor unnötigem Ballast zu schützen. Dadurch generiert dieses Verfahren im Vergleich zu anderen sehr reichhaltige und hochwertige Ausgangsdaten, und ermöglicht die Zusammenführung von Sessions sowie die Zuordnung dieser Sessions zu einzelnen Benutzern.

Vorteilhaft ist auch, dass der Webserver von der Aufgabe befreit wird, die Log-files aufzuzeichnen. Eine ständige Verbindung zur zentralen Datenbank ist nicht notwendig, so dass diese eine deutlich geringere Belastung bewältigen muss.

⁵ Vgl. Gentsch et al. (2001), S. 43-46

⁶ Vgl. Gentsch et al. (2001), S. 142-144

6 Fazit und Ausblick

Die dargestellten Business Cases zeigen, dass sich für das betrachtete e-Portal zahlreiche Ansatzpunkte bieten, um die Beziehungen zu Kunden, bzw. Usern mit Hilfe von Web Mining Analysen zu verbessern. Ziel ist es dabei meist, die Seite zu personalisieren, Inhalte und Angebote an den individuellen Nutzer anzupassen.

Auch ohne das beschriebene Tracking-System konnten bereits Ergebnisse generiert werden, die das Verständnis der Nutzerstruktur deutlich erhöht haben. Hierzu haben im Wesentlichen die Segmentierung und die Assoziationsanalyse beigetragen, die beide auf Basis der Registrierungsdaten durchgeführt worden sind. Weitere Verbesserungspotentiale bzgl. der Usability der Seiten hat die Analyse der Navigationspfade geliefert.

Die wahren Potentiale einer Web Mining Analyse hinsichtlich der Entdeckung verborgener Informationen über die Nutzer und der Erklärung und Prognose des Kundenverhaltens können jedoch jetzt erst durch die Implementierung der beschriebenen Tracking-Lösung realisiert werden.

Durch die geplante Verbindung der Logfiles und dem Registrierungsprozess, können in einem weiteren Schritt die personalisierte Ansprache und individuelle Behandlung der Kunden durch gezielte CRM-Maßnahmen eingeführt werden kann. Die gezeigten Inhalte können sich dann besser am Profil der Kunden orientieren und nur die Banner, die für einen Kunden auch mit hoher Wahrscheinlichkeit von Interesse sind, werden eingeblendet. Die Navigation innerhalb der Seite wird auf den einzelnen Besucher zugeschnitten, so dass die Suche nach Informationen und passenden Angeboten vereinfacht und beschleunigt wird. Dadurch, dass der Kundenwert prognostiziert und bei der Auswahl der Kundschaft berücksichtigt werden kann, steht dem Informationsportal die Möglichkeit zur Verfügung, sich einen profitablen Kundenkreis aufzubauen.

Es ist insbesondere hervorzuheben, dass ein entsprechendes User-Tracking mit aufbauendem Web Mining nicht nur wertvollen Input für die Personalisierung darstellt, sondern insbesondere auch wichtige Informationen für das Reporting und Controlling liefert. Kennzahlen wie z.B. „Clickzahl je Bereich im Verhältnis zur Verweilzeit je Bereich“, „Stickiness“ oder der „Personalisierungsindex“, die durch die viel zitierte E-Metrics⁷ postuliert werden, lassen sich nicht exakt auf Basis konventioneller Logfiles ermitteln. Auf Basis des erweiterten semantischen Tracking lassen sich bereits mit einfachen Reporting-Abfragen einige wichtige Kennzahlen generieren. Kunden-relevante Aussagen wie z.B. das Verhalten von bestimmten Kundenclustern, die Prognose von Warenkörben oder die Bestimmung signifikanter Navigationspfade benötigen jedoch zusätzlich sophistiziertere Analysen wie die hier beschriebenen Methoden des Web Mining.

