JOHANNES KEPLER
UNIVERSITÄT LINZ

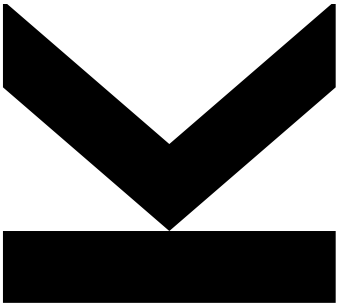
Eingereicht von
Manuel Steiner, 1055166

Angefertigt am
**Institut für
Anwendungsorientierte
Wissensverarbeitung**

Beurteiler
**A.Univ.-Prof. DI Dr.
Wolfram Wöß**

März 2016

Personalisierung von web-basierten Informationssystemen und deren Applikation im Tourismus



Masterarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Masterstudium

COMPUTER SCIENCE

**JOHANNES KEPLER
UNIVERSITÄT LINZ**
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696

Kurzfassung

Diese Arbeit beschäftigt sich mit der Personalisierung von web-basierten Informationssystemen. Dabei wird der Prozess zur Auswahl von individuell auf Anwender zugeschnittenen Inhalten sowie deren Darstellung anhand des Beispieles Tourismus erläutert. In solchen Systemen findet eine Vermittlung von Unterkünfte an Benutzer basierend auf deren Bedürfnissen und Interessen statt. Neben der Definition des Begriffes Personalisierung werden die Notwendigkeit für die Verwendung sowie Vorteile wie auch Probleme im Zusammenhang mit der Adaption erläutert. Anschließend werden unterschiedliche Möglichkeiten zur Verwaltung von Benutzerprofilen präsentiert. Unterschiedlichen Methoden zur Personalisierung behandelt die Arbeit gleichermaßen wie die notwendige Repräsentation von Dokumenten in einer aussagekräftigen Form. Den Abschluss bildet ein Vergleich Techniken, welche in bekannten Webapplikationen in Verwendung sind.

Abstract

This work is concerned with the personalisation of web-based informatin systems. The process of individual selected content as well as its representation is elucidated based on the example tourism. In such systems, accommodations are offered to users based on their needs and interests. The term personalisation is defined. Furthermore, necessities for the usage as well as advantages and issues in relation to adaptation are presented. Subsequently, different possiblities for user profiling are explained. Different personalisation methods are discussed, as is the necessary representation of documents in a meaningful way. Finally, a comparison between techniques, which are used in well-known web applications, is made.

Inhaltsverzeichnis

Aufgabenstellung

Der folgende Abschnitt beschreibt die Hintergrundinformationen bezüglich der Motivation zur Verfassung dieser Arbeit sowie die Zielsetzung.

Hintergrundinformation

Ohne den Einsatz von Methoden zur Personalisierung in web-basierten Informatinonssystemen wird jeder Anwender auf gleiche Weise behandelt. Dies bedeutet, dass nicht nur die Informationen für alle Benutzer gleich aufbereitet werden, sondern auch die gleiche Navigationsstruktur, das gleiche Layout, die gleichen Suchfunktionalitäten besitzen. Die Benutzeransprache kann mit diesem einheitlichen Konzept nur suboptimal erfolgen.

In Tourismussystemen besteht eine hohe Anzahl an unterschiedlichen, relevanten Kriterien zu einem personalisierten Ansatz. Sie können von regionalen (Berge, Seen, etc.) bis themenbezogenen (wie Skifahren, Snowboarden, etc.) Ausprägungen, von unterkunftsbezogenen (Unterkunftstypen, -kategorien, -ausstattungskriterien etc.) bis anderen touristischen Objekten (wie Sehenswürdigkeiten, Sport- und Freizeiteinrichtungen etc.), von zeitabhängigen (wie Veranstaltungen, Pauschalen etc.) bis anreisebedingten (Dauer der Anreise, Bahn, etc.) und vielen mehr reichen.

Zielsetzung

Anhand des Beispieles Unterkunftsvermittlung im Tourismus soll der Prozess zur Personalisierung von web-basierten Informationssystemen evaluiert werden. Dabei soll erläutert werden, warum die Adaption von Inhalten sowie deren Darstel-

lung notwendig sind. Weiters soll auf die einzelnen Phasen des Prozesses näher eingegangen werden. Dies umfasst die folgenden Punkte.

- Die Darstellung von Anwendern im Informationssystem sowie die Verwaltung der resultierenden Benutzerprofile.
- Die Aufbereitung von Dokumenten, um diese repräsentativ darstellen sowie mit Anforderungen von Anwender vergleichen zu können, um relevante Inhalte zu bestimmen.
- Unterschiedliche Möglichkeiten zur Individualisierung.

Für die Personalisierung bestehen vornehmlich folgende möglichen Methoden.

Content-Individualisierung bedeutet, dass den unterschiedlichen Benutztypen unterschiedlicher Content an ein und derselben Stelle eines Objekts angezeigt werden soll. Zum Beispiel sollte bei einem themenorientierten Benutzertyp bei einem Hotel, das sich auf Businessgäste genauso wie Wellnessgäste und Wanderer beziehungsweise Skifahrer orientiert hat, die themenspezifischen Ausprägungen wie Wellness, Wandern und Skifahren besonders hervorgehoben werden (im Homepagetext, in der Bildsprache auf der Willkommensseite der Unterkunft etc.).

Navigationsindividualisierung bedeutet die Anpassung der Navigationsstruktur an den jeweiligen Benutzertyp. Ein themenorientierter Benutzer, der sich inspirieren lassen will, erwartet von der Navigation her andere Elemente als ein Geschäftsreisender, der rasch seine vielleicht wöchentliche Geschäftsreise einplanen will.

Unter der Individualisierung der Such- beziehungsweise Filterfunktionalitäten wird die Anpassung dieser Funktionalitäten an die Benutzertypen verstanden. Während Geschäftsreisende zum Beispiel die Such- oder Filtermöglichkeit nach Lage der Unterkunft, Unterkunft mit 24-Stunden-Rezeption etc. wünschen, bevorzugen themenorientierte Benutzergruppen die Möglichkeit zum Suchen und Filtern nach speziellen Ausprägungen, wie das Vorhandensein eines absperrbaren Fahrradraums, einer Werkstatt, einem Sportlerfrühstück, einem Mountainbikeguide etc.

