September 2001

Web-Mining mit Methoden des Information Retrievals - Personalisierung von Web-Sites auf Basis von Webtracking Daten

Jürgen Schackmann
Universität Augsburg, juergen.schackmann@wiso.uni-augsburg.de

Matthias Knobloch
Universität Augsburg, matthias.knobloch@gmx.de

In: Buhl, Hans Ulrich, u.a. (Hg.) 2001. *Information Age Economy*; 5. Internationale Tagung Wirtschaftsinformatik 2001. Heidelberg: Physica-Verlag

ISBN: 3-7908-1427-X

Web-Mining mit Methoden des Information Retrievals – Personalisierung von Web-Sites auf Basis von Webtracking Daten

Jürgen Schackmann, Matthias Knobloch
Universität Augsburg

Zusammenfassung: Die Verwendung sog. Webtracking Daten wird häufig als wesentlicher Bestandteil einer Lösung propagiert, um über den Kunden Informationen und Wissen zu generieren, welches sowohl für ein ganzheitliches Multi-Channel-Customer-Relationship-Management verwendet werden kann, als auch zur Generierung personalisierter Empfehlungen im WWW-Kanal. In dieser Arbeit wird ein Rahmen skizziert, wie bisher bekannte Methoden im Bereich des Information Retrievals und der Recommender Systems im Web-Mining anwendbar sind. Es wird gezeigt, wie sog. Webtracking Daten semantisch angereichert, aggregiert repräsentiert und im Rahmen eines Customer Relationship Management für die Personalisierung eingesetzt werden können.

Schlüsselworte: Web-Mining, Webtracking, Personalisierung, Content Based Filtering, Collaborative Filtering, Recommender Systems, Information Retrieval

1 Einleitung

Die Anzahl der Angebote im WWW ist in den letzten Jahren ebenso drastisch angestiegen wie deren Nachfrager, und der Trend scheint ungebrochen [oV01]. Fast alle Produkte, Dienstleistungen, Informationen etc., die in der Realwelt existent sind, haben mittlerweile auch ein Pendant im WWW. Doch gerade diese Flut an Informationen führt bei den Nachfragern zu einem “Information Overload”, der zu steigenden Suchkosten und Frustration bei erfolgloser Suche führt. So konnten 20% der Befragten einer Studie eine Web-Site, auf der sie bereits waren, nicht wiederfinden oder 55% konnten eine Seite nicht finden, von der sie wussten, dass sie existiert [FiMa97].

Vor diesem Hintergrund gewinnt die Personalisierung [PiZa01] von Web-Sites zunehmend an Bedeutung [LiSc01,Lued97] bzw. ist für einige Branchen wie z.B. die Finanzdienstleistungsbranche mittlerweile ein wesentlicher Teil der Geschäftsstrategie [BuWo00]. Die Strategie der Personalisierung geht zurück auf das “One-to-One-Marketing” [PeRo97], bei dem jedem Kunden genau das Produkt angebo-

ten werden soll, welches seinen Zielen, Bedürfnissen und Präferenzen am besten entspricht [LiSc00]. Voraussetzung ist, das hierfür notwendige Wissen über den Kunden zu besitzen und dies in geeigneter Weise in elektronisch verarbeitbarer Form vorrätig zu halten. Die Verwendung sog. Webtracking Daten (WTD) wird dabei häufig als wesentlicher Bestandteil einer Lösung propagiert, um über den Kunden Informationen und Wissen zu generieren, welches sowohl für ein ganzheitliches Multi-Channel-Customer-Relationship-Management verwendet werden kann [FrSc00,PöSz01], als auch als Basis für die Generierung personalisierter Empfehlungen im WWW-Kanal dient [FrSt01].

Hierzu wird im zweiten Teil der Arbeit der Begriff des Webtracking (WT) und dessen Bedeutung für die Personalsieirung erläutert. Anschließend wird ein Framework für das Web-Mining auf Basis von WTD entwickelt, welches geeignet ist, die Ergebnisse sowohl im Rahmen eines Kundenmodells über verschiedene Kanäle zur Verfügung zu stellen als auch um hierauf aufbauend personalisierte Empfehlungen zu generieren. Im vierten Abschnitt werden Methoden aus dem Bereich der Recommender Systems diskutiert, die im Bereich des Web-Minings eingesetzt werden können.

