Universität Augsburg
Prof. Dr. Hans Ulrich Buhl
Kernkompetenzzentrum
Finanz- & Informationsmanagement
Lehrstuhl für BWL, Wirtschaftsinformatik,
Informations- & Finanzmanagement

Universität
Augsburg
University

Diskussionspapier WI-96

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

von

Jürgen Schackmann, Matthias Knobloch

Juni 2001

in: Buhl, H. U., Huther, A., Reitwiesner, B., Hrsg., Information Age Economy, Augsburg, 2001, Physica, Heidelberg, 2001, S.279-292

Universität Augsburg, 86135 Augsburg
Besucher: Universitätsstr. 12, 86159 Augsburg
Telefon: +49 821 598-4801 (Fax: -4899)
www.fim-online.eu

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

Jürgen Schackmann, Matthias Knobloch

Lehrstuhl für Betriebswirtschaftslehre mit Schwerpunkt Wirtschaftsinformatik, WiSo-Fakultät, Universität Augsburg, 86135 Augsburg

E-mail: juergen.schackmann@wiso.uni-augsburg.de,

matthias.knobloch@student.uni-augsburg.de

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

Zusammenfassung: Die Verwendung sog. Webtracking Daten werden häufig als wesentlicher Bestandteil einer Lösung propagiert, um Informationen und Wissen über den Kunden zu generieren, das sowohl für ein ganzheitliches Multi-Channel-Customer-Relationship-Management verwendet werden kann, als auch zur Generierung individueller Empfehlungen im WWW-Kanal. Gleichzeitig gewinnt die Individualisierung von Web-Sites zunehmend an Bedeutung bzw. ist für einige Branchen wie z.B. die Finanzdienstleisungsbranche mittlerweile ein wesentlicher Teil der Geschäftsstrategie. In dieser Arbeit wird ein Rahmen skizziert, wie bisher bekannte Methoden im Bereich des Information Retrieval und der Recommender Systems zum Web-Mining eingesetzt werden können. Es wird gezeigt, wie sog. Webtracking Daten semantisch angereichert, aggregiert repräsentiert und im Rahmen eines Customer Relationship Management für die Individualisierung eingesetzt werden können.

Schlüsselworte: Web-Mining, Webtracking, Individualisierung, 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 [Cowo01]. Fast alle Produkte, Dienstleistungen, Informationen, etc., die in der Realwelt existent sind, haben mittlerweile auch ein Pendant im WWW – auch wenn dies evtl. nur die Homepage eines Friseurs oder der Abstract eines wissenschaftlichen Artikels ist. 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 ein Web-Site, auf der sie bereits waren, nicht wiederfinden oder 55% konnten eine Seite nicht finden, von der sie wussten, dass sie existiert [W3B97].

Vor diesem Hintergrund gewinnt die Individualisierung von Web-Sites zunehmend an Bedeutung [LiSc01,Lued97] bzw. ist für einige Branchen wie z.B. die Finanzdienstleisungsbranche mittlerweile ein wesentlicher Teil der

Geschäftsstrategie [BuWo00]. Die Strategie der Individualisierung geht zurück auf das “One-to-One-Marketing” [PeRo97], bei dem jedem Kunden genau das Produkt angeboten 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) werden dabei häufig als wesentlicher Bestandteil einer Lösung propagiert, um Informationen und Wissen über den Kunden zu generieren, das sowohl für ein ganzheitliches Multi-Channel-Customer-Relationship-Management verwendet werden kann [FrSc00,PöSz01], als auch zur Generierung individueller Empfehlungen im WWW-Kanal [FrSt01].