Kapitel 1

Einleitung

Die Massen an Informationen, welche in einzelnen, modernen web-basierten Informationssystemen verfügbar sind, können zunehmend nicht mehr durch manuelle Navigation zwischen Webseiten vermittelt werden. Weiters steigt der Bedarf, auch in Systemen, welche über das Internet verfügbar sind, individuell behandelt zu werden. Ein Systembenutzer soll sich wie in einem kleinen Geschäft fühlen [MR01]. Dabei soll er persönliche Beratung zu angebotenen Produkten oder Dienstleistungen erhalten sowie bei der Suche nach speziellen Objekten unterstützt werden. Um dies in Webapplikationen zu verwirklichen, benötigt es unterschiedlicher Komponenten, welche zusammen ein personalisiertes System zur Verfügung stellen.

Diese Arbeit beschäftigt sich mit der Personalisierung von web-basierten Informationssystemen. Dabei wird als Beispiel zur Veranschaulichung ein System aus dem Tourismusbereich zur Vermittlung von Unterkünften verwendet. Nach der Definition des Begriffes Personalisierung werden dessen Bedeutung sowie Verwendungszwecke erläutert. Weiters wird der Prozess zur Adaption von Inhalten sowie deren Darstellung in Webapplikationen vermittelt. Dabei wird genauer auf die Verwaltung von Benutzerprofilen, die repräsentative Darstellung von Dokumenten sowie unterschiedliche Methoden zur Anpassung eingegangen. Zum Abschluss wird ein Vergleich von Techniken präsentiert, welche in bekannten, produktiven Websystemen zum Einsatz kommen. Die einzelnen Kapitel sind nach der Reihenfolge der jeweiligen Schritte im Prozess zur Personalisierung angeordnet. Das Dokument ist folgendermaßen gegliedert.

In Kapitel 2 wird der Begriff Personalisierung erklärt. Weiters wird die Verwendung der Adaption von web-basierten Systemen erläutert. Unterschiedliche

Gründe zum Einsatz sowie die Vor- und Nachteile für Betreiber sowie Anwender werden ebenfalls vermittelt. Außerdem wird der grundsätzliche Prozess zur Anpassung von Inhalten und deren Darstellung an Bedürfnisse und Eigenschaften von Benutzern präsentiert.

Den ersten Schritt zur Personalisierung bildet die Erstellung und Verwaltung eines Benutzerprofils. Dieses Thema wird in Kapitel 3 behandelt. Es werden unterschiedliche Inhalte von Profilen erklärt. Weiters werden die verschiedenen Arten der Benutzeridentifikation und Ermittlung von relevanten Daten erläutert. Anschließend werden die gebräuchlichsten Darstellungsarten für Profile vorgestellt. Zum Schluss wird vermittelt, welche Arten von Benutzerprofilen für den Einsatz in Systemen, welche im Tourismusbereich verwendet werden, geeignet sind.

Kapitel 4 beschäftigt sich mit der repräsentativen Darstellung von Dokumenteninhalten sowie unterschiedlichen Methoden zur Ermittlung der Relevanz einzelner Inhalte für Anwender. Hierbei wird näher auf die Vorbereitung von unstrukturiertem Text eingegangen, so dass dieser als Vektor von Stichwörtern dargestellt und gewichtet werden kann. Anschließend werden die zwei bekanntesten Modelle der Informationsbeschaffung (Boolesches und Vektorraum-Modell) erläutert. Den Abschluss bildet eine Übersicht über weitere Modelle.

In Kapitel 5 werden die gängigsten Techniken zur Personalisierung von Inhalten sowie deren Darstellung in web-basierten Informationssystemen vermittelt. Hierbei wird eine Einteilung in drei Klassen vorgestellt. Weiters werden personalisierte Suchen sowie Empfehlungssysteme genauer betrachtet. Unterstützung in der Navigation durch adaptive Links sowie die Vermittlung von relevanten Teilen innerhalb eines Dokumentes werden ebenfalls behandelt.

Abschließend werden in Kapitel 6 Methoden erläutert, welche in bekannten Webapplikationen eingesetzt werden. Nachdem über das Thema im Bereich Tourismus keine verwendbaren Informationen vorhanden sind, werden Systeme aus unterschiedlichen Bereichen evaluiert, weswegen ein direkter Vergleich zwischen diesen Systemen nicht möglich ist. Jede Plattform wird separat behandelt.

Kapitel 2

Personalisierung von web-basierten Informationssystemen

Dieser Abschnitt führt in das Thema Adaption in web-basierten Informationssystemen ein. Es wird der Begriff der Personalisierung definiert, sowie deren Notwenigkeit in modernen Webapplikationen erläutert. Weiters wird ein generischer Prozess zur Anpassung von Inhalten und deren Darstellung präsentiert. Den Abschluss bildet eine Übersicht von Vorteilen sowie Nachteilen für Betreiber und Anwender bei der Verwendung von solchen Techniken in Systemen.

2.1 Der Begriff Personalisierung und seine Bedeutung

Ganz allgemein bedeutet Personalisierung die Anpassung von Objekten oder Dienstleistungen an die Bedürfnisse von Personen. Laut Serino, Furner und Smatt lässt sich die Personalisierung in web-basierten Systemen wie folgt zusammenfassen.

> Personalization is the use of information about a particular user that provides tailored or personalized services [...] involves automatic changes of Web pages to accommodate [...] user's needs, interests, knowledge, goals or tasks. [...] might also include [...] recommending products [...]. ([SFS05])

Des weiteren definieren Kobsa, Könemann und Pohl ein personalisiertes Websystem wie folgt.

> We define a personalised hypermedia application as a hypermedia system which adapts the content, structure and/or presentation of the networked hypermedia objects to each individual user's characteristics, usage behaviour and/or usage environment. ([KKP01])

Aus den beiden Aussagen ist zu erkennen, dass Informationen von Anwendern einer Webapplikation dazu verwendet werden, diese im System darzustellen. Diese Repräsentation wird anschließend herangezogen, um die Inhalte sowie die Darstellung dieser entsprechend anzupassen. Hierbei kann diese Adaption auf unterschiedlichen Eigenschaften des Benutzers beruhen. Dazu gehören Bedürfnisse, Interessen, Wissen, Ziele sowie Aufgaben. Weiters besteht aufgrund der gesammelten Informationen die Möglichkeit, einem Anwender Inhalte aus dem Websystem zu empfehlen, welche für diesen als bedeutsam oder interessant erachtet werden.