2 Vom Webtracking zu personalisierten Empfehlungen

In dieser Arbeit werden WTD als Informationsquelle behandelt, um daraus in einem Web-Mining Prozess Wissen über den Kunden zu generieren. WT ist die automatische Generierung sog. Logfiles durch den jeweiligen Web-Server. In diesen werden in ihrer einfachsten und standardisierten Form (Common Logfile Standard) die von den Nutzern abgerufenen Dateien (Html und Bild), die Abrufzeitpunkte und die IP-Adresse des Abrufers festgehalten. Vielfach wird dieser Aufzeichnungsstandard von Websiteanbietern um weitere Informationen, beispielsweise Session-ID`s erweitert, um einen besseren Aussagegehalt über das Verhalten des Nutzers gewinnen zu können. So ermöglicht das Tracken der Session-ID die sichere Identifikation typischer Benutzerpfade während, bei alleiniger Nutzung der IP-Adressen durch Proxy-Server, Sammel-IP´s etc. die Validität der Aussagen stark sinkt. [FrSt01]

Der große erwartete Nutzen bei der Verwendung von WTD zur Personalisierung ist aber auch gleichzeitig deren Hauptproblem, da sie in sehr großer Anzahl je Kunden anfallen, zunächst wenig Semantik tragen, folglich von einem menschlichen Berater bzw. Verkäufer, nicht verwertbar sind und deshalb automatisiert verarbeitet werden müssen [AnKo00]. Im Rahmen des Web-Minings wurden bisher verschiedene Methoden zur Verarbeitung von WTD untersucht. So verwendet [ZuAl99] Markovmodelle oder [Cool00] Transaktionscluster. Des weiteren wer-

den in letzter Zeit auch Bayesnetze [SuKo01] und Assoziationsregeln [SrAg95] als geeignete Methoden diskutiert. Eine vergleichende Analyse dieser Verfahren findet sich bei [Fran01]. Den Autoren ist jedoch noch kein durchgängiges Konzept bekannt, welches die folgenden erfolgskritischen Kriterien erfüllt (vgl. auch [AnKo00]):

1. Die aktuelle Entwicklung zeigt, dass der Vertriebskanal Internet zunehmend zu einem wichtigen Kanal in einem Multi-Channel-Mix wird. Folglich ist es für ein kanalübergreifendes, konsistentes Multi-Channel-Customer-Relationship-Management unerlässlich, dass das gewonnene Wissen über den Kunden in aggregierter, semantik-tragender, konsistenter und situationsunabhängiger Form allen Kanälen zur Verfügung gestellt werden kann.
2. Dieses Wissen muss derart repräsentiert werden, dass es auch automatisiert für die Generierung personalisierter Empfehlungen weiterverarbeitet werden kann.
3. Die den Autoren bekannten Verfahren sind nicht oder nur bedingt in der Lage, den Prozess der Generierung von Wissen über den Kunden und der personalisierten Empfehlung durchgängig automatisiert und ohne menschliche Unterstützung wie Pre- oder Postprocessing oder der Analyse der Ergebnisse durchzuführen. Dies ist aber eine wesentliche Bedingung, um die in sehr großen Mengen und sehr häufig anfallenden WTD effizient und zeitnah zu verarbeiten.

Fridgen et al. haben gezeigt, dass Kundenmodelle geeignet sind, um eine Menge von Kundendaten zu abstraktem, aber allgemein einsetzbaren Wissen über den Kunden zu aggregieren sowie dieses Wissen persistent und konsistent über verschiedene Distributionskanäle verfügbar zu machen [FrSc00]. Gleichzeitig ist das Kundenmodell das elektronische Repository, aus dem personalisierte Empfehlungen generiert werden. Folglich sind Kundenmodelle insbesondere für das WT relevant, da die gewonnenen Informationen aus den Logfiles schon von Natur aus auf einem niedrigen Abstraktionsniveau sind und für die Personalisierung nur wenig Semantik enthalten. Gleichzeitig werden Produktmodelle benötigt und eingesetzt, um aus einer großen Menge von Produkten in automatisierten Prozessen personalisierte Produkte erstellen zu können [KuWo01]. [1]

Auf Basis dieser Ergebnisse wird in Kapitel 3 zunächst ein Framework entwickelt, welches in einem zweistufigen Inferenzprozess WTD in einem Kundenmodell verarbeiten kann und auf Basis eines Produktmodells personalisierte Empfehlungen generiert. In Kapitel 4 werden die geeigneten Methoden diskutiert, die mit den WTD in den Inferenzprozessen eingesetzt werden können.

---

[1] Für eine Übersicht über die verschiedene Ansätze, Produkte zu repräsentieren, siehe [SaFo83, Pazz01] bzgl. automatischer Klassifikationsmethoden bzw. [KuWo01] bzgl. analytischer Vorgehensweisen.

3 Framework

Im Folgenden wird ein Framework für die beschriebene Problemstellung entwickelt, welches die in Kapitel 2 erarbeiteten Kriterien erfüllt.