Es stellt sich herbei jedoch das Problem, dass WTD in sehr großer Anzahl je Kunden anfallen, diese zunächst wenig Semantik tragen, händisch deshalb nicht verwertbar sind und deshalb automatisiert verarbeitet werden müssen. Bisher existiert jedoch noch kein durchgängiges Konzept, wie WTD automatisiert verwertet werden können, so dass einerseits die Ergebnisse direkt in die Individualisierung einfließen können und andererseits die sehr große Menge an Daten aggregiert und semantiktragend in ein Kundenmodell integriert werden können, so dass das so gewonnene Wissen über alle Kanäle verfügbar gemacht werden kann.

In dieser Arbeit wird ein Konzept entwickelt, wie WTD im Rahmen eines Data Mining Prozesses nutzbar gemacht werden können, um diese im Rahmen des CRM über alle Kanäle zur Verfügung zu stellen, also auch um dem Kunden individualisierte Empfehlungen zu generieren. Hierzu wird im zweiten Teil der Arbeit der Begriff des Webtracking und dessen Bedeutung für die Individualisierung 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 individualisierte Empfehlungen zu generieren. Im 4 Abschnitt werden Methoden diskutiert, die im Bereich der Recommender Systems bereits ausführlich erforscht wurden, und diskutiert wie diese für das Web-Mining eingesetzt werden können.

2 Vom Webtracking zu individualisierten Empfehlungen

In dieser Arbeit werden sogenannte Webtracking Daten als neue Infomations-quelle behandelt, um daraus in einem Web-Mining Prozess Wissen über den Kunden zu generieren. Das Webtracking (WT) findet statt durch 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 Aufzeichnungs-standard 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]

WTD liefern also einen Fülle von Informationen, die ohne geeignete Auswertung weder für CRM noch für die Generierung individualisierter Empfehlungen große Aussagekraft besitzen. Ziel dieser Arbeit ist folglich, einen Web-Mining Prozess zu definieren, um die Informationen zu aggregieren und daraus geeignete Schlüsse zu ziehen,. Fridgen et al. haben gezeigt, dass Kundenmodelle das geeignet Mittel sind, um eine Menge von Kundendaten über Inferenzmechanismen zu abstraktem, aber allgemein einsetzbaren Wissen über den Kunden zu aggregieren, dieses Wissen persistent und konsistent über verschiedene Distributions-Kanäle verfügbar zu machen [FrSc00]. Gleichzeitig ist das Kundenmodell das elektronische Repository, um daraus individualisierte Empfehlungen generieren.

Sowohl für die Generierung der individualisierten Empfehlungen als auch für die Ableitung des Kundenmodells sind Inferenzmechanismen notwendig. In dieser Arbeit wird eine spezielle Form der Inferenzmechanismen untersucht: Collaborative Filtering und Content based Filtering sind zwei Methoden die sowohl isoliert als auch in Kombination in letzter Zeit intensiv untersucht und erfolgreich angewendet wurden, um Kunden individualisierte Empfehlungen zu generieren (Beispiele im WWW finden sich bei [PaBi97, ClGo01]). Im Gegensatz zu dem in dieser Arbeit verfolgten Ansatz wurde jedoch bisher keine dieser Methoden verwand, um speziell WTD zu verarbeiten.

3 Framework

Im Folgenden wird ein Framework beschriebene Problemstellung entwickelt. Das Framework und die darin enthaltene formale Darstellung stellt die Grundlage für die weitere Untersuchung dar. Elementar ist dabei die Aufteilung in Kundenmodell auf der einen -und Produktmodell auf der anderen Seite. Die Schlußfolgerungen finden über zwei Inferenzmechanismen I1 und I2 statt. I1 hat die Aufgabe, von bereits gewonnen WTD auf das Kundenmodell zu schließen. I2 hingegen übernimmt das Matching von Kundenmodell und Produktmodell. Die Vorteile dieses mehrstufigen Vorgehens wurde bereits von Fridgen et. al herausgearbeitet und gründen vor allem in der Reduktion von Komplexität und der domänenunabhängigen Verwendbarkeit solcher Modelle [FrSc00].