Personalisierung kann in gewisser Weise als Abbild der Realität gesehen werden. Wenn eine Person Dienstleistungen oder Produkte in einem Geschäft in Anspruch nimmt, bekommt sie dort üblicherweise eine Beratung von Verkäufern [MR01]. Dabei werden Kunden von Mitarbeitern eines Geschäftes mit dem Namen angesprochen, weil sich die Menschen persönlich kennen [MR01]. Die Art der Beratung reicht von Hilfe bei der Suche nach bestimmten Produkten bis hin zu Empfehlungen, falls sich der Kunde nicht sicher ist, was er genau benötigt [MR01]. Gute Beratung hat zur Folge, dass die entsprechenden Anwender zufrieden sind und in weiterer Folge den Betrieb wieder aufsuchen, falls sie ähnliche Produkte oder Dienstleistungen benötigten [MR01]. Zwischen Verkäufer und Kunde wird eine Beziehung aufgebaut. Genau diese Art von Verkaufsprozess soll durch Webapplikationen, welche Methoden zur Anpassung an Benutzer einsetzen, nachgeahmt werden [MR01]. Auch hier lässt sich feststellen, dass die häufigsten Methoden zur Unterstützung die Suche nach Informationen sowie deren Empfehlung ist. Im Unterschied zu physikalischen Geschäften werden diese Prozesse der Anpassung jedoch hauptsächlich vom System selbst verwaltet und durchgeführt [MR01]. Es benötigt demnach nur eine geringe Interaktion von der Seite des Betreibers. Der Kunde soll das Gefühl haben, dass er individuell betreut wird und seine speziellen Bedürfnisse berücksichtigt werden [MR01].

Betreiber andererseits wollen mit der Verwendung von personalisierten Inhalten und Darstellungen nicht nur die Zufriedenheit ihrer Kunden aufrecht erhalten. In erster Linie sollen diese Techniken zur Steigerung des Gewinnes beitragen [KKP01]. Dies wird erreicht, indem einem Anwender so rasch wie möglich zu gesuchten Informationen verholfen wird. Weiters soll dieser langfristig an die Webapplikation gebunden werden, damit weitere Transaktionen abgewickelt werden [KKP01]. Diese Ziele sollen dabei so effizient und kostensparend wie möglich erreicht werden. Dies bedeutet, mit möglichst wenig Personalaufwand ein maximales Ergebnis zu erzielen. Dafür werden effektive Methoden benötigt, die auf Benutzer zugeschnittene Informationen zur Verfügung stellen.

2.2 Notwendigkeit für die Adaption von Inhalten im World Wide Web

Ohne jegliche Personalisierung würden in Webapplikationen jedem Anwender die gleichen Inhalte auf die selbe Art und Weise präsentiert werden. Diese Vereinheitlichung funktioniert jedoch im Web nicht, weil wie in der Realität Personen unterschiedliche Charaktere, Eigenschaften und Bedürfnisse besitzen [MR01]. Um diese speziell zu adressieren, muss von der Massenorientierung abgewichen und ein Anwender als Individuum betrachtet werden. Nur dadurch kann eine Beziehung zwischen Betreiber und Kunde aufgebaut werden und aus Sicht des Betreibers eine Gewinnsteigerung erzielt werden. Neben der bereits beschriebenen Notwendigkeit für die Bindung der Kunden sowie deren Zufriedenstellung ergibt sich für einen Betreiber weiters der Vorteil, dass dieser nur einen geringen Aufwand hat, um die Inhalte im Informationssystem zu pflegen. Der Personalisierungsprozess geschieht zum Großteil automatisch. Dies bedeutet, dass keine weiteren Eingriffe vom Betreiber notwendig sind. Voraussetzung für einen reibungslosen Ablauf sind jedoch qualitativ hochwertige Daten sowie gut durchdachte Methoden zur Anpassung dieser. Nur durch eine korrekt eingespieltes System funktionieren Techniken wie Empfehlungen und Suchunterstützung [KKP01].

Anwender des Systems haben ebenfalls Vorteile bei der Verwendung eines personalisierten Systems. Einerseits werden diese bei der Informationsbeschaffung unterstützt [KKP01]. Dies bedeutet, dass die gesuchte Information schnellstmöglich verfügbar sein soll. Weiters kann dieser bisher unbekannte Produkte oder Dienstleistungen entdecken und zusätzliche Information aufnehmen. Langwieri-

ge Navigation durch die womöglich große Menge an Webseiten bleiben dadurch erspart. Durch Filterung von Inhalten, welche für den Benutzer von Bedeutung sind, wird dieser nicht mit unnötigen Informationen belastet [KKP01]. Somit kann er sich auf die wichtigen Daten konzentrieren und wird nicht abgelenkt. Die kognitive Überladung sinkt.

2.3 Probleme der Personalisierung

Der Prozess zur Adaption von Inhalten in Webapplikationen bringt nicht nur Vorteile für Betreiber und Anwender. Es bestehen auch einige Probleme, die in Kauf genommen werden müssen.

Aus Sicht des Betreibers wird der Aufwand, welcher in Personalkosten eingespart wird, möglicherweise in Computerhardware für Rechenleistung benötigt. Umso größer ein Informationssystem wird (durch die Anzahl der Benutzer und der vorhandenen Informationen), desto mehr Berechnungen müssen angestellt werden, damit Anwendern auf sie zugeschnittener Inhalt dargestellt werden kann [MR01]. Ab einem gewissen Datenaufkommen sind diese in Echtzeit nicht mehr zu bewältigen. Es müssen dann diverse Algorithmen im Hintergrund ausgeführt werden, um weiterhin Personalisierung betreiben zu können. Dies bedeutet, dass Einflüsse aus einer aktuellen Sitzung für den Anwender erst bei der nächsten Interaktion mit dem System in der Anpassung berücksichtigt werden. Bei einem automatisierten Prozess ist dies tolerierbar. Wenn der Benutzer jedoch interaktiv daran teilnimmt, erwartet er sich in der Regel sofortige Ergebnisse. Nicht zuletzt müssen alle zur Personalisierung benötigten Daten gespeichert werden [MR01]. Die Menge an Information, welche dabei zu verarbeiten ist, kann sehr rasch ausarten.