Ein Aspekt, der in Zukunft eine immer wichtigere Rolle bei der Generierung von Kundenwissen spielen wird, ist die Verbindung reaktiver Verfahren (klassische Online-Marktforschung) mit nicht-reaktiven Verfahren (Web Mining).⁸ Um personalisierte Inhalte und Produkte bedarfsgerecht liefern zu können, benötigen Unternehmen differenziertes Wissen über ihre Kunden: Welche Interessen haben sie? Welche Kommunikationsform präferieren sie? Wie unterscheiden sich bestimmte Kundengruppen? Die Antworten hierzu kommen in der traditionellen Offline-Welt in der Regel aus der Marktforschung. Das Internet bietet eine einzigartige Vielzahl von Informationen, die es ermöglichen, den Puls des Kunden

⁷ Whitepaper zur E-Metrics ist unter www.netgen.com abrufbar.

⁸ Gentsch/ Roth/ Faulhaber, (2001), S. 349-367.

in Realtime zu messen. Den Puls des Kunden zu messen, heißt aus Sicht des Web Mining, zu beobachten und zu messen, was der Kunde tatsächlich in der Vergangenheit gemacht hat und aktuell gerade macht. Welche Inhalte interessieren ihn? Welche Produkte präferiert er? Damit kann die Frage beantwortet werden, was er macht oder gemacht hat. Die Fragen nach seiner Intention, nach seinen Wünschen, Sehnsüchten und Ängsten bleiben jedoch weitgehend unbeantwortet. Um wirklich fundiertes und differenziertes Wissen über Kunden und Märkte zu erhalten, sollten zukünftig stärker die Verhaltensebene und die Motivebene miteinander verbunden werden (siehe Abbildung 7). So kann der datengetriebene Ansatz des Web Mining zum Einen helfen, im Rahmen der Marktforschung auf Basis des tatsächlichen Verhaltens die richtigen Fragen zu stellen. Zum anderen lassen sich durch das tatsächlich erfasste Verhalten Aussagen validieren, die durch die Marktforschung gewonnen wurden. Genau diese Integration von klassischer Online-Marktforschung und Web-Mining wird derzeit für das vorliegende e-Portal im Form einer ASP-Lösung realisiert.