3.1 Kundenmodell

Eine Eigenschaft von Kundenmodellen ist, eine große Menge an Informationen über Kunden bewältigen zu können und verwertbar zu machen. Für WT ist das Vorhandesein eines Kundemodells insbesondere relevant, da die gewonnen Informationen aus den Log-Files schon von Natur aus auf einem niedrigen Abstraktionsniveau sind und für die Individualisierung nur wenig Semantik enthalten. Dabei stellt sich das Problem, das Informationen oft nur implizit vorliegen, als das sie allgemein wiederverwertbar wären. Aus diesem Grund werden Informationen im Kundenmodell aggregiert und auf ein höheres Abstraktionsniveau gebracht. Kundenmodelle sind in der Lage, sowohl quantitative Information (wie zum Beispiel das Einkommen) als auch qualitative Informationen (zum Beispiel der Geschmak) zu kombinieren. Einstellungen (wie beispielsweise die Risikobereitschaft) haben sich dabei als geeignet erweisen, um als abstrakte Repräsentationsform das Wissen über den Kunden zu beschrieben und daraus dessen Präferenzen abzuleiten. Ausführlich wurde der Inhalt ein solchen Modell bereits von [FrVo00] entwickelt.

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

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

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

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

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

K = {Menge aller Kunden}

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

3.2 Produktmodell

Produktmodelle werden benötigt und eingesetzt, damit aus einer großen Menge von Produkten in automatisierten Prozessen individualisierte Produkte erstellt werden können [KuWo01]. Unter Produkten sollen sowohl physiche Produkte als als auch digitale Produkte verstanden werden, d.h. auch eine Information oder ein bestimmter Content einer Web-Site wird als Produkt verstanden. Wie die Einstellungen des Kundenmodells werden bei Produktmodellen, möglichst unabhänige Kategorien definiert, die alle zur Beschreibung der Produkte

relevanten Attribute abdecken sollen. Es existieren verschiedene Ansätze, Produkte zu repräsentieren, wobei keine allgemeingültige Aussage über den richtigen Ansatz zulässig ist. Vielmehr hängt dies von der gegebenen Domäne ab. So sind für die Repräsentation von Content oder von Informationen automatische Klassifikationsmethoden bekannt und erforscht; die am häufigsten verwendete Methode ist dabei der TFIDF-Algorithmus (TermFrequency/ InverseDocument-Frequency) [SaBu88, SaFo83, Pazza01]. Hier wurden in letzter Zeit jedoch auch verstärkt analytische Vorgehensweisen vorgeschlagen [Dubl01]. Für die Modellierung von Produkten im eigentlichen Sinne gibt es bisher noch wenig Erfahrungswerten. Die Autoren sind jedoch der Meinung, dass hier analytische Vorgehensweisen geeigneter sind. Kundisch et.al. haben sich bereits ausführlich mit der Identifikation sowie der optimalen Zusammensetzung von Contentattributen im Finanzdienstleisungssektor beschäftigt [KuWo01].

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

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

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

i e {1, ... , l}, l Anzahl der Produkte

j ∈ {1, ..., m}, m Anzahl der Produkteigenschaften

P = {Menge aller Produkte}

3.3 Webtracking Daten

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

(3) $$C_{i,j} = (K^i, P^j) \in C$$

mit: $K^i \in K$

$P^j \in P$

C = {Menge aller Klicks}

Problematisch bei diesem Verständnis ist allerdings, das heutige Log-Files im Common Logfile Standard diese Informationen (derzeit) so nicht liefern. Es muss folglich sichergestellt sein, dass ein Kunde identifiziert ist, wenn er sich auf der Site befindet, bspw. durch Login oder Cookies, und diese eindeutige

Identifizierung auch im Log-file gespeichert wird, und dass jedes Produkt eindeutig identifzierbar ist, bspw. durch einenen Primärschlüssel, und diese Identifizierung im Log-file gespeichert wird sowie dass über die eindeutige Bezeichnung im Log-file sowohl die Kundenmodellinforamtionen als auch die Produktmodellinformationen abgreifbar sind.