Kundenmodell K

Formal lässt sich das Kundenmodell als Kundenvektor darstellen, wobei jedes Element des Vektors eine Einstellung repräsentiert:

$$K^i = (K_1^i, \ldots, K_j^i, \ldots, K_n^i) \in K \quad (1)$$

mit: K_j^i ist die j-te Einstellung des i-ten Kunden

$i \in \{1, ..., k\}$, k = Anzahl der Kunden

$j \in \{1, ..., n\}$, n = Anzahl der Einstellungen

K = {Menge aller Kunden}

Der Begriff des Kunden beschränkt sich dabei nicht nur auf den Kunden einer Unternehmung, sondern kann ganz allgemein als Kunde von Web-Sites verstanden werden, der sowohl Produkte als auch nur die Informationen der Site nachfragen kann.

Produktmodell P

Formal lässt sich das Produktmodell als Vektor darstellen, dessen Elemente jeweils eine Produkteigenschaft beschreiben:

$$P^i = (P_1^i, \ldots, P_j^i, \ldots, P_m^i) \in P \quad (2)$$

mit: P_j^i ist die j-te Eigenschaft des i-ten Produkts

$i \in \{1, ... , h\}$, h = Anzahl der Produkte

$j \in \{1, ..., m\}$, m = Anzahl der Produkteigenschaften

P = {Menge aller Produkte}

Für die Problemstellung ist es jedoch notwendig, den Produktbegriff breiter zu fassen, als dies generell üblich ist: Unter Produkten werden im Folgenden alle nachfragbaren Angebote im WWW verstanden, d.h. dies können sowohl atomare Informationen, wie bspw. Wechselkurse oder Nachrichten, aber auch klassische Produkte sein.

Webtracking Daten C

Betrachtet man die WTD im Kontext des Kunden- und Produktmodells, so kann formal jedem Klick zugeordnet werden, von welchem Kunde dieser getätigt wurde und welches Produkt hinter diesem Klick zur Verfügung steht. Ein Klick C spezifiziert folglich den Kunden sowie das ausgewählte Produkt:

$$C_{i,j} = \begin{cases} +1 & \text{wenn Kunde i Produkt j angeklickt hat} \\ -1 & \text{sonst} \end{cases} \quad (3)$$

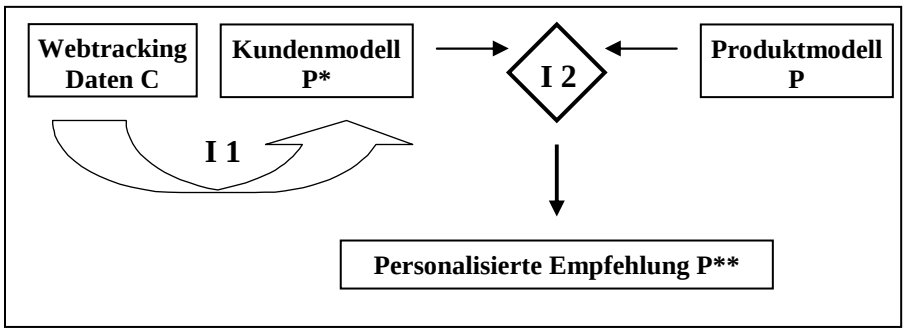

Abbildung 1: Framework mit zweistufigem Inferenzmechanismus

Auf Basis des Kunden- und Produktmodells sowie den WTD kann nun gemäß Fridgen et al. [FrVo00] ein zweistufiger Inferenzmechanismus definiert werden (vgl. Abbildung 1). Inferenzmechanismus 1 leitet aus den WTD des Kunden das jeweilige Kundenmodell ab.

$$\text{I1}: F(C) \Rightarrow K \quad (4)$$

Da im vorliegenden Fall die WTD nur Informationen über die Präferenzen des Kunden bzgl. Verschiedener Produkte vorliegen, kann nur auf die Produkteinstellungen des Kunden geschlossen werden. Folglich ist das Ergebnis von I1 ein ideales Produktmodell P* und es gilt: P*$\subset$ K. Da das allgemeine Kundenmodell nicht Gegenstand dieser Arbeit ist, kann (4) folglich vereinfacht werden:

$$\text{I1}: F(C) \Rightarrow P^* \quad (4')$$

Inferenzmechanismus 2 repräsentiert demgegenüber den Beratungs- bzw. Verkaufsprozess. Aufbauend auf dem Kundenmodell werden diejenigen Produkte gesucht, von denen angenommen wird, dass sie dem Kunden maximalen Nutzen generieren. Hierbei können über die WTD hinaus noch andere Daten, Informationen und Methoden mit einfließen, das Ergebnis ist jedoch wiederum ein optimales Produktmodell P**.

$$\text{I2}: F(P^*) \Rightarrow P^{**} \quad (5)$$

Ziel des gesamten Inferenzprozesses (I1 und I2) ist es, dem Kunden diejenigen Produkte anzubieten, die seinen Nutzen maximieren.

$$U(P^{**}) \xrightarrow{!} Max \tag{6}$$

4 Methodik

Nachdem keine der gängigen Web-Mining Methoden die in Kapitel 2 aufgestellten Kriterien erfüllt, werden im Folgenden Methoden aus dem Information Retrieval untersucht. Es wird gezeigt, wie diese eingesetzt werden können, damit diese Kriterien erfüllt sind. Im Rahmen des Frameworks sind jedoch auch andere Methoden beliebig einsetzbar.