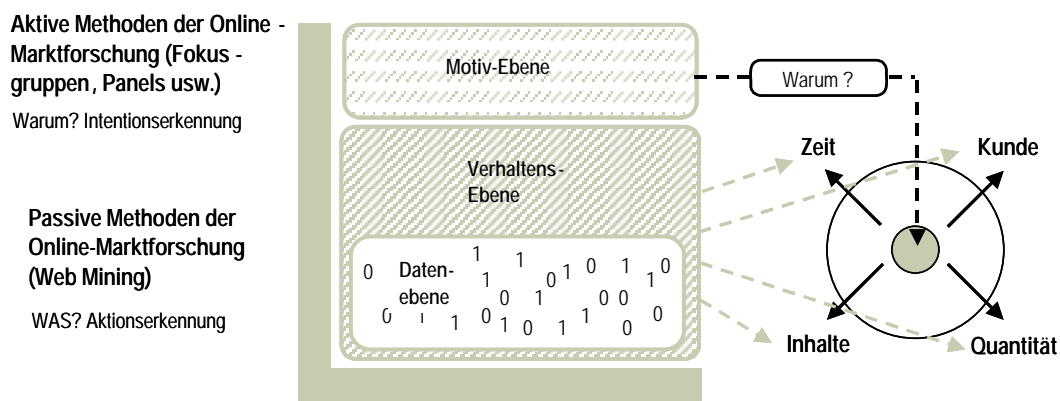


Abbildung 7: Verhaltensebene und Motivebene als Basis für die Personalisierung

Im Rahmen des Web Mining wird sich zunehmend auch die Diskussion um den gläsernen Kunden verschärfen. Dies betrifft in erster Linie die zugrunde liegende vertragliche Basis für die Sammlung und Nutzung von Daten für die Online Marktforschung und Personalisierung. Gibt es im Internet eine vertragliche Tradition? Impliziert bereits der Besuch einer Website einen „mündlich“ formlosen Vertrag des Einverständnisses zur Speicherung des Clickstreams, oder bedarf es explizit einer schriftlichen Vereinbarung? Wie ausdrücklich müssen Nutzungsvereinbarungen über Daten sein, die im Internet erfasst werden? Und folglich, wie ausdrücklich muss eine regelnde Politik sein, um die Privatsphäre schützen zu können? Die Unternehmen werden sich auf die Entwicklung der zunehmenden Sensibilität hinsichtlich personenbezogener Daten weiter einstellen müssen, um sich das Vertrauen der Konsumenten zurückzugewinnen und persönliche, langfristige Kundenbeziehungen aufbauen zu können. Vertrauensmanagement muss dem Konsumenten Klarheit darüber verschaffen, dass seine Daten ihm gehören, welchen Gegenwert er für die Preisgabe seiner Daten erhält und dass seine Daten vor dem Zugriff Dritter geschützt sind.⁹

⁹ Vgl. Gentsch, (2001).

Aus Sicht des analytischen CRM bzw. eCRM werden neben den strukturierten Daten zunehmend die unstrukturierten, qualitativen Daten im Internet wichtig werden. So enthalten Web-Seiten, E-Mails sowie Äußerungen in Chats und Newsforen wertvollen Input zur Analyse von Kunden und deren Verhaltensweisen. Die zum Data Mining analog für die weniger formatierten Daten einsetzbaren Analysetechniken werden unter dem Begriff „Text Mining“ oder auch „Content Mining“ diskutiert. Mit Hilfe dieser Analyseverfahren lassen sich z.B. früh Trends in Communities weitgehend automatisiert erkennen.

Insbesondere die Integration der verschiedenen Mining-Ansätze (Data, Text und Web Mining) wird es zukünftig ermöglichen, systematisch automatisierte Daten- und Marktforschungsanalysen im E-Business durchzuführen zu können. Ziel dieser Analysen ist die möglichst umfassende individuelle Ansprache des Kunden, die Optimierung des Leistungsangebotes sowie die Erhöhung der Kundenzufriedenheit und –bindung.

Mit zunehmenden internen und externen Daten- und Dokumentenvolumina wird sich das Potential des integrierten Mining für die Online-Marktforschung und Personalisierung weiter erhöhen. Die zunehmende Verbreitung des XML-Standards wird die Semantik der strukturierten und unstrukturierten Daten erhöhen und damit insbesondere die Anwendungsmöglichkeiten und ‚Entdeckungspotentiale‘ des integrierten Mining im Internet deutlich vergrößern. Insbesondere wird die voranschreitende Standardisierung im E-Business (z.B. Open Profiling Standard (OPS), Customer Profile Exchange (CPEX), E-Commerce Modeling Language (ECML), Common Log Format (CLF)) mit der einhergehenden Vereinheitlichung von Datentypen und Datensemantik die Möglichkeiten des Knowledge Mining im E-Business erweitern.¹⁰

¹⁰ Vgl. Gentsch et al. (2001), S. 204 ff.

Literatur:

Gentsch, Peter/ Roth, Martin/ Faulhaber, Nina.: Data Mining in der Online-Marktforschung – Auf dem zu gläsernden Märkten und Kunden? in: Theobald/ Dreyer/ Starsetzki (Hrsg.): Online-Marktforschung – Theoretische Grundlagen und praktische Erfahrungen Gabler, 2001, S. 349-367.

Gentsch, Peter et al.: Web-Personalisierung und Web-Mining für eCRM: 12 Tools im Vergleich OXYGON VERLAG, 2001.

Gentsch, Peter: Kundengewinnung und –bindung im Internet: Möglichkeiten und Grenzen des Analytischen eCRM, in: Schögel, M.: Report Electronic Customer Relationship Management (E-CRM) – eine neue Dimension der Kundenbeziehung Symposium, erscheint Ende 2001.

Grothe, Martin/ Gentsch, Peter: Business Intelligence „Aus Informationen Wettbewerbsvorteile gewinnen, ADDISON-WESLEY, 2000.

Kurzbiographien:

Dr. Peter Gentsch ist Director CRM/ Analytics der Business Intelligence Group (Frankfurt, Berlin), einem führenden Dienstleister im Bereich Business Intelligence/ Data Mining. Zudem hat er Lehraufträge an der Universität München, Aalen und Berlin.