Die größte Einschränkung für den Benutzer liegt in dessen Privatsphäre. Für ein Websystem, welches Methoden zur Personalisierung einsetzt, ist es unumgänglich, dass Informationen über den Anwender gesammelt werden [MR01]. Dabei kann es sich um Stammdaten, Interessen, Bewegungsprofile und andere Daten handeln. Diese werden benötigt, damit ein Benutzerprofil verwaltet werden kann. Das Profil repräsentiert einen Anwender im System und die darin gespeicherten Informationen bilden die Grundlage für die Anpassung von Inhalten sowie der Darstellung dieser [KKP01]. Hierbei sollte von Betreiberseite natürlich darauf geachtet werden, dass diese Daten vertraulich behandelt sowie gespeichert werden [MR01]. Dies ist allerdings vom Anwender schwer zu überprüfen. Weiters ist

diesem in der Regel nicht exakt bekannt, welche Informationen über ihn gesammelt werden. Dies ist speziell bei Anwendungen der Fall, welche auf dem Endgerät der jeweiligen Person ausgeführt werden. Dadurch können alle möglichen Dateien an die Webapplikation gesendet werden. Ähnlich dazu kann bei der Anmeldung mittels sozialem Netzwerk an einer Applikation diese üblicherweise viele dort gespeicherten Daten abrufen und für eigene Zwecke verwenden. In der Regel wird ein Anwender zwar darüber informiert, meistens wird diese Meldung jedoch ignoriert, weil die Anwendung benutzt werden möchte. Ein weiteres Problem mit der Preisgabe von Daten ist, dass die meisten Benutzer nichts dagegen einzuwenden haben, wenn man ihnen gewisse Services zur Verfügung stellt und erklärt, wofür die Informationen benutzt werden [MR01]. Ob diese Angaben korrekt sind, kann ebenfalls schwer überprüft werden. Eine weitere Unannehmlichkeit für viele Anwender ist, dass sie Daten explizit zur Verfügung stellen müssen. Normalerweise möchten diese eine Anwendung benutzen, ohne viel zusätzlichen Aufwand zu betreiben [MR01]. Sie fühlen sich durch die Frage nach Feedback oder anderen Eigenschaften belästigt. Dies kann dazu führen, dass sie keine Interaktion mit dem jeweiligen System mehr ausüben.

2.4 Der Prozess zur Personalisierung

Ein Prozess zur Personalisierung kann in drei unterschiedliche Aufgaben unterteilt werden [KKP01]. In der Informationsbeschaffung werden Eigenschaften von Benutzern ermittelt [KKP01]. Diese werden in der Darstellungsphase durch eine bestimmte festgelegte Form ausgedrückt [KKP01]. In der finalen Phase kann diese Information für die Adaption verwendet werden. In diesem Fall ist die Produktion von personalisiertem Inhalt gegeben [KKP01]. Weitere Informationen zur Verwaltung von Benutzerprofilen finden sich in Kapiten 3, während unterschiedliche Methoden zur Personalisierung in Abschnitt 5 behandelt werden.

- Bei der Informationsbeschaffung werden Daten über Benutzer gesammelt [KKP01]. Dies können Charakteristiken, Interessen oder auch der aktuelle Kontext sein. Die Sammlung erfolgt durch die Analyse der Interaktion von Anwendern mit dem System. Alternativ können auch externe Quellen verwendet werden [KKP01]. Vom Benutzer wird ein Profil erstellt, welches diesen im System darstellen [KKP01]. Dabei soll die Repräsentation so genau wie möglich der Realität entsprechen. Die gesammelten Daten werden

weiters der Komponente zur Verfügung gestellt, welche für die Anpassung von Inhalten und der Darstellung zuständig ist [KKP01].

- In der Phase der Darstellung werden die Informationen aus dem Benutzerprofil formal aufbereitet, so dass sie in einer definierten Form vorliegen [KKP01]. Diese Art der Daten kann weiterverarbeitet werden. So kann etwa auf weitere Annahmen über den jeweiligen Benutzer aus den verfügbaren Daten geschlossen werden [KKP01]. Dadurch wird das Profil angereichert und es stehen erweiterte Möglichkeiten für die Personalisierung zur Verfügung.

- In der Produktion werden Inhalte sowie deren Darstellung und Navigationselemente an einen Benutzer angepasst, indem die verfügbaren Daten aus dem entsprechenden Profil verwendet werden [KKP01]. Basierend auf diesen können etwa Inhalte gefiltert oder die Darstellung nach Bedürfnissen angepasst werden.

Die Informationen in einem Benutzerprofil werden in einer repräsentativen Form gespeichert. Weiters werden auch Dokumente aus dem Informationssystem klassifiziert. Dazu können etwa Vektoren von Stichwörtern verwendet werden. Durch die Darstellung im gleichen Format besteht die Möglichkeit eines Vergleichs [MSM07]. Durch diesen kann die Ähnlichkeit zwischen einem Profil und einem Dokument ermittelt werden [MSM07]. Der errechnete Wert kann danach verwendet werden, um festzustellen, ob entsprechende Inhalte für den jeweiligen Anwender von Bedeutung sind [MSM07]. Diese Methode wird beispielsweise in personalisierten Suchen sowie Empfehlungssystemen eingesetzt. Eine genauere Beschreibung der Darstellung von Dokumenten erfolgt in Kapitel 4.

Kapitel 3

Erstellung und Verwaltung von Benutzerprofilen

In diesem Kapitel werden Benutzerprofile für die Personalisierung in webbasierten Applikationen erläutert. Dabei wird erklärt, weshalb solche Profile notwendig sind und welche Informationen darin gespeichert werden. Weiters werden der Prozess für die Profilierung eines Benutzers veranschaulicht sowie unterschiedliche Arten von Profilen vorgestellt. Diese werden im Abschluss für den Einsatz in Systemen mit Bezug auf den Tourismus verglichen.