4.1 State-of-the-Art

Die aufgezeigte Problemstellung – die Selektion bestimmter Objekte aus einer großen Menge potenzieller Objekte nach den Bedürfnissen eines einzelnen Nutzers – entstand nicht erst mit der Entwicklung des WWW. Derartige Problemstellungen werden bereits seit einigen Jahrzehnten unter dem Stichwort „Information Retrieval“ diskutiert [SaFo83]. Die hierbei entwickelten Methoden und Lösungen erfuhren in letzter Zeit eine Renaissance in sog. „Recommender Systems“, die sich i.d.R. speziell mit der Generierung von personalisierten Empfehlungen für bestimmte Problemdomänen im WWW beschäftigen [ReVa97]. Ein prominentes Beispiel ist hier der Internetbuchhändler Amazon (www.amazon.com). Die zwei wesentlichen Methoden zur Generierung personalisierter Empfehlungen sind einerseits das Content based Filtering (CbF), und andererseits das Collaborative Filtering (CF) [Bala97].

a) Content based Filtering

CbF generiert Empfehlungen auf Grund der Inhalte des bisher von einem Nutzer betrachteten Contents sowie des Inhalts des potenziell zu empfehlenden Contents [UnFo00, Pazz01, Bala97]. D.h. im ersten Schritt aufgrund der vom Nutzer bisher angeklickten und nicht angeklickten Seiten ein Nutzerprofil erstellt, auf dessen Basis im nächsten Schritt bestimmt wird, welchen Content dieser Nutzer als nächstes auswählen würde, wenn er eine komplette Übersicht hätte. Zwingende Voraussetzung für die Anwendung des CbF ist folglich, dass eine **Repräsentationsform der Beschreibung des Contentinhalts** existiert, da ansonsten keine Informationen hierüber zur Verfügung stehen. Des weiteren muss eine **Methode zur Erstellung der Nutzerprofile** und ein **Mechanismus zur Vorhersage**, welchen

noch nicht betrachteten Content der Nutzer als nächstes sehen möchte, definiert werden.

Diese Vorgehensweise führt zu einigen Nachteilen [ClGo01]:

- Die inhaltliche Beschreibung des Contents enthält keine Aussagen über die Qualität.
- Es herrscht eine sehr statische Nutzersicht vor, d.h. es wird von einem Nutzer ausgegangen, dessen Bewertung von Content sich nur selten und langsam ändert
- Mit steigender Anzahl an potenziell zu empfehlendem Content wird CbF immer ineffektiver, da trennschwächer.
- Grundsätzlich wertvolle, abstrakte Zusammenhänge über das Verhalten aller Kunden wird nicht berücksichtigt.

b) ***Collaborative Filtering***

CF gibt Empfehlungen auf Basis des Verhaltens oder der Empfehlungen anderer im System vorhandener Nutzer [ClGo01, UnFo00, Pazz01]. D.h. einem Nutzer wird derjenige Content empfohlen, der von anderen Nutzern - die dem Nutzer sehr ähnlich sind - bereits angeklickt wurde. Die zwingende Voraussetzung für die Anwendung von CF ist die **Existenz** und eine **Repräsentationsform für das Kundenprofil**. Des weiteren muss ein **Ähnlichkeitsmaß zur Identifizierung ähnlicher Kunden** sowie ein **Mechanismus zur Vorhersage**, welchen noch nicht betrachteten Content der Nutzer als nächstes sehen möchte, definiert sein.

Nachteile [ClGo01, Pazz01]:

- Bei Content, der neu zur Verfügung steht, oder Nutzern, die noch keine oder nicht ausreichend viele Bewertungen abgegeben haben, funktioniert CF nur unzureichend oder gar nicht.
- Daten über den bisher von einem Kunden betrachteten Content werden beim CF überhaupt nicht berücksichtigt.
- „Herdentrieb"-Phänomene und selbstverstärkende Effekte, die zu irrationalem Verhalten führen, werden erhöht.

c) Collaboration via Content (CvC)

Aufgrund der diskutierten Nachteile wurden in letzter Zeit vermehrt Ansätze entwickelt, die beide Methoden miteinander kombinieren, sog. Collaboration via Content, und hierdurch die aufgezeigten Nachteile erheblich reduzieren können. [ClGo01], Pazz01, ClGo01, Bala97] haben empirisch belegt, dass diese Kombination zu signifikanten Performanceverbesserung führt. Grundsätzlich kommt diese dadurch zustande, dass das im CbF generierte Nutzerprofil zur Berechnung des

