

# Personalisierung mittels informationsbewertender Profile\*

*Tobias Galliat*

Institut für Informationswissenschaft,  
Fachhochschule Köln,  
Claudiusstr. 1, 50678 Köln  
*tobias.galliat@fb-koeln.de*

## **Zusammenfassung**

Die automatisierte Personalisierung von Web-Angeboten erfordert eine schnelle Profilbildung. Der Open Profiling Standard versucht das Kaltstart-Problem zu vermeiden, indem Nutzer-Profile normiert werden, so dass sie zwischen unterschiedlichen Profilingssystemen übertragen werden können. In Fällen wo dies nicht praktikabel ist, müssen zur anfänglichen Profilbildung auch Informationen niedrigen Qualitätsgrades – beispielsweise Prognosen – verwendet werden. Dies setzt jedoch ein Profilmodell voraus, bei dem der Informationsgrad geeignet bewertet wird. Hierzu wird ein informations-theoretisch motiviertes Konzept vorgestellt, das zudem eine beschleunigte Profilbildung durch Wechselwirkung zwischen Profilen vorsieht.

## **I Einleitung**

Die Attraktivität des Internets hängt offensichtlich eng damit zusammen, dass es prinzipiell für alle Bedürfnisse von Menschen in Bezug auf den Erwerb und den Austausch von Information, Waren und Dienstleistungen entsprechende Angebote bereithält. Allerdings ist in der Regel ein erheblicher, nicht immer von Erfolg gekrönter Suchaufwand erforderlich, bevor der Internet-Nutzer auf die für ihn relevanten Angebote und Inhalte zugreifen kann. Ob und in welchem Umfang die Suche erfolgreich ist, hängt dabei stark von seinen individuellen Fähigkeiten im Umgang mit dem Internet sowie der Zeit ab, die er bereit ist in die Suche zu investieren.

---

\* Veröffentlicht in: OSSWALD, Achim; STEMPFHUBER, Maximilian; WOLFF, Christian (2007). Open Innovation. Proc. 10. Internationales Symposium für Informationswissenschaft. Konstanz: UVK, 27-41.

Es ist deshalb wenig verwunderlich, dass neben der Bereitstellung von effizienten Suchmaschinen die Personalisierung von Internet-Angeboten mittels Profilen – zum Beispiel via Portal-Seiten – als Ausweg aus diesem Dilemma angesehen wird. Zu unterscheiden ist dabei zwischen einer manuellen und einer automatisierten Personalisierung<sup>1</sup>. Die manuelle Personalisierung, die das Festlegen von bevorzugten Web-Seiten, aber auch die detaillierte Auswahl von Angeboten auf einer Portal-Seite umfassen kann, nimmt der Nutzer bewusst und unmittelbar selber vor. Dies setzt voraus, dass er seine Bedürfnisse genau kennt und den entsprechenden Angeboten zuordnen kann. Offensichtlich verkürzt die manuelle Personalisierung hauptsächlich die Zeit der Suche nach im Wesentlichen bekannten Angeboten und bereitet diese in strukturierter Form auf. Für den weniger erfahrenen Nutzer bietet sie hingegen keine wirkliche Lösung seines Suchproblems. Er ist sich seiner Bedürfnisse teilweise gar nicht bewusst bzw. kann diesen keine oder nur wenige relevante Angebote unmittelbar zuordnen. Bei der automatisierten Personalisierung mittels Profilen ist dies auch nicht notwendig, denn hier wirkt der Nutzer nur mittelbar, explizit oder implizit mit, indem er beispielsweise durch Angabe seines Geburtsdatums, die Bewertung von Angeboten oder sein aufgezeichnetes Klickverhalten, die Ausprägungen vorgegebener Attribute – wie z. B. das Alter, die Interessen oder die mittlere Verweildauer im Internet – festlegt. Hieraus lässt sich ein Nutzer-Profil ableiten, dem dann automatisch Angebote und Inhalte zugeordnet werden, die in der Vergangenheit für Nutzer mit ähnlichen Profilen – die über das so genannte „Profil-Matching“ ermittelt werden – relevant gewesen sind. Dieser Vorgang wird in der Literatur üblicherweise als „kollaboratives Filtern“ [Breese et al. 98], die entsprechende Personalisierungs-Software als „Recommender Systeme“ [Terveen et al. 01] bezeichnet<sup>2</sup>.

Nun muss es das Ziel jedes Anbieters von personalisierten Web-Services sein, das so genannte Kaltstartproblem zu lösen, d. h. eine möglichst schnelle Profilbildung bei jedem neuen Nutzer – optimalerweise nach dem ersten Klick – zu erreichen und danach die Profile der wiederkehrenden Nutzer – durch das so genannte „Profil-Influencing“ – immer weiter zu verfeinern und aktuell zu halten. Mit dem „Open Profiling Standard (OPS)“ hat es schon frühzeitig Versuche gegeben eine Normierung der Profile zu schaffen [Dunn et al. 97], so dass Web-Anbieter ein bereits be-

---

<sup>1</sup> In der Literatur zum Thema „Benutzermodellierung“ werden alternativ oft die – im Wesentlichen jeweils inhaltlich deckungsgleichen – Begriffe „(benutzer)-adaptierbar“ und „(benutzer)-adaptiv“ verwendet [Kobsa 04].