In der Literatur werden Benutzerprofile mit Benutzermodellen oftmals gleichgesetzt, manche Autoren unterscheiden diese Begriffe [Koc01]. Diese bezeichnen die Einheit, welche die gespeicherten Informationen enthält, als Benutzermodell [Koc01]. Ein Benutzerprofil wird als eine Art zur Erstellung und Verwaltung eines Modells angesehen [Koc01] [ZG07]. In den folgenden Abschnitten werden die Begriffe Benutzerprofil und Benutzermodell als synonym angesehen.

Ein Benutzerprofil repräsentiert einen Anwender im Informationssystem [Koc01]. Es beinhaltet alle personenbezogenen Daten, welche für die Applikation von Bedeutung sind [Koc01] in maschinenlesbarer Form. Dies ermöglicht die Verwendung im System. Solch ein Profil ist üblicherweise spezifisch für eine Anwendung. Es wird im Normalfall von dieser erstellt und verwaltet. Durch diese Möglichkeit zur Darstellung von Anwendern können diese von einander unterschieden werden [Kay01]. Es werden für jede Person unterschiedliche Charakteristiken gespeichert [Kay01]. Diese Unterscheidbarkeit bildet die Grundlage für mögliche Personalisierungen.

Der zugehörige Prozess zur Verwaltung von Benutzerprofilen stellt sicher, dass geänderte Charakteristika der Anwender im System widergespiegelt werden [ZG07]. Somit entsprechen die gespeicherten Informationen zu jedem Zeitpunkt möglichst genau den aktuellen Eigenschaften der Personen. Neben dieser wichtigen Aufgabe gehören zum Lebenszyklus der Profilierung die Informationsbeschaffung, Erstellung und Löschung von Profilen sowie die Aufbereitung der Daten für die Verwendung im System [ZG07].

Informationssysteme, welche keine Anwenderdaten verwalten, behandeln jeden Benutzer auf die gleiche Weise [Koc01]. Es besteht keine Möglichkeit zur Personalisierung, weil die Benutzer nicht unterschieden werden können [ZG07]. Dies bedeutet für web-basierte Systeme vor allem, dass die Darstellung der Informationen statisch ist und eventuell für den Anwender unwichtige Daten präsentiert werden. Dies kann dazu führen, dass durch die subjektive Empfindung von Unübersichtlichkeit oder Reizüberflutung das Angebot nicht weiter in Anspruch genommen wird.

3.1 Benutzerdaten für Personalisierungszwecke

Informationen über Systemanwender, welche in Profilen gespeichert werden, sind applikationsspezifisch. Jedes System speichert jene Daten, welche später weiterverarbeitet und für Personalisierungszwecke verwendet werden können. Die wichtigsten Merkmale hierbei sind Wissen, Interessen, Ziele und Aufgaben, Hintergrundinformation, individuelle Charakteristik sowie der Kontext eines Benutzers [BM07] [SD10]. Diese werden in den folgenden Abschnitten genauer erläutert.

3.1.1 Wissen

Der Wissensstand eines Anwenders kann in Bezug auf das gesamte Domänenwissen repräsentiert werden [BM07]. Diese Art von Information wird vorwiegend in adaptiven Lernsystemen oder wissensbasierten Systemen verwendet [BM07]. Das System kann somit entscheiden, welche Inhalte dem Benutzer bereits bekannt oder geläufig sind und welche ihm basierend auf dieser Information als nächstes angezeigt werden [BM07] [SD10]. Dieses Merkmal ist ideal für Lernplattformen, in denen Inhalte auf anderen aufbauen. Dadurch werden dem Anwender keine

Aufgaben gestellt, die für ihn eventuell noch nicht lösbar sind. Weiters kann im Verlauf von Interaktionen mit dem System falsches Wissen festgestellt werden und entsprechende Inhalte zu späteren Zeitpunkten erneut in den Lernprozess integriert werden.

Im Unterschied zu anderen Merkmalen muss bereits vor dem Betrieb des Systems ein Domänenwissen vorhanden sein [BM07]. Dieses wird als Referenz verwendet und Inhalte werden den Wissenskategorien zugeordnet. Diese Aufgabe muss von Experten im jeweiligen Bereich durchgeführt werden [BM07]. Erst durch diese Zuordnung kann eine Personalisierung erfolgen, indem das Benutzerprofil einen Teil dieses Wissens repräsentiert.

Für Informationssysteme im Bereich des Tourismus ist das Merkmal Wissen von geringer Bedeutung. Um Angebote wie Unterkünfte zu personalisieren, sind Eigenschaften wie Interssen oder Ziele und Aufgaben besser geeignet.

3.1.2 Interessen

Interessen der Anwender sind in kommerziellen Websystemen das wichtigste Merkmal [BM07]. Weiters sind diese Informationen oftmals die einzigem, welche zu Benutzern gespeichert werden [BM07]. Dies ergibt sich daraus, dass in typischen Informationssystemen die zur Verfügung stehenden Inhalte in irgendeiner Form mit Interessen von Benutzern verknüpfen lassen [BM07] [SD10]. Während Wissen für Verkaufsplattformen, wie etwa Unterkünfte im Tourismus, geringe Bedeutung haben, können die angebotenen Objekte Kategorien zugeordnet werden. Diese können wiederum mit Benutzerprofilen assoziiert werden, um etwa dem Benutzer Empfehlungen vorzuschlagen.

Im Bereich des Tourismus sind Interessen wohl das wichtigste Merkmal für die Personalisierung. Basierend auf Vorlieben wie etwa Freizeitaktivitäten können passende Unterkünfte angeboten werden.

3.1.3 Ziele und Aufgaben

Dieses Merkmal bezeichnet den Grund, den ein Anwender hat, um mit dem System zu interagieren [BM07]. Das Ziel einer aktuellen Sitzung kann etwa das Auf-

finden von spezifischen Informationen oder die Erledigung einer Aufgabe sein [BM07]. Der Benutzer will ein bestimmtes Ziel erreichen, welches das Ergebnis auf die Frage nach dem Grund der Interaktion liefert [BM07]. Durch Erkennung dieses Zieles kann dem Benutzer gezielte Information zur Verfügung gestellt werden.