Ähnlichkeitsmaßes des CF herangezogen wird. Hieraus ergibt sich die folgende Vorgehensweise:

1. Erstellung der Nutzerprofile auf Basis des Contentinhalts des betrachteten Contents.
2. Berechnung der Ähnlichkeit zwischen verschiedenen Nutzern auf Basis der Profile.
3. Vorhersage der erwünschten Contents und damit Generierung einer personalisierten Empfehlung.

d) Bewertung

Sowohl das CF als auch das CbF weisen erhebliche Nachteile auf, die in Einzelfällen zu unbrauchbaren Ergebnissen führen. Das CvC hingegen kann durch die Kombination der beiden vorhergehenden Ansätze deren Nachteile fast vollständig eliminieren, insbesondere wenn über die Art der Kombination situativ entschieden werden kann. Diese theoretischen Vorteile konnten auch in empirischen Untersuchungen bestätigt werden. Folglich scheint für die Anwendung im Framework nur das CvC lohnenswert.

4.2 Collaboration via Content im Framework

Im Folgenden soll nun gezeigt werden, wie die zwei in Kapitel 3 definierten Inferenzstufen 1 und 2 mit Hilfe des kombinierten Ansatzes (CvC) durchgeführt werden können.

4.2.1 Ableitung des Kundenmodells

Gemäß des Frameworks wird im ersten Schritt in I1 aus den vorhandenen WTD auf das Kundenmodell geschlossen. Die bekanntesten und aus dem Information Retrieval stammenden Methoden sind Rocchios oder Winnows Algorithmus. Ein anderer, in diesem Zusammenhang bisher kaum verwendeter Ansatz, ist die klassische Regression.

Rocchios's Algorithmus ist die mit am weitesten verbreitete und angewendete Lernmethode im Information Retrieval [Joac97]. Er wurde bspw. erfolgreich zur Erzeugung von Kundenmodellen für Nachrichten [Lang95] oder für Web-Sites [Bala97] verwendet. Rocchios Algorithmus hat jedoch den Nachteil, dass die Anzahl der verwendeten Produktkategorien bekannt sein muss. Bei Anwendungsfällen, die diese Bedingung nicht erfüllen, kann der **Winnows Algorithmus** [BlHe95] verwendet werden, der bspw. bei Kundenmodellen für Restaurants eingesetzt wurde [Pazz01].

Die **Regression** ist eine Methode der Statistik, die bisher in den Disziplinen des Information Retrieval und der Recommender Systems kaum eingesetzt wurde. Durch die Verwendung einer mittels der Regression zu schätzenden Nutzenfunktion können jedoch auch hier diskrete Kundenentscheidungen modelliert werden [Gree97, Davi83]. Die Gewichte der Nutzenfunktion stellen somit das Kundenmodell dar. Bei der Schätzung diskreter Funktionen sollten jedoch nicht-lineare Regressionsmodelle verwendet werden, die erheblich komplexer in einem iterativen Prozess berechnet werden müssen [Gree97].

4.2.2 Personalisierte Empfehlung

Im Inferenzprozess 2 wird von den Kundenmodellen auf personalisierte Empfehlungen geschlossen. Zunächst wird die Ähnlichkeit oder auch nicht Ähnlichkeit zwischen den einzelnen Kunden auf Basis von P* festgestellt. Hierauf aufbauend wird P** berechnet und eine Empfehlung gegeben.

a) Ähnlichkeitsberechnung – Proximitätsmaße

Prinzipiell werden Ähnlichkeitsmaße als Grundlage für die im nächsten Schritt durchzuführende Prognose benötigt. Sie werden dabei in Form von Gewichten eingesetzt, so dass über sie gesteuert werden kann, welche Kunden wie stark in die Prognose miteinbezogen werden sollen. In der Literatur finden sich eine Fülle von Algorithmen zur Bestimmung von Proximitätsmaßen, wobei hier nur auf zwei Ansätze näher eingegangen werden soll. Eine ausführliche Behandlung der distanz- sowie korrelationsbasierten Ähnlichkeitsmaße findet sich bei [Runte00]. Zur Vergleichbarkeit der verschiedenen Ansätze haben die Proximitätsmaße Q folgende Bedingung zu erfüllen:

(6) $-1 < Q < +1$, wobei gilt: je höher Q, desto ähnlicher.

Distanzbasierter Ansatz

Distanzbasierte Ansätze gehören sicherlich zu den am häufigsten verwendeten Proximitätsmaßen. Zu nennen sind vor allem die euklidische Distanz sowie die City-Block-Metrik, die jeweils eine spezielle Form der allgemeinen Minkowski-L_q-Metrik sind. Distanzbasierte Ansätze summieren die jeweiligen Differenzen zweier Merkmale, jeweils mit einen konstanten Faktor gewichtet, auf. Daraus ergibt sich ein Wert für die Distanz (Unähnlichkeit) zweier Objekte, der mittels linearer Transformation in ein Ähnlichkeitsmaß überführt werden kann.