<sup>2</sup> Es gibt allerdings auch Recommender Systeme, die mit profilunabhängigen Filterverfahren, wie z. B. eigenschaftsbasierten Filtern arbeiten [van Meteren et al. 00].

stehendes Profil – bei entsprechender Freigabe durch den Nutzer – zur Personalisierung ihres Angebotes verwenden können, ohne dass der Nutzer die Anbieter-Seite bereits früher besucht bzw. dort einen umfangreichen Prozess der Profilbildung absolviert hat. In der Praxis hat der OPS-Ansatz mit diversen Schwierigkeiten zu kämpfen, wie beispielsweise der zunehmenden Sensibilität der Nutzer bezüglich der Weitergabe ihrer persönlichen Daten, der in der Normierung begründet liegenden Beschränktheit der Profilbildung oder der Aktualität der Profile. Damit bietet der OPS-Ansatz in vielen Fällen keine wirklich befriedigende Lösung für das Kaltstartproblem von Profilingssystemen.

Im Folgenden soll deshalb das Konzept einer automatisierten Personalisierung mittels informationsbewertender Profile vorgestellt werden, bei dem die Profilbildung dadurch beschleunigt wird, dass zum einen nicht nur den Nutzern, sondern auch den Informations- oder Warenangeboten – den so genannten „Items“ – Profile zugeordnet werden, die sich wechselseitig beeinflussen können<sup>3</sup>. Dadurch wird die Trennung zwischen der Profilbildung und dem Prozess des kollaborativen Filterns aufgehoben und im Gegensatz zu vielen herkömmlichen Ansätzen, besteht die Möglichkeit neue Angebote sofort in das System zu integrieren<sup>4</sup>. Zum anderen sieht das Konzept vor, dass jede zur Profilbildung verwendete Attribut-Ausprägung mit einem so genannten „Informationsgrad“ versehen wird, der bewertet wie zuverlässig die Festlegung der Ausprägung ist<sup>5</sup>. Um die Interpretation zu erleichtern, bietet es sich an, den Wertebereich für den Informationsgrad auf das Intervall  $[0, 1]$  festzusetzen. Eine Attribut-Ausprägung erhält dann einen Informationsgrad nahe bei 0, wenn die Aussage, dass die Attribut-Ausprägung dem Profil zugeordnet werden kann, keinen oder nur eine sehr geringe Validität besitzt. Ein Wert nahe bei 1 ist hingegen als eine (fast) sichere Information zu interpretieren. Der Informationsgrad ist dabei nicht unveränderlich über die Zeit. Beispielsweise lässt sich eine Adressangabe durch den erfolgreichen Versand eines bestellten Produktes verifizieren oder das Nutzer-Interesse an einem bestimmten Angebot durch dessen regelmäßige Besuche der entsprechenden Web-Seiten bestätigen. Damit wird es möglich, instantan Profile auf der Basis von Informationen niedrigen Qualitätsgrades – beispielsweise Hypothesen – zu bilden und sie dann im Laufe der Zeit zu verfeinern und zu validieren.

---

<sup>3</sup> In der neueren Literatur um Thema „Collaborative Filtering“ finden sich zunehmend Ansätze zur Verbindung von Nutzer- und Item-Profilen [Wang et al. 06].

<sup>4</sup> Es existieren auch Verfahren, die dies auf anderem Wege erreichen [Schein et al. 02].

<sup>5</sup> In anderen Ansätze wird die Nutzer-Zuverlässigkeit bewertet [Massa et al. 04].

Basis des entwickelten Konzeptes ist ein Modell für Profilingssysteme, bei dem – in Analogie zum aus dem Information Retrieval bekannten Vektorraummodell – Profile durch Objekte in einem mehrdimensionalen Raum – im Folgenden als Profil-Raum bezeichnet – repräsentiert werden, wobei jedem Attribut mit seiner endlichen Anzahl von zulässigen Ausprägungen jeweils eine Dimension des Raumes zugeordnet ist<sup>6</sup>. Neu ist hier die Modellierung des Informationsgrades als so genannte „charakteristische Dimension“. In diesem Kontext lässt sich dann das Profil-Influencing als die regelbasierte Veränderung der Objekte des Profil-Raumes über die Zeit und durch Wechselwirkung untereinander interpretieren. Das Profil-Matching wiederum lässt sich als die Bestimmung von im Profil-Raum nahe beieinander liegenden Objekten auffassen. Dabei wird der Grad der „Nähe“ mittels eines Distanzmaßes bestimmt, das den Informationsgrad sinnvoll berücksichtigt und auch der Tatsache gerecht wird, dass die Mehrzahl der Objekte im Profilraum für viele Dimensionen keine echte Ausprägung besitzen bzw. vielen Ausprägungen ein Informationsgrad nahe bei 0 zugeordnet wird. Diese Vorgaben führen zur Verwendung einer bereits früher entwickelten informationsbewertenden Distanz [Galliat 99] im Zusammenspiel mit einer Adaption des Ähnlichkeits-Maßes von Tanimoto.

Die Präsentation des Konzeptes erfolgt in drei Schritten: In einem ersten Schritt wird durch Formulierung der gewünschten Eigenschaften eines Profils ein geeigneter Profil-Raum konstruiert, dann werden die Regeln für das Profil-Influencing beschrieben und im letzten Schritt ein zum Profil-Matching geeignetes informationsbewertendes Distanzmaß vorgestellt. Abschließend erfolgt noch ein Ausblick auf mögliche Erweiterungen des Konzeptes im Hinblick auf eine praktische Realisierung.

## 2 Konstruktion des mehrdimensionalen Profil-Raumes

Bei der Konstruktion des Profil-Raumes, sind folgende gewünschte Eigenschaften eines Profils zu berücksichtigen:

- Die Ausprägungen der Attribute besitzen nicht zwingend eine natürliche Ordnung, d. h. die Dimensionen des Profil-Raumes sind im Allgemeinen weder ordinal noch metrisch skaliert.

---

<sup>6</sup> Vorausgesetzt wird dabei zunächst, dass es zu jedem Attribut eine „Missing“- Ausprägung gibt, d. h. einen Wert der gesetzt wird, wenn keine der echten Ausprägungen zutrifft. Durch die Berücksichtigung des Informationsgrades wird diese künstliche Ausprägung überflüssig.