3.4 Inferenzmechanismen

Auf Basis des Kunden- und Produktmodells sowie den WTD kann nun gemäß Fridgen et al. [FrVo00] ein zwei-stufiger Inferrenzmechanismus definiert werden. Die wesentlichen Eigenschaften der Infernezmechanismen sollen an dieser Stelle kurz dargestellt werden.

Inferenzmechanismus 1 leitet aus Informationen die über den Kunden vorliegen, das jeweilige Kundenmodell ab. Dadurch erwächst unmittelbar die Anforderung, aus den vorliegende Kundendaten eine Beziehung zu dessen Einstellungen im Kundenmodell herstellen zu können. I1 schließt von den WTD des Kunden i auf das Kundemodell des Kunden i:

(4) $$I1: F(C) \Rightarrow K$$

Inferenzmechanismus 2 repräsentiert demgegenüber den Beratung- bzw. Verkaufsprozess. Aufbauend auf ein Kundenmodell werden diejenigen Produkte gesucht, von denen angenommen wird, sie stiften dem Kunden maximalen Nutzen.

(5) $$I2: F(K\ ) \Rightarrow P$$

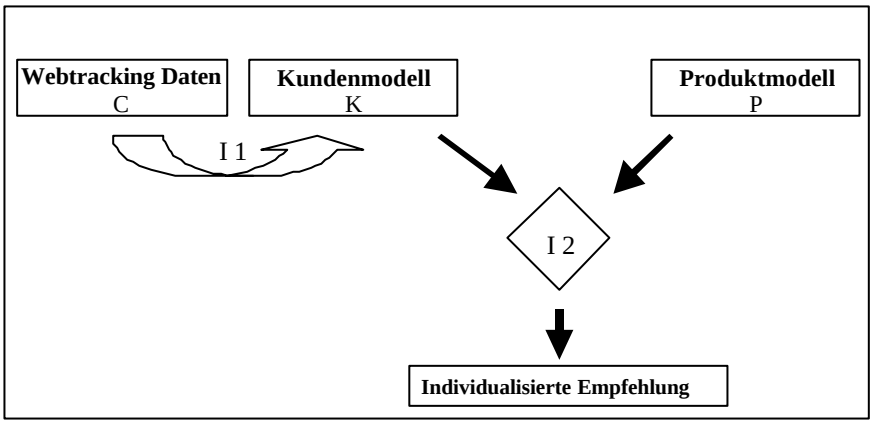

Abbildung 1: Zweistufiger Inferenzmechanismus

In Abschnitt 4 werden nun Methoden besprochen, um in I1 von Webtracking Daten auf das Kundenmodell und vom Kundenmodell auf Produkte zu schließen.

4 Methodik

In diesem Kapitel werden Methoden diskutiert, die im Rahmen des entwickelten Frameworks verwendet werden können, um für den Nutzer in einem zweistufigen Inferenzprozess individualisierte Web-Sites mit dem für ihn passenden Content zu generieren. Als Ergebnis wird im Anschluss eine Gestaltungsempfehlung getrennt nach Ableitung des Kundenmodells und individualisierter Empfehlung erarbeitet.

4.1 State-of-the-Art

Die aufgezeigte Problemstellung – die Selektion bestimmter Objekte aus einer großen Menge potentieller 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 indvidualisierten Empfehlungen für bestimmte Problemdomänen im WWW beschäftigen [ReVa97], z.B. der Internetbuchhändler Amazon (www.amazon.com). Im Folgenden werden ausgewählte Methoden der Recommender Systems und angrenzender Themengebiete diskutiert, die im dargestellten Framework einsetzbar wären.