Um Ziele identifizieren zu können, muss ein Katalog von möglichen Zielen zur Verfügung stehen [BM07]. Diese werden von Experten im jeweiligen Bereich festgelegt. Das System versucht, Ziele von Sitzungen zu erkennen und einem der vorgefertigten Ziele zuzuordnen [BM07]. Dieser Prozess ist jedoch sehr anspruchsvoll, nicht zuletzt weil sich das Ziel eines Anwenders üblicherweise von einer Sitzung zur anderen verändert [BM07]. Die einfachste Variante zur Bestimmung ist jene durch den Benutzer selbst. Dieser kann eines aus den vordefinierten Zielen auswählen [BM07]. Eine Alternative ist die Verwendung von Wahrscheinlichkeiten, um das mögliche Ziel der aktuellen Sitzung einem definierten zuzuordnen [BM07].

Im Fall von Systemen mit Bezug auf den Tourismus ist das Ziel eines Anwender meist das Finden einer passenden Unterkunft nach seinen Vorgaben. Deswegen sind Ziele und Aufgaben für die Personalisierung in diesem Anwendungsgebiet eher von untergeordneter Bedeutung. Ein möglicher Einsatz wäre die Unterscheidung zwischen Suche nach Unterkünften, Bearbeitung einer reservierten Unterkunft oder Änderung der Einstellungen. Da diese unterschiedlichen Aufgaben jedoch von Beginn an relativ klar unterscheidbar sind, werden diese Möglichkeiten normalerweise direkt auf der Startseite durch Links geboten. Eine Zielidentifikation ist dadurch überwiegend unnötiger Mehraufwand in der Applikation.

3.1.4 Hintergrundinformationen

Hintergrundinformationen von Benutzern bezeichnen allgemeine Charakteristiken, welche nicht in Verbindung mit dem System stehen, in dem sie verwendet werden [BM07]. Diese Merkmale können auch nur explizit durch den Anwender zur Verfügung gestellt werden, weil sie nicht von der Applikation eruiert werden können [BM07]. Zu diesen Daten zählen etwa das Alter, der Familienstand oder der Beruf.

Üblicherweise werden solche Informationen in Stereotypen eingeteilt, um keine detaillierten Daten speichern zu müssen [BM07]. Das Alter von Personen lässt

sich etwa in Intervalle einteilen, welche für die Applikation sinnvoll sind. Weiters ändern sich diese Merkmale nicht durch die Interaktion mit dem System [BM07]. Deshalb ist eine Aktualisierung der Daten nicht nötig und es sind keine individuellen Profile gerechtfertigt.

Merkmale wie der Beruf oder der Familienstatus können für Plattformen im Tourismusbereich nützlich sein, um etwa Empfehlungen für Anwender zu ermitteln oder die Ergebnisse bei der Suche nach Unterkünften entsprechend anzupassen. Verheirateten Anwendern können etwa standardmäßig Zimmer mit Doppelbetten vorgeschlagen werden. Für ältere Personen könnten Eigenschaften wie barrierefreier Zutritt in Betracht gezogen werden.

3.1.5 Individuelle Charakteristik von Benutzern

Merkmale zur Beschreibung von individuellen Charakteristika eines Anwenders definieren zusammengefasst ein Individuum. Ähnlich zu Hintergrundinformationen können solche Daten nur über lange Zeit oder gar nicht verändert werden. Informationen, welche diesem Bereichen von Eigenschaften zugeordnet werden, sind etwa die Persönlichkeit (introvertiert, extrovertiert), kognitive Eigenschaften und Lernstile. Diese Charakteristika werden üblicherweise durch psychologische Tests ermittelt und können nicht durch einmalige Befragung eines Benutzers festgestellt werden. Während individuelle Merkmale für die Personalisierung wichtig sind, werden größtenteils nur kognitive Stile und Lernstile berücksichtigt. [BM07]

Unter dem kognitiven Stil versteht man die bevorzugte Weise, Informationen darzustellen und zu organisieren. Hierbei bestehen einige Unterscheidungsmöglichkeiten. Für die Personalisierung, speziell für den Navigationssupport, ist die Aufteilung in feldabhängig bzw. feldunabhängig sowie holistisch bzw. serialistisch von Bedeutung. Bei ersterer Kategorisierung bedeutet feldabhängig, dass ein vorhandenes Umfeld bewusst beeinflusst, was von einer Person wahrgenommen wird. Feldunabhängigkeit bedeutet, dass eine Person weniger vom Umfeld abgelenkt werden kann. Im zweiten Unterscheidungsfall bedeutet holistisch, dass sich eine Person Inhalte des gesamten Systems merken und serialistisch veranlagte Personen Information seriell aufnehmen. [BM07]

Lernstile bezeichnen die Art, auf welche Personen bevorzugt neue Information erlernen [BM07]. Dies ermöglicht etwa unterschiedliche Darstellung von Informa-

tion, welche für den jeweiligen Typ von Bedeutung ist oder Filterung des Inhaltes passend zum Lernstil [BM07]. Da dieses Merkmal auf adaptive Lernsysteme beschränkt ist, wird es hier nicht genauer erläutert.

3.1.6 Aktueller Kontext

Der Kontext eines Benutzers in einer Sitzung bezeichnet Eigenschaften, welche nicht direkt dem Benutzer zugeordnet sind, sondern durch andere Gegebenheiten bestimmt werden [BM07]. Zu diesen zählen etwa die verwendete Plattform (das Gerät, welches vom Anwender benutzt wird) oder der aktuelle Aufenthaltsort des Benutzers. Obwohl diese Informationen nicht direkt eine Person widerspiegeln, sind sie für die Personalisierung von Informationssystemen von Bedeutung [BM07]. Weiters werden üblicherweise einige kontextbezogene Eigenschaften in Benutzermodellen gespeichert [BM07].

Im folgenden werden die wichtigsten Kontexteigenschaften erläutert [BM07].

- Die Benutzerplattform identifiziert die Eigenschaften des verwendeten Endgerätes. Dazu zählen etwa Anzeigegröße, Betriebssystem, Art der Benutzereingabe repräsentieren. Durch diese Angaben kann vorwiegend die Darstellung der Webseiten angepasst werden um eine optimale Benutzerinteraktion zu gewährleisten.