- Ein Profil kann mehrere Ausprägungen eines Attributes besitzen<sup>7</sup>, wird also nicht durch einen Punkt, sondern durch ein komplexeres Objekt im Profil-Raum repräsentiert.
- Damit einem Profil eine Attribut-Ausprägung zugeordnet werden kann, ist ein rechtfertigendes Ereignis – der so genannte „Hit“ – notwendig. Je nach dem ob dieses Ereignis beobachtet<sup>8</sup>, aufgrund einer gezielten Befragung ermittelt<sup>9</sup> oder durch eine Prognose – oft auf der Basis von soziodemographischen Daten – ausgelöst wird<sup>10</sup>, spricht man von einem Beobachtungsereignis, einem Befragungsereignis oder einem Prognoseereignis. Prognoseereignisse sind dabei in der Regel die Folge von Beobachtungs- oder Befragungsereignissen. Deshalb bestehen hier auch oft Abhängigkeiten zwischen den zugehörigen Attribut-Ausprägungen.
- Die Aussage, dass einem Profil eine bestimmte Attribut-Ausprägung zugeordnet werden kann, ist oft mit Unsicherheit behaftet. Deshalb soll für jede Attribut-Ausprägung die einem Profil zugeordnet wird, die Zuverlässigkeit der Information bewertet und als Informationsgrad vermerkt sein, d. h. jeder Attribut-Ausprägung wird eine zusätzliche charakteristische Dimension zugeordnet, die den Informationsgrad widerspiegelt.
- Die Attribut-Ausprägungen und die ihnen zugeordneten Informationsgrade können sich im Laufe der Zeit ändern<sup>11</sup>. Die Objekte im Profil-Raum sind also nicht statisch, sondern dynamisch. Aus diesem Grund ist es zweckmäßig den Attribut-Ausprägungen weitere charakteristische Dimensionen zuzuordnen. So ist es sinnvoll die Kategorie des rechtfertigenden Ereignisses als „HitCategory“, den Zeitpunkt des letzten rechtfertigenden Ereignisses als „HitTime“ und die durchschnittliche Dauer zwischen zwei rechtfertigenden Ereignissen als „AverageHitToHitTime“ zu vermerken<sup>12</sup>.

---

<sup>7</sup> Ein Nutzer interessiert sich für Reisen nach Frankreich und Schweden. Seinem Profil müssen also für das Attribut „bevorzugte Reiseziele“ mindestens zwei gültige Ausprägungen zugeordnet werden.

<sup>8</sup> Z. B. durch das Anklicken eines Werbebanners.

<sup>9</sup> Z. B. im Rahmen des einmaligen Anmeldeprozesses für einen Web-Service

<sup>10</sup> Beispielsweise lässt sich die Ausprägung für das Attribut „Kreditwürdigkeit“ aus einem Abgleich von Personendaten mit der Datenbank einer Kredit-Auskunftei ermitteln.

<sup>11</sup> Eltern interessieren sich z. B. in der Regel nur für einen gewissen Zeitraum für Babysachen. Sobald die Kinder älter sind, werden entsprechende Attribut-Ausprägungen für ihre Profile keine oder nur noch eine sehr geringe Gültigkeit besitzen.

<sup>12</sup> Üblicherweise gilt: je häufiger ein rechtfertigendes Ereignis beobachtet wird, umso zuverlässiger ist die Zuordnung zu einem Profil. Allerdings sollte berücksichtigt werden, wie häufig ein derartiges rechtfertigendes Ereignis überhaupt zu erwarten ist.

Aus den aufgeführten Eigenschaften lässt sich folgendes Modell für den Profil-Raum ableiten:

Sei  $A_1, \dots, A_k$  die Menge der Attribute, die zur Profilbildung verwendet werden sollen und für jedes Attribut  $A_i$  bezeichne  $M_i := M(A_i) := \{a_{i1}, \dots, a_{in(i)}\}$  die Menge der zulässigen Ausprägungen. Diese Attribute bilden die Attribut-Dimensionen des Profil-Raumes und die Mengen  $M_i$  die zugehörigen Wertebereiche. Zusätzlich existieren für jede Attribut-Ausprägung  $a_{ij}$  charakteristische Dimensionen für den Informationsgrad, die Kategorie und den Zeitpunkt des letzten rechtfertigenden Ereignisses sowie die durchschnittliche Dauer zwischen zwei rechtfertigenden Ereignissen.

Besitzt ein Profil mehr als eine Ausprägung für ein Merkmal lässt es sich offensichtlich nicht als Punkt im Profil-Raum interpretieren. Wir modellieren deshalb ein Profil als komplexes Objekt des Profil-Raumes, dem für jede (Attribut-)Dimension eine Menge von Ausprägungen zugeordnet werden können.

Wie nun diese Objekte – ausgelöst durch rechtfertigende Ereignisse – konkret gebildet werden und sich über die Zeit verändern bzw. sich wechselseitig beeinflussen können, wird im nächsten Schritt erläutert.

### 3 Regeln für das Profil-Influencing

Beim Profil-Influencing lassen sich die folgenden Situationen unterscheiden:

- Erstmalige Zuordnung von Attribut-Ausprägungen zu einem Profil, sowie die Festlegung der zugehörigen charakteristischen Werte, aufgrund eines rechtfertigenden Ereignisses.
- Änderung der charakteristischen Werte von bereits zugeordneten Attribut-Ausprägungen durch das Ausbleiben bzw. das Eintreten von weiteren rechtfertigenden Ereignissen.
- Wechselwirkung zwischen Profilen, die zur Zuordnung von Attribut-Ausprägungen und Änderung von charakteristischen Werten führt.