Die Distanz $D_{i,j}$ zwischen zwei Kunden K^i und K^j berechnet sich mit Hilfe von Minkowskis L_q-Metrik:

$$D_{i,j} = \left(\sum_{r=1}^{k} \left| P**_r^j - P**_r^i \right|^q \right)^{\frac{1}{q}}, \qquad (7)$$

wobei q, mit q>0, als ein Maß der Nicht-Linearität interpretiert werden kann

Damit Bedingung (6) erfüllt ist, muss folgende Transformation vorgenommen werden:

$$Q_{i,j} = 1 - \frac{2 \cdot D_{i,j}}{\max\limits_{i} D_{i,j}} \tag{8}$$

Korrelationskoeffizient

Häufig wird auch der Bravais-Pearson-Korrelationskoeffizient zur Berechnung der Ähnlichkeit verwendet. Der Wert für die Ähnlichkeit ergibt sich unmittelbar und nicht wie beim distanzbasierten Ansatz über den Umweg eines Distanzmaßes. Es ist allerdings zu beachten, dass solche Korrelationskoeffizienten immer einen linearen Zusammenhang zwischen den betrachteten Objekten untersuchen. Der Korrelationskoeffizienten wird wie folgt berechnet:

$$Q_{i,j} = \frac{\sum_{r=1}^{k} \left(P*^{i}_{r} - \overline{P}*^{i}\right) \cdot \left(P*^{j}_{r} - \overline{P}*^{j}\right)}{\sqrt{\sum_{r=1}^{k} \left(P*^{i}_{r} - \overline{P}*^{i}\right)^2 \cdot \sum_{r=1}^{k} \left(P*^{j}_{r} - \overline{P}*^{j}\right)^2}} \tag{9}$$

Eine Transformation ist nicht notwendig, da der Pearsonsche Korrelationskoeffizient die Bedingung (6) bereits erfüllt.

b) Personalisierte Empfehlung

Aufgrund des Ähnlichkeitsmaßes Q kann nun für Kunde i die Produktdimension k $P**_k^{\,i}$ folgendermaßen berechnet werden:

$$P**_k^{\,i} = \frac{1}{\sum_{l=1, l\neq i}^{k} Q_{i,l}} \sum_{j=1, j\neq i}^{k} P*_k^{\,j} \cdot Q_{i,j} \tag{10}$$

Aus den in 4.1 genanten Gründen kann für die Generierung der personalisierten Empfehlung je nach Situation sowohl auf P* als auch P** zurückgegriffen werden. Dabei wäre es einerseits möglich, eine Konvexkombination zu verwenden, deren Gewichtung statistisch geschätzt wurde. Andererseits könnte die Gewichtung auch situativ-dynamisch bestimmt werden, je nach dem, wie stark sich der kollaborative Ansatz für einen Kunden in einer bestimmten Situation eignet.

Mit Hilfe der L_q–Metrik (7) wird nun die Distanz D zwischen einem vorhandenen, potenziell zu empfehlenden Produkt und dem ermittelten optimalem Produktmo-

dell berechnet. Der Kunde bekommt dann die Produkte empfohlen, die die geringste Distanz D besitzen.

5 Bewertung und Ausblick

In dieser Arbeit konnte sowohl ein Rahmen skizziert werden, mit dessen Hilfe WTD mit entsprechenden Methoden zur Personalisierung und zum Customer Relationship Management eingesetzt werden können, als auch gezeigt werden, wie bisher bekannte Methoden im Bereich des Information Retrieval und der Recommender Systems zum Web-Mining eingesetzt werden können. Dieser Ansatz hat gegenüber bisherigen Ansätzen vor allem die folgenden Vorteile:

- Das Verfahren reagiert dynamisch auf neue WTD, der gesamte Prozess ist automatisiert und bedarf keines händischen Pre- bzw. Postprocessings.
- Mit dem Kundenmodell existiert eine abstrakte und aggregierte Repräsentationsform des in den WTD vorhandenen Wissens über den Kunden, welches über verschiedene Kanäle situationsunabhängig für verschiedene Zwecke zur Verfügung gestellt werden kann.
- Auf Basis des Kundenmodells können automatisiert personalisierte Produktempfehlungen generiert werden.