- Der Aufenthaltsort des Benutzers kann verwendet werden, um etwa die Sprache entsprechend einzustellen. Weiters kann der Ort für Empfehlungen verwendet werden. Speziell im Tourismus können Unterkünfte in der Nähe zum aktuellen Aufenthaltsort bei Suchergebnissen ganz nach oben gereiht werden.

- Der persönliche Kontext besteht aus Dimensionen der Umgebung und menschlichen Dimensionen. Zu den Eigenschaften der Umgebung zählen etwa Licht oder Temperatur. Persönliche Informationen sind etwa Puls oder Blutdruck.

- Der Gefühlsstatus einer Person kann durch die Interaktion von Anwendern mit dem System ermittelt werden. Somit lassen sich etwa Motivation, Frustration oder Engagement feststellen.

Einige kontextbezogene Eigenschaften sind für den Bereich Tourismus sehr nützlich. Der Aufenhaltsort kann etwa für die Reihung von Suchergebnissen verwendet werden. Weiters kann die Darstellung an das jeweilige Endgerät angepasst werden.

3.1.7 Benutzerspezifische Applikationseinstellungen

Aktuelle Präferenzen eines Anwenders können ebenfalls in dessen Benutzerprofil gespeichert werden [KKP01]. Solche Einstellungen sind normalerweise vom Anwender festgelegt und ändern sich nur langfristig. Beispiele für solche Eigenschaften sind etwa Vorzüge für gewisse Anzeigeelemente oder das Aussehen der Applikation. Für solche Präferenzen ergibt es in der Regel keinen Sinn, sie durch den Personalisierungsprozess zu evaluieren und entsprechend anzupassen. Das Endergebnis ist zwar auf eine Person abgestimmt, jedoch ist die Auswahl über die unterschiedlichen Möglichkeiten fix vorgegeben und für alle Systemanwender gleich (wie zum Beispiel eine gewisse Anzahl an auswählbaren Seitendesigns). Weiters sind auch die Adaptionsschritte fest vorgegeben. Für die Einstellung, ob ein Anwender Suchergebnisse von passenden Unterkünften etwa nach der Relevanz oder dem Preis sortiert angezeigt bekommen möchte, gibt es nur zwei Möglichkeiten. Die Sortierung kann direkt ohne speziellen Prozess in die Darstellung implementiert werden. Dieses Beispiel zeigt außerdem, dass es unwahrscheinlich ist, diese Eigenschaft von einem Besuch zum nächsten zu verändern. Üblicherweise setzt man solche Vorlieben einmalig in den Benutzereinstellungen, etwa nach der Registrierung im System.

3.2 Der Prozess zur Verwaltung sowie Verwendung von Benutzerprofilen

Um Benutzerprofile zu verwalten sowie für die Personalisierung von Inhalten in Informationssystemen einzusetzen, sind insgesamt drei Schritte notwendig [Gau+07]. Diese werden iterativ angewandt, um bei jeder Sitzung neue oder geänderte Merkmale von Anwendern in das Profil einfließen zu lassen. Dieses wird somit ständig aktuell gehalten und dem Personalisierungssystem stehen die bestmöglichen Daten zu Verfügung. Die folgenden Schritte beschreiben diesen

Prozess, welcher als Profilierung bezeichnet wird [Gau+07]. Abbildung 3.1 stellt die Profilverwaltung sowie -verwendung grafisch dar.

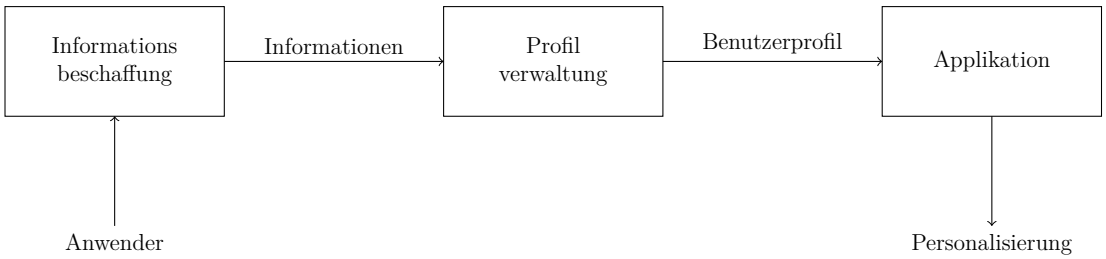


Abbildung 3.1: Der Prozess zur Verwaltung von Benutzerprofilen

1. Zuerst müssen Benutezrdaten von Anwendern gesammelt werden. Diese Informationen können entweder explizit vom Anwender zur Verfügung gestellt werden oder impliziert durch Beobachtung ermittelt werden [Gau+07]. Wichtig ist, dass die erfassten Daten für die Applikation von Bedeutung sind. Dadurch wird die spätere Verwendung für die Personalisierung ermöglicht. Weiters kann die Informationsbeschaffung durch unterschiedliche Methoden erfolgen, welche im Verlauf dieses Kapitels noch genauer erläutert werden. Um gesammelte Daten einem Anwender zuordnen zu können, müssen diese vom System eindeutig identifizierbar sein. Auch hierfür gibt es einige verschiedene Möglichkeiten.

2. Die gesammelten Informationen werden verwendet um ein Profil aufzubauen beziehungsweise zu verwalten [Gau+07]. Dies geschieht in der Regel durch ein eigenes Teilsystem der Applikation [Gau+07]. Die Verwaltungskomponente sorgt nicht nur für ein einmaliges Erstellen des Benutzerprofils, sondern ist auch für die ständige Änderung zuständig, um eine möglichst realitätsnahe Abbildung der Anwender im System zu gewährleisten. Unterschiedliche Arten wie etwa Vektoren von Stichwörtern oder Semantische Netze können zur Speicherung der Daten verwendet werden. Diese repräsentieren Merkmale von Benutzern in strukturierter, maschinenlesbarer Form, sodass sie von Personalisierungskomponenten weiterverarbeitet werden können. Die wichtigsten Typen werden in diesem Kapitel behandelt.

3. Die Applikation kann nun die vorhandenen Benutzerinformationen zur Personalisierung des Systems verwenden [Gau+07]. Hierfür gibt es unterschiedliche Möglichkeiten von der Adaption der Darstellung über Empfehlungs-

systeme zu personalisierter Suche. Die bedeutendsten Methoden sind in Kapitel 5 beschrieben.