#### 3.1 Erstmalige Zuordnung von Attribut-Ausprägungen

Betrachten wir zunächst die Situation einer erstmaligen Zuordnung von Attribut-Ausprägungen.

Zu einem Zeitpunkt  $T$  erfolge ein rechtfertigendes Ereignis für die Zuordnung der o.b.d.A eindeutigen Ausprägung  $x$  für das Attribut  $A$  zum Profil  $P$ . Sei  $x^*(P)$  der Informationsgrad mit dem  $x$  dem Profil  $P$  zugeordnet werden kann. Mit  $HC(x,P)$ ,

$HT(x,P)$  und  $\emptyset H_2 HT(x,P)$  seien ferner die HitCategory, die HitTime und die AverageHitToHitTime von  $x$  bezüglich des Profils  $P$  bezeichnet.

Zusätzlich sei  $GHT(x)$  bzw.  $G\emptyset H_2 HT(x)$  als globale durchschnittliche HitTime bzw. globale durchschnittliche AverageHitToHitTime der Ausprägung  $x$  über alle bereits im System befindlichen Profile definiert<sup>13</sup>:

$$(1) \quad GHT(x) := \bigcap_{HT(x,P) \text{ definiert}} HT(x,P) \text{ bzw.}$$

$$(2) \quad G\emptyset H_2 HT(x) := \bigcap_{\emptyset H_2 HT(x,P) \text{ definiert}} \emptyset H_2 HT(x,P) .$$

Als erstes wird die Kategorie  $HC(x,P)$  des rechtfertigenden Ereignisses für  $x$  bezüglich  $P$  vermerkt. Handelt es sich hierbei um ein Beobachtungs- oder Befragungsergebnis, setzt man als nächstes den Informationsgrad auf einen a priori festgelegten Wert. Wie die späteren Ausführungen zur Erhöhung des Informationsgrades bei der Beobachtung weiterer rechtfertigender Ereignisse zeigen werden, bietet sich hier der Wert 0,5 an. Durch Auswertung externer Quellen bzw. die Anwendung von Validierungsregeln<sup>14</sup> kann der Informationsgrad ggf. noch herauf oder herabgesetzt werden. Als HitTime  $HT(x,P)$ , wird der aktuelle Zeitpunkt  $T$  vermerkt. Für  $\emptyset H_2 HT(x,P)$  setzt man den Wert  $G\emptyset H_2 HT(x)$  ein<sup>15</sup>. Handelt es sich hingegen um ein Prognoseereignis, geht man anders vor. Derartige Ereignisse sind immer mit früher stattgefundenen Ereignissen verknüpft, d. h. die Zuordnung einer Ausprägung  $x$  zum Profil  $P$  erfolgt aufgrund anderer Ausprägungen, die  $P$  bereits früher zugeordnet worden sind. Bezeichne deshalb im Folgenden  $D(x,P)$  die Menge der Attribut-Ausprägungen von denen  $x$  abhängt<sup>16</sup>.

Die charakteristischen Werte von  $x$  bezüglich  $P$  sind natürlich ebenfalls abhängig von den entsprechenden Werten der Ausprägungen in  $D(x,P)$ . In einem konservativen Ansatz wählen wir jeweils den minimalen Wert:

$$(3) \quad x^*(P) := \min_{\tilde{x} \in D(x,P)} \tilde{x}^*(P) ,$$

$$(4) \quad HT(x,P) := \min_{\tilde{x} \in D(x,P)} HT(\tilde{x}, P) ,$$

$$(5) \quad \emptyset H_2 HT(x,P) := \min_{\tilde{x} \in D(x,P)} \emptyset H_2 HT(\tilde{x}, P) .$$

<sup>13</sup> Das  $\cap$ -Symbol ist hier wie folgt zu interpretieren:

$$\bigcap_{\lambda \in I} f(\lambda) := \sum_{i=1}^r \frac{1}{r} f(\lambda_i) \quad \text{mit } I = \{\lambda_1, \dots, \lambda_r\}.$$

<sup>14</sup> Z. B. empirischer Zusammenhang zwischen Vornamen und Geburtsjahrgang.

<sup>15</sup> Der Wert von  $G\emptyset H_2 HT(x_i)$  darf anfangs undefiniert bzw. eine Konstante sein.

<sup>16</sup> Beispielsweise kann die Ausprägung für das Attribut „Kaufkraft“ von den Ausprägungen der Attribute „Wohnort“ und „Alter“ abhängen.

Als nächstes stellt sich nun die Frage, welche Auswirkungen das Auftreten bzw. das Ausbleiben von rechtfertigenden Folgeereignissen für  $x$  bezüglich  $P$  auf die zugehörigen charakteristischen Werte haben.

### 3.2 Regelmäßige Aktualisierung von charakteristischen Werten

Beobachtete Ereignisse die eine Aufnahme von  $x$  ins Profil  $P$  rechtfertigen sind im Allgemeinen nicht singular, sondern können im Laufe der Zeit immer wieder auftreten. Vielmehr sollten sie sogar regelmäßig auftreten, denn sonst muss von einer abnehmenden Gültigkeit der Attribut-Ausprägung für das betrachtete Profil ausgegangen werden. Die Häufigkeit des Auftretens in einem bestimmten Zeitraum sollte dabei nicht absolut bewertet werden<sup>17</sup>, sondern relativ zu der in der Vergangenheit beobachteten Frequenz und der Häufigkeit, mit der entsprechende Beobachtungen bei anderen Profilen erfolgen. Der Informationsgrad einer Attribut-Ausprägung für ein Profil hängt also in geeigneter Weise vom Zeitpunkt des letzten rechtfertigenden Ereignisses sowie der in der Vergangenheit beobachteten durchschnittlichen Zeit zwischen zwei rechtfertigenden Ereignissen ab, und zwar sowohl bezogen auf das gegebene Profil als auch über alle Profile hinweg.