Die zwei wesentlichen Methoden zur Generierung individueller Empfehlungen sind einerseits das Content based Filtering (CbF) und andererseits das Collaborative Filtering (CF) [Bala97]. Die unterschiedlichen Annahmen und Vorgehensweisen dieser beiden Ansätze und die daraus resultierende Vor- und Nachteilen sollen kurz diskutiert werden.

a) Content based Filtering

CbF generiert Empfehlungen auf Grund der Inhalte des bisher von einem Nutzer betrachteten Contents und des Inhalts des potentiell zu empfehlenden Contents [UnFo00, PaBi97, Bala97]. D.h. dass im ersten Schritt auf Grund der vom User bisher angeklickten und nicht angeklickten Seiten ein Nutzerprofil erstellt wird, 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 **Repräsentationform 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 folgenden Nachteilen von CbF [ClGo01]:

- Die Beschreibung des Contents erfolgt i.d.R. nur inhaltlich, d.h. eine Bewertung der Qualität des Contents fließt in die Vorhersage nicht mit ein.
- CbF hat eine sehr statische Nutzersicht, d.h. dass von einem Nutzer ausgegangen wird, dessen Bewertung von Content sich nur selten und langsam ändert. (Tages-)Aktuelle temporäre Einflüsse können nur schwer berücksichtigt werden.
- Mit steigender Anzahl an potentiell zu empfehlendem Content wird bei einem rein CbF-basierten Ansatz die Menge des in Frage kommenden Contents immer größer und der Ansatz damit immer ineffektiver, da trennschwächer.
- Das CbF berücksichtigt bei der Vorhersage nur die Daten über einen bestimmten Kunden. Die Daten aller anderen Kunden, werden jedoch ignoriert, obwohl diese wertvolle grundsätzliche, abstrakte Zusammenhänge über das Kundenverhalten enthalten können.

b) Collaborative Filtering

CF empfiehlt auf Basis des Verhaltens oder der Empfehlungen anderer im System vorhandener Nutzer [ClGo01, UnFo00, PaBi97]. D.h. dass einem Nutzer derjenige Content empfohlen wird, der von anderen Nutzern bereits angeklickt wurde, die dem Nutzer sehr ähnlich sind. Die zwingend 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

- CF funktioniert nur, wenn ausreichend Nutzer vorhanden sind, zwischen denen signifikante Ähnlichkeiten bestimmt werden können und wenn der Content von ausreichend vielen Nutzern bewertet wurde. Im Umkehrschluss heißt dies, dass bei Content, der neu zur Verfügung steht, oder Nutzern, die noch keine oder nicht ausreichend viele Bewertungen abgegeben haben, CF nur unzureichend oder gar nicht funktioniert.
- Daten, über den bisher von einem Kunden betrachteten Content, werden beim CF überhaupt nicht berücksichtigt.
- Eine Grundannahme von CF ist, dass das, was ähnliche Nutzer tun, auch richtig ist. Es gibt jedoch viele Domänen oder Situationen, in denen dies nicht

der Fall sein muss. Bspw. gibt es Herdentriebe und selbstverstärkende Effekte, die zu irrationalem Verhalten führen.

c) Collaboration via Content (CvC)

Auf Grund 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]. [PaBi97, Pazz01, ClGo01, Bala97] haben empirisch belegt, dass die Kombination zu signifikanten Performanceverbeseerung führt. Grundsätzlich kommt die Kombination 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 individualisierten Empfehlung.

Gemäß des in Kapitel 3 erarbeiteten Frameworks wird nun gezeigt, wie der kombinierte Ansatz zur Ableitung des Kundenmodells und zur Generierung der individualisierten Empfehlung eingesetzt werden kann.