Im Gegensatz zu den bisherigen Recommender Systems wird im Kontext dieser Arbeit statt des (ordinalen) Ratings eines Kunden für einen bestimmten Content nur die binäre Information verwendet, ob Produkte, die dem Kunden – als Link, Abstract o.ä. – angezeigt wurde, von dem Kunden angeklickt wurden (vgl. (3)). Aus diesem Grund lassen die bisher erarbeiteten empirischen Ergebnisse sich nicht notwendiger Weise unbesehen auf die Problemstellung übertragen. Folglich wurden bisher einige Fragen offengelassen bzw. unterschiedliche Alternativen nicht diskutiert, die sich durch eine analytische Vorgehensweise nicht klären lassen, sondern weitere empirische Forschung erfordern: So muss überprüft werden, welche der vorgeschlagenen Methoden (Regression, Rocchio, Winnow) zur Bestimmung von P* am geeignetsten ist, ob ein transformiertes lineares Distanzmaß oder ein Ähnlichkeitsmaß zur Bestimmung der Proximität eher geeignet ist oder wie die freien Parameter, bspw. q zur Berechnung der L_q-Metrik, gewählt werden sollten.

Allerdings wird die notwendige empirische Datenerhebung für das Framework im Zuge zukünftiger Forschung durch derzeitige technische Einschränkungen des WT erschwert: Heutige Logfiles im Common Logfile Standard enthalten keine detaillierten (Meta-) Informationen über den jeweiligen Kunden bzw. das angeklickte Produkt. Diese Informationen müssen nachträglich mit den Logfile Informationen gematcht werden. Dies ist zwar technisch problemlos möglich, die hierfür benö-

tigte Infrastruktur ist bei den meisten Anbietern jedoch noch nicht standardisiert vorhanden. Aus diesen Gründen wird unsere zukünftige Forschung in drei Stufen ablaufen. Im ersten Schritt werden auf Basis von angenommenen Nutzenfunktionen Logfile-Daten simuliert, die zu einer Überprüfung und ersten Verfeinerung der Methodik eingesetzt werden. Vorausgesetzt der erste Schritt verläuft erfolgreich, werden im zweiten Schritt Laborexperimente durchgeführt. Nach Abschluss dieser beiden Schritte wird letztendlich mit realen Logfiles gearbeitet (in der Hoffnung, dass bis zu diesem Zeitpunkt die technischen Einschränkungen durch entsprechende Weiterentwicklung und evtl. die Einführung einer entsprechenden Standardsoftware aufgehoben sind).

Literatur

[Bala97] Balabanovic, Marko: An Adaptive Web Page Recommendation Service. In: First International Conference on Autonomous Agents, Marina del Ray CA, USA (1997).

[BlHe95] Blum, A.; Hellerstein, L. et al.: Learning in the Presence of Finitely or Infinitely Many Irrelevant Attributes. In: Journal of Computer and System Sciences, 50 (1995), S. 32-40.

[BuWo00] Buhl, Hans Ulrich; Wolfersberger, Peter: Neue Perspektiven im Online- und Multichannel Banking. In: Locarek-Junge, H.; Walter, B. (Hrsg.): Banken im Wandel: Direktbanken und Direct Banking, Berlin-Verlag, Berlin 2000, S. 247-268.

[ClGo01] Claypool, M.; Gokhale, A. et al.: Combining Content-Based and Collaborative Filters in an Online Newspaper. In: ACM SIGIR Workshop on Recommender Systems – Implementation and Evaluation, Berkeley CA, USA 1999.

[Cool00] Cooley, R.: Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. Ph.D. Thesis, Department of Computer Science, University of Minnesota, 2000.

[Davi83] Davison, Mark: Multidimensional Scaling. John Wiley&Sons, New York NY, USA 1983.

[FiMa97] Fittkau, S.; Maaß, H: W3b-Uni-Ergebnisband, WWW-Benutzeranalyse, Oktober/November 1997, Hamburg 1997.

[FrSc00] Fridgen, M.; Schackmann, J. et al.: Preference Based Customer Models for Electronic Banking. In: Hansen, H.-R., Bichler, M., Mahrer H., Hrsg., Proceedings of the 8th European Conference on Information Systems ECIS 2000, Wien, Österreich 2000, Volume 2, S. 819-825.

[Fran01] Frank, Melanie: Web Usage Mining zur Personalisierung. Diplomarbeit an der Wirtschafts-und Sozialwissenschaftlichen Fakultät der Universität Augsburg, 2001.

[FrSt01] Fridgen, Michael; Steck, Werner: New Perspectives of Website Controlling by Usage Mining. Eingereicht für: The Thirty-Fifth Annual Hawaii International Conference on System Sciences (HICSS-35), 2001.

[FrVo00] Fridgen, M.; Volkert, S. et al.: Kundenmodell für eCRM – Repräsentation individueller Einstellungen. In: 3.FAN-Tagung 2000, Siegen 2000.

[Gree97] Greene, William: Econometric Analysis. Prentice Hall, Upper Saddle River NJ, USA 1997, S. 871-947.