Befragungsereignisse sind hingegen tendenziell singularer Natur<sup>18</sup>. Hier bleibt in der Folge der Informationsgrad über die Zeit konstant, es sei denn es gibt konkrete Hinweise, die eine Änderung begründen.

Anders sieht die Situation bei Attribut-Ausprägungen  $x$  aus, die aufgrund eines Prognoseereignisses dem Profil  $P$  zugeordnet worden sind. Solange sich die Kategorie von  $x$  nicht ändert, müssen die charakteristischen Werte bei jeder regelmäßigen Aktualisierung von Attribut-Ausprägungen aus der Menge  $D(x,P)$  entsprechend den Regeln für die erstmalige Zuordnung<sup>19</sup> ebenfalls aktualisiert werden.

Im Folgenden sei nun  $x$  eine Attribut-Ausprägung für die bezüglich des Profils  $P$  letztmalig zum Zeitpunkt  $HT(x,P)$  ein rechtfertigendes Ereignis beobachtet worden ist. Ferner sei  $T$  der aktuelle Beobachtungszeitpunkt.

---

<sup>17</sup> Nutzer A besucht die Web-Seiten eines E-Commerce-Anbieters täglich, Nutzer B jedoch nur einmal wöchentlich. Die Häufigkeit von rechtfertigenden Ereignissen wird somit bei Nutzer A größer sein als bei Nutzer B. Daraus lässt sich aber keine entsprechende Aussage über die Gültigkeit der Attribut-Ausprägungen treffen, die den Profilen zugeordnet sind.

<sup>18</sup> Die wiederholte Befragung führt bei Kunden oft zu Irritationen und wird von diesen in der Regel abgelehnt. Sollte dennoch erfolgreich eine erneute Befragung durchgeführt worden sein, so werden einfach die alten durch die neuen Attribut-Ausprägungen ersetzt.

<sup>19</sup> Siehe Formeln (3) – (5).



### 3.2.1 Kein rechtfertigendes Ereignis zum Zeitpunkt $T$

Wir betrachten zunächst den Fall, dass zum Zeitpunkt  $T$  kein rechtfertigendes Ereignis für  $x$  bezüglich  $P$  beobachtet wird. Wenn nun der Zeitraum zwischen  $HT(x,P)$  und  $T$  größer ist als die AverageHitToHitTime von  $x$  bezüglich  $P$ , d. h. wenn gilt:  $\emptyset H_2 HT(x,P) \leq T - HT(x,P)$ , dann reduzieren wir den Informationsgrad („regelmäßiger Downgrade“):

$$(6) \quad x^*(P) := x^*(P) - (0,5 / (2t - \emptyset H_2 HT(x,P))) x^*(P).$$

Dabei ist  $t$  das Maximum aus  $\emptyset H_2 HT(x,P)$  und dem halben globalen durchschnittlichen Zeitraum zwischen zwei rechtfertigenden Ereignissen für  $x$ , d. h.  $t := \max(\emptyset H_2 HT(x,P), G\emptyset H_2 HT(x) / 2)$ .

Außer für Profile, bei denen sehr selten rechtfertigende Ereignisse beobachtet werden, gilt  $t = \emptyset H_2 HT(x,P)$ , d. h. der Informationsgrad wird um den Faktor  $0,5/\emptyset H_2 HT(x,P)$  reduziert<sup>20</sup>. Je regelmäßiger also in der Vergangenheit ein rechtfertigendes Ereignis für  $x$  bezüglich  $P$  zu beobachten war, umso stärker ist der Downgrade, wenn es im weiteren Verlauf ausbleibt.

Sinkt der Informationsgrad von  $x$  bezüglich  $P$  nach dem Downgrade auf einen Wert nahe bei 0, d. h. gilt  $x^*(P) < \varepsilon$  für sehr kleines  $\varepsilon > 0$ , wird die Ausprägung automatisch aus dem Profil entfernt und  $HT(x,P)$ , sowie  $\emptyset H_2 HT(x,P)$  als undefiniert gekennzeichnet.

Wenn der nächste erwartete Zeitpunkt für ein rechtfertigendes Ereignis hingegen noch nicht erreicht ist, wenn also gilt  $\emptyset H_2 HT(x,P) > T - HT(x,P)$ , dann bleibt  $x^*(P)$  unverändert<sup>21</sup>.

Die charakteristischen Werte  $HT(x,P)$  und  $\emptyset H_2 HT(x,P)$  ändern sich nicht.

### 3.2.2 Rechtfertigendes Ereignis zum Zeitpunkt $T$

Betrachten wir nun die Situation, dass zum Zeitpunkt  $T$  ein rechtfertigendes Ereignis für  $x$  bezüglich  $P$  beobachtet wurde. In diesem Fall setzen wir:

$$(7) \quad x^*(P) := x^*(P) + (1 - x^*(P)) / 2.$$

<sup>20</sup> Sinnvollerweise arbeitet man mit einem diskreten Zeitgeber, d. h.  $T$  wird immer um eine Zeiteinheit erhöht. Es gilt also  $\emptyset H_2 HT(x,P) \geq 1$  und damit nach dem regelmäßigen Downgrade auch immer  $x^*(P) \geq 0$ .

<sup>21</sup> Dies ist insbesondere auch dann der Fall, wenn  $\emptyset H_2 HT(x,P)$  noch undefiniert ist.

Der Informationsgrad wird also um den Wert  $(1-x \cdot (P))/2$  erhöht<sup>22</sup>. Zusätzlich müssen auch noch die Werte  $HT(x,P)$  und  $\emptyset H_2 HT(x,P)$  angepasst werden.