4.2 Erweiterung

Um den Kombinationsansatz aus Kapitel 4.1 auf die Problemstellung anwenden zu können, werden die folgenden Erweiterungen und Einschränkungen vorgenommen bzw. Annahmen getroffen:

1. Es wird keine Einschränkung auf die Empfehlung für Nutzer vorgenommen, sondern es werden Empfehlungen für Kunden generiert, wie sie in Kapitel 3 definiert wurden, und schließt Nutzer damit mit ein.
2. Aus den in Kapitel 3 und bei Fridgen et al. diskutierten Gründen, ziehen wir den Begriff “Kundenmodell” dem in der Recommender Systems Literatur verwendeten Begriff des Nutzerprofils vor.
3. Es wird keine Einschränkung auf Contentempfehlungen vorgenommen, sondern es werden Empfehlungen für Produkte generiert, wie sie in Kapitel 3 definiert wurden, und schließt Content damit mit ein.
4. Im Gegensatz zu den bisherigen Recommender Systems wird im Kontext dieses Papers anstatt des (ordinalen) Ratings eines Kunden für einen bestimmten Content nur die binäre Information verwendet, ob Produkte, die

dem Kunden – als Link oder als Abstract – angezeigt wurde, von dem Kunden angeklickt wurde oder nicht mit der automatischen Bewertung +1 oder -1. Intuitiver und z.T. auch so verwendet wäre das Nicht-Klicken mit 0 zu bewerten. Hierbei würden jedoch Informationen über das Nicht-Klicken und damit über nicht erwünschte Produkte verlorengehen.

In den beiden folgenden Kapiteln soll nun gezeigt werden, wie die zwei in Kapitel 3 definierten Inferenzstufen 1 und 2 mit Hilfe des kombinierten Ansatzes 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. Wie in Kapitel 4.1 beschrieben wird beim CbF das Kundenmodell auf Basis der vom Nutzer bisher bereits betrachteten Produkte erstellt. In der konkreten Problemstellung sind nur Daten darüber vorhanden, welche Produkte der Kunde ausgewählt hat. Somit kann über WTD nur auf die Produkteinstellungen des Kunden geschlossen werden. Folglich ist das Ergebnis des I1 ein optimales kundenindividuelles Produktmodell P*. Zu beachten ist hierbei, dass die nicht bedeutet, dass das Kundenmodell ausschliesslich aus P* besteht. Es wäre bspw. durchaus denkbar, dass K noch anderes Wissen über den Kunden enthält, welches aus anderen Quellen stammt. Es gilt also: $P^* \subset K$. Da dies jedoch nicht Teil der Untersucht ist, wird im Folgenden K durch P* ersetzt.

Die bekanntesten und aus dem Information Retrieval stammenden Methoden sind Rocchio's oder der Winnow Algorithmus. Ein anderer in diesem Zusammenhang bisher kaum verwendeter Ansatz ist die klassische Regression. Diese sollen im folgenden näher erläutert und diskutiert werden.

Rocchios's Algorithmus ist mit die am weitesten verbreitete und angewendete Lernmethode im Information Retrieval [Joac97]. Er wurde bspw. erfolgreich verwendet zur Erzeugung von Kundenmodellen für Nachrichten [Lang95, HaKn88] oder für Web-Sites [Bala97]. Rocchio's Algorithmus hat jedoch den Nachteil, dass die Anzahl der verwendeten Produktkategorien bekannt sein muss. Bei Anwendungsfällen, die diese Bedingung nicht erfüllen, kann der **Winnow Algorithmus** [BlHe95] verwendet werden, der bspw. eingesetzt wurde bei Kundenmodellen für Restaurants [Pazz00].

Die **Regression** ist ein 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,

JuHi85]. [PaBi97] haben jedoch gezeigt, dass nicht-lineare Methoden bei ähnlichen Problemstellungen zu keiner signifikanten Verbesserung führen. Auch bei Regressionen exisitiert die Einschränkung, dass die Anzahl der Produktdimensionen vorher bekannt und konstant sein muss. Eine abschließende Beurteilung dieser Methoden ist jedoch nur empirisch möglich.