[ItLe95] Ittner, D.; Lewis, D.; et al.: Text Categorization of low quality images. In: Symposium on Document Analysis and Information Retrieval, Las Vegas 1995, S. 301-315.

[Joac97] Joachims, Thorsten: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: Proceedings of International Conference on Machine Learning (ICML), 1997.

[KuWo01] Kundisch, Dennis; Wolfersberger, Peter et al.: Enabling eCCRM: Content Model and Management for Financial eService. In: Sprague, Ralph (Hrsg.): Proceedings of the Thirty-Fourth Annual Hawaii International Conference on System Sciences (HICSS-34), Vol. VII, Internet and the Digital Economy Track, IEEE Computer Society Press, Los Alamitos, Hawaii 2001.

[Lang95] Lang, K.: News Weeder: Learning to Filter Netnews. In: International Conference on Machine learning, 1995.

[LiSc00] Link, Hubert; Schackmann, Jürgen: Ein ökonomisches Modell für die Produktion individueller digitaler Produkte. In: Bodendorf, F.; Grauer, G. (Hrsg.): Verbundtagung Wirtschaftsinformatik 2000, Siegen, Oktober 2000, Shaker, Aachen 2000, S. 192 - 207.

[LiSc01] Link, Hubert; Schackmann, Jürgen: Individuelle digitale Güter und Leistungen im Electronic Commerce. Diskussionspapier des Lehrstuhls für Betriebswirtschaftslehre, Universität Augsburg, 2001.

[Lued97] Lüdi, Ariel: Personalize or Perish. In: Schmid, Beat F.; Selz, Dorian et al. (Hrsg): EM - Electronic Product Catalogs. EM - Electronic Markets, Vol. 7, No. 3, 1997.

[oV01] o.V: Studie: Jeder dritte Deutsche ist „drin“. In: Computerwoche, Nr.6, 9. Februar 2001.

[Pazz01] Pazzani, Michael: A Framework for Collaborative, Content-Based and Demographic Filtering. Erscheint in: Artificial Intelligence Review. http://www.ics.uci.edu/~pazzani/Publications/AIREVIEW.pdf., Abruf am 2001-02-28.

[PeRo97] Peppers, Don; Rogers, Martha: The one to one future. Currency Doubleday, New York 1997.

[PiZa01] Piller, Frank; Zanner, Stefan: Mass Customization und Personalisierung im Electronic Business. In: WISU 1/01 (2001), S. 88-96.

[PöSz01] Pöttgens, Ulrich; Szinovatz, Andreas et al.: Bankkunden erwarten individuelle Leistungen und Informationen. In: Computerwoche Nr. 6, 9. Februar 2001.

[ReVa97] Resnick, Paul; Varian, Hal: Recommender Systems. In: ACM 1997-03-00, Vol. 40(3).

[Rocc71] Rocchio, J.: Relevance Feedback in Information Retrieval. In: Salton, Gerard (Hrsg): The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, 1971, S. 313-323.

[Runt00] Runte, M.: Personalisierung im Internet – Individualisierte Angebote mit Collaborativ Filtering, Dissertation an der Universität Kiel, Wiesbaden 2000.

[SaFo83] Salton, Gerard; Fox, E.A. et al.: Advanced Feedback Methods in Information Retrieval. In: Journal of the American Society for Information Science, 36(3) 1983, S. 200-210.

[ScLi01] Schackmann, Jürgen; Link, Hubert: Mass-Customization of Digital Products in Electronic Commerce. In: Innovation Science Innovation, Workshop on Information Systems for Mass Customization, Dubai 2001.

[ScKo00] Schafer, B.; Konstan, J. et al.: E-Commerce Recommendation Applications. In: Journal of Data Mining and Knowledge Discovery, vol. 5 nos. 1/2, pp 115-152.

[SrAg95] Sikant, R.; Agrawal, R.: Mining Generalized Association Rules. In: Proceedings of the 21th VLDB Conference, Zürich 2000.

[SuKo01] Ansari, Suhail; Kohavi, Ron et al.: Integrating E-Commerce and Data Mining: Architecture and Challenges. http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/ronny.pdf, Abruf am 2001-05-31.

[UnFo00] Ungar, Lyle; Foster, Dean: A Formal Statistical Approach to Collaborative Filtering. In: Conference on Automated Learning and Discovery (2000).

[ZuAl99] Zuckermann, I.; Albrecht, D.W.; Nicholson, A.E.: Predicting Users' Requests on the WWW. In: Proceedings of the 7th International Conference of User Modelling UM99, S. 275-284, 1999.

[AnKo00] Ansari, Suhail; Kohevi, Ron et al.: Integrating E-Commerce and Data Mining: Architecture and Challenges. In: WEBKDD'2000: Web Mining for E-Commerce-Challenges and Opportunities, 2000, Boston, MA.