Falls  $\emptyset H_2 HT(x,P)$  undefiniert ist, setzen wir

$$(8) \quad \emptyset H_2 HT(x,P) := T - HT(x,P),$$

andernfalls wird ein neuer Durchschnittswert berechnet<sup>23</sup>:

$$(9) \quad \emptyset H_2 HT(x,P) := (\emptyset H_2 HT(x,P) + T - HT(x,P)) / 2.$$

Im Anschluss aktualisieren wir dann noch die HitTime auf den aktuellen Zeitpunkt, d. h.  $HT(x,P) := T$ .

### 3.3 Wechselwirkung zwischen Profilen

Der Aufbau von Profilen durch die bisher beschriebenen Regeln ist sehr zeitaufwendig, da sehr viele rechtfertigende Ereignisse benötigt werden. Durch den Einsatz von Prognosen, lässt sich dieser Prozess schon deutlich beschleunigen. Allerdings ist die Formulierung von Prognosen nicht trivial, sondern es müssen geeigneten Datenquellen zur Verfügung stehen und anspruchsvolle Techniken der statistischen Datenanalyse und des Data-Mining, wie z. B. Cluster-Verfahren [Ungar et al. 98] genutzt werden. Des Weiteren sind Attribut-Ausprägungen, die aufgrund von Prognosen einem Profil zugeordnet werden, auch immer nur so valide, wie es die der Prognose zugrunde liegenden Attribut-Ausprägungen sind. Können einem Profil am Anfang also nur wenige durch Beobachtung oder Befragung ermittelte Attribut-Ausprägungen zugeordnet werden – wie dies üblicherweise oft der Fall ist – so lassen sich auch nur schwer Prognosen mit hohem Informationsgrad abgeben.

Einen Ausweg aus diesem Dilemma bietet die Wechselwirkung zwischen Profilen. Hierzu ist es erforderlich nicht nur Nutzer-, sondern auch Item-Profile anzulegen, die sich dann wechselseitig beeinflussen können<sup>24</sup>. Dadurch erfolgt ein schneller Aufbau von Profilen sowohl für neue Nutzer, als auch für neue Items. Bei der Mo-

---

<sup>22</sup> Offensichtlich wird so sichergestellt, dass stets gilt:  $x \cdot (P) \leq 1$ . Für den Fall, dass  $x$  dem Profil  $P$  zuvor nicht zugeordnet werden konnte, also  $x \cdot (P) = 0$  gilt, erhält man nach dem rechtfertigenden Ereignis den Wert  $x \cdot (P) = 0,5$ . Dies ist konsistent mit der vorgeschlagenen Bewertung bei erstmaliger Beobachtung eines rechtfertigenden Ereignisses für  $x$  bezüglich  $P$ .

<sup>23</sup> Es handelt sich um keine echte Durchschnittsbildung, sondern der letzte Zeitraum wird immer stark übergewichtet. Damit wird eine schnellere Änderung der charakteristischen Werte ermöglicht und auch die Datenhaltung vereinfacht.

<sup>24</sup> Web-Links können z. B. mit Profilen versehen werden. Wenn nun der Nutzer auf einen solchen Link klickt, beeinflussen sich sein Profil und das Link-Profil wechselseitig.

dellierung gibt es keinen Unterschied zwischen einem Nutzer- und einem Item-Profil, beides sind Objekte im Profil-Raum. Wir formulieren deshalb als nächstes Regeln für die Wechselwirkung zwischen Objekten P und Q des Profil-Raumes. Dabei betrachten wir jede Attribut-Dimension separat.

Profil P besitze für Attribut A die Ausprägungen  $x_1, \dots, x_s$  und Profil Q die Ausprägungen  $y_1, \dots, y_r$  mit jeweils positivem Informationsgrad. Alle Attribut-Ausprägungen in Q die nicht in P liegen, werden auch in P eingefügt, allerdings mit Informationsgrad 0. Entsprechend verfährt man mit Attribut-Ausprägungen in P die nicht in Q liegen. Anschließend sortiert man die Ausprägungen so um, dass  $x_j = y_j$  für  $j = 1, \dots, n$ , wobei n die Anzahl der gemeinsamen Attribut-Ausprägungen in Profil P und Q bezeichne.

Die wechselseitige Beeinflussung von P und Q erfolgt nun über eine Veränderung des Informationsgrades. Dies ist gerechtfertigt, da es sich bei der Wechselwirkung ja offensichtlich um die Folge eines beobachteten Ereignisses handelt. Für  $j = 1, \dots, n$  setzt man:

$$(10) \quad x_j^*(P) := x_j^*(P) + \Delta(y_j^*(Q), x_j^*(P))$$

$$\text{mit } \Delta(y_j^*(Q), x_j^*(P)) := (y_j^*(Q) - x_j^*(P)) \left( (1 - d_j^*(P, Q)) + (1 - x_j^*(P)) \right) / 2,$$

$$(11) \quad y_j^*(Q) := y_j^*(Q) + \Delta(x_j^*(P), y_j^*(Q))$$

$$\text{mit } \Delta(x_j^*(P), y_j^*(Q)) := (x_j^*(P) - y_j^*(Q)) \left( (1 - d_j^*(Q, P)) + (1 - y_j^*(Q)) \right) / 2.$$

Dabei bezeichnet  $d_j^*(P, Q)$  die informationsbewertende Distanz [Galliat 99] zwischen den Profilen P und Q für die Ausprägung  $x_j (= y_j)$  entlang der betrachteten (Attribut-)Dimension. Wir definieren:

$$(12) \quad d_j^*(P, Q) := \log_2 \left( 2 - (x_j^*(P) y_j^*(Q))^{1/2} \right)$$

Offensichtlich gilt  $d_j^*(P, Q) \in [0, 1]$  sowie  $d_j^*(P, Q) = 0$  gdw.  $x_j^*(P) y_j^*(Q) = 1$  und  $d_j^*(P, Q) = 1$  gdw.  $x_j^*(P) y_j^*(Q) = 0$ .