4.2.2 Individualisierte Empfehlung

Im Inferenzprozess 2 wird von den Kundenmodellen auf individualisierte Empfehlungen geschlossen, wobei die individualisierte Empfehlung ein optimales Produktmodell als Ergebnis hat, welches mit P** bezeichnet wird. Es werden dem Kunden also die Produkte empfohlen, die P** am nähesten sind. Den Ergebnissen aus Kapitel 4.1 folgend wird P** durch Collaboration über die Kundenmodelle bestimmt. Die Ableitung von P** erfolgt in zwei Stufen. Zunächst wird die Ähnlichkeit oder auch nicht Ähnlichkeit zwischen den einzelnen Kunden festgestellt. Anschließend wird hierauf aufbauend P** berechnet und eine Empfehlung angegeben.

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 Ähnlichkeitsmaße 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 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 Proxmimitätsmaße Q folgende Bedingung zu erfüllen:

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

Distanzbasierter Ansatz

Distanzbasierte Ansätze gehören sicherlich zu den am häufigsten verwendeten Proximitätsmaßen. Zu nennen sind vor allem die euklidsche 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:

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

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

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

(8) $$Q_{i,j} = 1 - \frac{2 \cdot D_{i,j}}{\max\limits_i D_{i,j}}$$

Korrelationkoeffizient

Häufig wird auch der Bravais-Pearson-Korrelationskoeefizient 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 untersucht. Die Berechnung des Korrelationskoeffizienten ist wie folgt:

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

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

b) Individualisierte Empfehlung

Aufgrund des Ähnlichkeitsmaßes Q kann nun für Kunde i die Produktdimension j P_j^{**i} folgendermaßen berechnet werden:

(10) $$P_j^{i**} = (1-x) \cdot \frac{1}{\sum\limits_{\substack{r=1 \\ r \neq i}}^{k} Q_{i,l}} \sum_{\substack{r=1 \\ r \neq i}}^{k} P_j^{*r} \cdot Q_{i,l} + x \cdot P_j^{*i},$$

mit: $Q_{i,l}$ ist die Ähnlichkeit zwischen Kunde i und l

x ∈ [0,1] ist ein Gewichtungsfaktor, der bestimmt wie P* und P** gegeneinander gewichtet werden.

x kann einerseits als statischer Gewichtungsfaktor zwischen den beiden Ansätzen verstanden werden, der empirisch festgelegt werden muss. Andererseits könnte x aber auch situativ-dynamisch bestimmt werden, je nach dem, wie stark sich der Collaborative Ansatz für einen Kunden eignet (vgl. die Diskussion der Nachteile des CF in Kapitel 4.1). Bspw. wäre denkbar, x höher setzen, wenn $Q_{i,j}$ nahe bei Null und damit nicht aussagekräftig ist.

Mit Hilfe der L_q–Metrik (7) wird nun die Distanz D^i_j eines vorhandenen potenziell zu empfehlenden Produkts P^j und P^i_j** P_j^{**i} berechnet. Der Kunde bekommt dann die Produkte empfohlen, die die gerinste Distanz D^i_j besitzen.

4.3 Diskussion

In Kapitel 4 wurde gezeigt, wie der empirisch bewährte Ansatz Collaboration via Content auf die im Framework definierte Problemstellung angewendet werden kann. Es stellt sich hierbei jedoch das Problem, dass die bisher erarbeiteten empirischen Ergebnisse sich nicht notwendiger Weise unbesehen auf die Problemstellung übertragen lassen. Folglich wurden bisher einige Fragen offengelassen bzw. unterschiedliche Alternativen nicht diskutiert, die sich durch eine analytische Vorgehensweise nicht klären lassen, sondern empirische Untersuchungen erfordern, die die speziellen Gegebenheiten des Web-Minings berücksichtigen. Die empirisch noch zu klärenden Punkte sind:

- Ableitung des Kundenmodells: Es muss überprüft werden, welche der vorgeschlagenen Methoden (Regression, Rocchio, Winnow) zur Bestimmung von P* am geeignetsten ist.
- Proximitätsmaß: Ist ein transformiertes lineares Distanzmaß oder ein Ähnlichkeitsmaß zur Bestimmung der Proximität zu verwenden.
- Freie Parameter: Wie sollen die freien Parameter x (Gewichtung von CvC und CbF) und q (Berechnung der L_q-Metrik) eingestellt werden.