Man verifiziert leicht, dass keine Änderung des Informationsgrades von  $x_j$  bezüglich P erfolgt, wenn  $y_j^*(Q) = x_j^*(P)$  oder  $(x_j^*(P) = 1 \text{ und } y_j^*(Q) = 0)$ . Ansonsten erfolgt eine Änderung in Abhängigkeit von der Distanz  $d_j^*(P, Q)$  und der Differenz  $y_j^*(Q) - x_j^*(P)$ . Dabei fällt die Änderung umso stärker aus, je kleiner  $x_j^*(P)$ . Wenn  $x_j^*(P) < y_j^*(Q)$ , dann wird der Informationsgrad von  $x_j$  bezüglich P erhöht, ansonsten reduziert. Analoge Aussagen gelten für  $y_j^*(Q)$ .

Auch die anderen charakteristischen Werte von  $x_j$  und  $y_j$  werden beeinflusst. Sei  $T$  der aktuelle Zeitpunkt der Wechselwirkung. Für  $x_j$  ergeben sich folgende Änderungen (entsprechend für  $y_j$ ):

Fall 1:  $\emptyset H_2 HT(x_j, P)$  und  $HT(x_j, P)$  undefiniert, d. h.  $x_j$  ist nicht im Profil  $P$

Setze  $\emptyset H_2 HT(x_j, P) := \emptyset H_2 HT(y_j, Q)$  und  $HT(x_j, P) := T$ .

Fall 2:  $\emptyset H_2 HT(x_j, P)$  undefiniert,  $HT(x_j, P) < T$

Setze  $\emptyset H_2 HT(x_j, P) := T - HT(x_j, P)$  und  $HT(x_j, P) := T$ .

Fall 3:  $\emptyset H_2 HT(x_j, P)$  definiert,  $HT(x_j, P) < T$

Setze  $\emptyset H_2 HT(x_j, P) := (\emptyset H_2 HT(x_j, P) + T - HT(x_j, P)) / 2$  und  $HT(x_j, P) := T$ .

Falls  $HT(x_j, P) = T$  ändern sich die charakteristischen Werte nicht.

Im letzten Schritt definieren wir nun ein geeignetes Maß zur Bestimmung der Distanz zwischen Profilen.

#### 4 Maß zur Bestimmung der Distanz von Profilen

Um die Ähnlichkeit von Profilen  $P$  und  $Q$  im Rahmen des Profil-Matching zu bestimmen, benötigen wir ein Distanzmaß<sup>25</sup>  $d(P, Q)$  innerhalb des Profil-Raumes. Hierzu definieren wir zunächst für jede (Attribut-)Dimension  $A$  eine separates aber allgemeines, d. h. von der Skalierung der Dimension unabhängiges, Distanzmaß  $d_A(P, Q)$ . Offensichtlich sollte  $d_A(P, Q)$  von sämtlichen Ausprägungen abhängen die  $P$  und  $Q$  für das Attribut  $A$  besitzen. Bei der Distanz-Berechnung ist zu berücksichtigen, dass Profile im Sinne des Profil-Matching als ähnlich gelten, wenn sie gemeinsame Attribut-Ausprägungen mit einem hohen Informationsgrad besitzen. Profile die nur Attribut-Ausprägungen mit einem Informationsgrad nahe bei 0 besitzen, sollten hingegen eine besonders hohe Distanz aufweisen<sup>26</sup>.

Für Attribut  $A$  seien wiederum Profil  $P$  die Ausprägungen  $x_1, \dots, x_s$  und Profil  $Q$  die Ausprägungen  $y_1, \dots, y_r$  mit positivem Informationsgrad zugeordnet. Alle Attribut-Ausprägungen in  $Q$  die nicht in  $P$  liegen, werden auch in  $P$  eingefügt, allerdings mit Informationsgrad 0. Entsprechend verfährt man mit Attribut-Ausprägungen in  $P$

---

<sup>25</sup> Der Begriff Maß wird hier nicht im streng mathematischen Sinne verwandt, sondern beschreibt allgemein eine Berechnungsvorschrift für einen Vergleichswert.

<sup>26</sup> Hier würde eine "Ähnlichkeit" ja nur im Fehlen von echten Attribut-Ausprägungen bestehen. Dies ist aber beim Profil-Matching gerade nicht erwünscht.

die nicht in  $Q$  liegen. Anschließend sortiert man die Ausprägungen so um, dass  $x_j = y_j$  für  $j = 1, \dots, n$ . Dabei bezeichne  $n$  erneut die Anzahl der gemeinsamen Attribut-Ausprägungen in Profil  $P$  und  $Q$ .

Basierend auf den gemäß Formel (12) definierten Distanzen zwischen  $P$  und  $Q$  bezüglich der Ausprägungen des Attributes  $A$ , ermitteln wir die informationsbewertende Attribut-Distanz wie folgt:

$$(13) \quad d_A(P, Q) := 1 - \frac{\sum_{j=1}^n (1 - d_{j^*}(P, Q))}{\sum_{j=1}^n x_{j^*} + \sum_{j=1}^n y_{j^*} - \sum_{j=1}^n (1 - d_{j^*}(P, Q))}$$

Wie man sich leicht überlegt gilt  $d_A(P, Q) \in [0, 1]$ , mit  $d_A(P, Q) = 0$  gdw.  $P$  und  $Q$  dieselben Ausprägungen für Attribut  $A$  besitzen (d. h.  $s = r = n$ ) und allen diesen Ausprägungen der Informationsgrad  $1$  zugeordnet ist.