5 Bewertung und Ausblick

In dieser Arbeit konnte ein Rahmen skizziert werden, wie bisher bekannte Methoden im Bereich des Information Retrieval und der Recommender Systems zum Web-Mining eingesetzt werden können. Es wurde gezeigt, wie sog. Webtracking Daten semantisch angereichert, aggregiert repräsentiert und für die Individualisierung eingesetzt werden können. Dennoch hat der aufgezeigte Ansatz noch die folgenden Einschränkungen:

- Die heute von Web-Servern erzeugten Standard-Log-files enthalten nicht die in dieser Arbeit geforderten Informationen. Es ist deshalb notwendig, dass entweder die Log-files um diese Informationen erweitert werden oder in einem zweiten Verarbeitungsschritt die noch fehlenden Informationen gematcht werden.
- Die Kategorisierung der Produkte über Produktmodelle ist noch keineswegs üblich, so dass es viele Problemstellungen geben wird, in denen diese Informationen gar nicht oder nur teilweise vorhanden sein werden.

Im Zuge zukünftiger Forschung sollte der hier aufgezeigte Rahmen empirisch überprüft und entsprechend den Ergebnissen erweitert bzw. konkretisiert werden. Da die Datenerhebung auf Grund der schon angesprochenen Problematik der unzureichenden Log-file-Informationen relativ aufwendig ist, wird unsere zukünftige Forschung in zwei Stufen ablaufen. Im ersten Schritt werden im Rahmen einer Monte-Carlo-Simulation Log-file-Daten erzeugt, die zu einer ersten Verfeinerung der Methodik eingesetzt werden. Vorausgesetzt der erste Schritt verläuft erfolgreich, werden im zweiten Schritt Realweltdaten erhoben.

6 Literatur

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

[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, August (1999), Berkeley CA, USA.

[CoWo01] o.V: Studie: Jeder dritte Deutsche ist „drin“. In: Computerwoche, Nr.6, 9. Februar 2001.

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

[Dubl01] Dublin Core Element Set. http://dublincore.org/index.shtml, Abruf am 2001-02-28.

[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), Volume 2, S. 819-825

[FrSt01] Fridgen, Michael; Steck, Werner: Webtracking als Datenquelle für Kundenmodelle im FDL. Diskussionspapier des Lehrstuhls für Betriebswirtschaftslehre mit Schwerpunkt Wirtschaftsinformatik, Universität Augsburg, 2001.

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

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

[HaKn88] Hayes, P.; Knecht, L. et al.: A news story categorization system. In: Second Conference on Applied Natural Language Processing, 1988, S. 9-17.

[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.

[JuHi85] Judge, G.; Hill, C. et al.: The Theory and Practice of Econometric, John Wiley and Sons, New York (NY), USA 1985.

[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.

[PaBi97] Pazzani, Michael; Billsus, Daniel: Learning and Revising User Profiles: The identification of Interesting Web Sites. In: Maching Learning 27 (1997), S. 313-331.

[Pazza01] Pazzani, Michael: A Framework for Collaborative, Content-Based and Demographic Filtering, 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.

[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 imInternet – Individualisierte Angebote mit Collaborativ Filtering, Dissertation an der Universität Kiel, Wiesbaden 2000.

[SaBu88] Salton, Gerard; Buckley, C.: Term Weighting Approaches in Automatic Text Tretrieval. In: Information Processing and Management, 24 (1988), S. 513-523.

[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):200-210, 1985-05-00.

[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.

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

[W3B] o.V: 5. W3B Umfrage, 1997, http://cf.nci.de/dmmv_studien/studien_update_oeff.cfm?Aus=2, Abruf am 2001-02-28.