Die Attribut-Distanz wird also in Anlehnung an das – aus dem Information Retrieval bekannte – Ähnlichkeits-Maß von Tanimoto [Kohonen 97] definiert. Anders als viele andere entsprechend adaptierte Distanz- oder Ähnlichkeitsmaße vergrößert dieses Maß wunschgemäß die Ähnlichkeit zwischen Profilen die gemeinsame Ausprägungen mit einem hohen Informationsgrad besitzen und senkt die Ähnlichkeit zwischen Profilen, die nur Ausprägungen mit einem niedrigen Informationsgrad besitzen.

Der Gesamtabstand  $d(P, Q)$  zwischen  $P$  und  $Q$  wird schließlich als mittlere Attribut-Distanz zwischen  $P$  und  $Q$  festgelegt:

$$(14) \quad d(P, Q) := \frac{1}{k} \sum_{i=1}^k d_{A_i}(P, Q) .$$

Dabei kann es u. U. sinnvoll sein nur diejenigen Attribut-Distanzen zu berücksichtigen, die unterhalb eines vorzugebenden Schwellwertes liegen<sup>27</sup>.

## 5 Ausblick

Es wurde ein Konzept für ein Profilingsystem vorgestellt, in dessen Zentrum ein Profilmodell steht, bei dem jeder Attribut-Ausprägung ein Informationsgrad zuge-

---

<sup>27</sup> In diesem Fall sollte allerdings gewährleistet sein, dass beim Profil-Influencing ebenfalls nur diejenigen Attribute angepasst werden, die eine geringe Attribut-Distanz besitzen.

ordnet ist, der sowohl beim Profil-Influencing, als auch beim Profil-Matching geeignet berücksichtigt wird. Dadurch ist es möglich bei der Profilbildung auch mit Informationen geringer Validität bzw. mit Prognosen zu arbeiten. Zusammen mit dem Ansatz, neben Nutzer-Profilen auch Item-Profile zu verwenden und diese sich wechselseitig beeinflussen zu lassen, liefert dies eine Lösung für das Kaltstartproblem von Systemen zur automatisierten Personalisierung mittels Profilen.

Noch offen ist die Frage, ob sich das Konzept auch auf Systeme übertragen lässt, die mit hierarchischen Relationen zwischen Attribut-Ausprägungen arbeiten. Für den praktischen Einsatz bleibt ferner zu klären, wie die gleichzeitige Wechselwirkung zwischen mehr als zwei Profilen, sowie die Verschmelzung von Profilen zu handhaben sind. In der Vergangenheit wurden Teile des Konzeptes in einem Recommender-System des ehemaligen Internet-Portals „yoolia“ implementiert. Eine vollständige Realisierung und Evaluierung auf vergleichbarer Basis [Herlocker et al. 00] ist für die Zukunft im Rahmen von Diplom- bzw. Masterarbeiten sowie einem Forschungsprojekt geplant, für das noch Partner aus der Wirtschaft gesucht werden.

## 6 Literatur

- [Breese et al. 98] Breese, J. S., Heckermann, D., Kadie, C. „Empirical Analysis of Predictive Algorithms for Collaborative Filtering“, Proc. of Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, Morgan Kaufmann, 1998
- [Dunn et al. 97] Dunn, M., Gwertzman, J., Layman, A., Partovi, H. „Privacy and Profiling on the Web“, W3C Working Draft, 1997
- [Galliat 99] Galliat, T. „Clustering Data of Different Information Levels“, International Data Analysis Conf. (IDA), Innsbruck, 2000 (available as ZIB-Preprint SC-99-42, Berlin, 1999)
- [Herlocker et al. 00] Herlocker, J., Konstan, J., Terveen, L., Riedl, J. „Evaluating Collaborative Filtering Recommender Systems“, ACM Transactions on Information Systems 22, p. 5-53, ACM Press, 2004
- [Kobsa 04] Kobsa, A. „Adaptive Verfahren – Benutzermodellierung“, In: Kuhlen, R., Seeger, T., Strauch, D. (Hrsg.) „Grundlagen der praktischen Information und Dokumentation“, 5. Auflage, K.G. Saur, München, 2004
- [Kohonen 97] Kohonen, T. „Self-Organizing Maps“, 2nd Ed., p. 16-17, Springer, Berlin, 1997
- [Massa et al. 04] Massa, P., Avesani, P. „Trust-aware Collaborative Filtering for Recommender Systems“, Proc. of the Intern. Conf. on Cooperative Information Systems (CoopIS), Larnaca, 2004
- [Schein et al. 02] Schein A. I., Popescul A., Ungar L. H., Pennock D. M., „Methods and Metrics for Cold-Start Recommendations“, Proc. of the 25<sup>th</sup> Annual Intern. ACM SIGIR

- Conf. on Research & Development on Information Retrieval (SIGIR02), Tampere, Finland, 2002
- [Terveen et al. 01] Terveen, L., Hill, W. „Beyond Recommender Systems: Helping People Help Each Other“, In Carroll, J. (ed.), HCI in the New Millennium. Addison Wesley, 2001.
- [Ungar et al. 98] Ungar, L. H., Foster, D. P. „Clustering Methods für Collaborative Filtering“, Proc. of the Workshop on Recommendation Systems. AAAI Press, Menlo Park California, 1998
- [van Meteren et al. 00] van Meteren, R., van Someren, M. „Using Content-Based Filtering for Recommendation“, MLnet / ECML2000 Workshop, Barcelona, 2000
- [Wang et al. 06] Wang, J., de Vries, A. P., Reinders, M. J. T. „Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion“, Proc. of the 29<sup>th</sup> Annual Intern. ACM SIGIR Conf. on Research & Development on Information Retrieval (SIGIR06), Seattle, 2006