Anzumerken ist, dass sich Benutzerinformationen von einer Sitzung zur nächsten oder sogar innerhalb einer Sitzung rasch ändern können [Gau+07]. Dabei ändern sich einige Charakteristiken von Benutzern schneller als andere [Gau+07]. Einige Systeme benutzen aus diesem Grund unterschiedliche zeitbasierte Profile für einen Anwender [Gau+07]. Somit ist es möglich, die Interessen innerhalb einer Sitzung gesondert zu behandeln, ohne dass die dabei gesammelten Informationen längerfristige Eigenschaften des Anwenders beeinflusst [Gau+07]. Wenn eine Person etwa üblicherweise ein Hotelbuchungssystem benutzt und dabei Zimmer mit Doppelbett bevorzugt, bedeutet eine einmalige Suche nach einer Unterkunft mit Einzelzimmer nicht automatisch, dass dies in Zukunft immer der Fall ist. Für die aktuelle Sitzung ist diese Eigenschaft jedoch von Relevanz.

3.2.1 Identifikation von Benutzern

Um gesammelte Informationen über Anwender diesen zuordnen zu können, müssen sie eindeutig innerhalb des Systems identifizierbar sein [Gau+07]. Dies kann über unterschiedliche Methoden geschehen. Jede besitzt andere Eigenschaften. Speziell der Benutzeraufwand und die Datensicherheit sind Faktoren, die für Anwender von hoher Bedeutung sind [Gau+07]. Weiters unterscheiden sich die Mechanismen darin, wie lange Informationen im System gespeichert und wie effektiv einzelne Benutzer voneinander unterschieden werden können [Gau+07]. Die Möglichkeiten für die Identifizierung reichen von Software, welche dediziert auf Endgeräten laufen muss bis zu reiner Verarbeitung auf Serverseite der web-basierten Applikation. Die fünf in diesem Abschnitt festgehaltenen Methoden können dabei benutzt werden.

3.2.1.1 Internet Protokoll Adressen

Eine für den Benutzer vollkommen transparente Möglichkeit zur Identifizierung sind Internet Protokoll (IP) Adressen. Ein Webserver kann die IP Adresse des anfragenden Gerätes bei jeder Kommunikation eruieren und weiterverarbeiten. Das bedeutet, dass diese auch an die Applikation weitergegeben und somit einem Profil im System zugeordnet werden kann. Hierbei wird ein Profil jeweils für ei-

ne IP Adresse verwaltet. Durch den Aufbau von IP Netzen und die Aufteilung der Adressen kann nicht garantiert werden, dass eine IP Adresse eindeutig einer Person zugeordnet ist [Rek+96]. Üblicherweise sind in einem privaten Netzwerk mehrere Endgeräte vorhanden [Rek+96]. Diese gelangen über den Internetanschluss zum gewünschten Informationssystem. Hierbei gibt es eine öffentliche IP Adresse [Rek+96]. Diese wird im öffentlichen Netz verwendet, um mit anderen Teilnehmern zu kommunizieren [Rek+96]. In diesem Fall werden alle Personen in einem lokalen Netzwerk, welche über die selbe öffentliche IP Adresse ins Internet gelangen, dem selben Profil zugeordnet [Rek+96]. Weiters ist es von Internetanbietern üblich, öffentliche IP Adressen dynamisch zu vergeben. Dies bedeutet, dass ein Anschluss mit wechselnden Adressen im Internet kommuniziert [Rek+96]. Somit werden sogar Personen von unterschiedlichen Anschlüssen einem Profil zugewiesen. Auf Grund dieser Nachteile sollte diese Methode nicht für zuverlässige Anwenderidentifizierung eingesetzt werden. Außerdem stehen bessere Verfahren zur Verfügung.

3.2.1.2 Cookies

Der Einsatz von Cookies ist ebenfalls mit keinem Aufwand für den Benutzer verbunden. Cookies ermöglichen Webanwendungen, diverse Einstellungen oder andere Eigenschaften über den Browser des Anwenders auf dessen Endgerät zu speichern [Gau+07]. Bei erstmaliger Kommunikation mit dem System kann dieses eine eindeutige Kennung generieren und den Client auffordern, diese zu speichern. Bei nachfolgenden Verbindungen kann diese ID wieder abgefragt werden, wodurch eine Zuordnung zum Benutzerprofil ermöglicht wird. Dies ermöglicht die Zuordnung zu einem Benutzerprofil. Eine eindeutige Unterscheidung von Anwendern ist auch mit dieser Methode nicht gegeben. Falls mehrere Personen ein Endgerät gemeinsam benutzen und weder unterschiedliche Systembenutzer noch mehrere Browserprofile verwenden, werden für diese Anwender gemeinsame Cookies verwendet [Gau+07]. Dies resultiert in der selben Benutzerkennung für das Informationssystem und dadurch erfolgt die Zuordnung von multiplen Personen zu einem Benutzerprofil [Gau+07]. Weiters besteht das Problem, dass ein Anwender mehrere Endgeräte verwenden kann, um mit der Applikation zu kommunizieren [Gau+07]. In diesem Fall wird für jedes Gerät eine Kennung generiert. Dies hat zur Folge, dass separate Benutzerprofile für eine Person verwaltet werden und die Personalisierung je nach verwendetem Endgerät unterschiedlich ausfällt [Gau+07]. Falls Cookies gelöscht werden, geht die Benutzerkennung ver-

loren und alle bisherig gesammelten Informationen sind wertlos [Gau+07]. Es muss ein neues Profil angelegt werden. Cookies sind heutzutage in nahezu jeder Webanwendung in Verwendung. Meist werden sie jedoch nur zur Speicherung von generellen, nicht kritischen Daten verwendet. Jedoch bleibt der Nachteil, dass Informationen auf Anwendergeräten gespeichert werden. Dies ist unter Umständen nicht erwünscht, da dieses Verfahren von vielen als Einschränkung der Privatsphäre angesehen wird. Für temporäre Identifikation von Anwendern sind Sitzungen besser geeignet.

3.2.1.3 Sitzungen

Mit Cookies haben Sitzungen gemeinsam, dass auch bei dieser Methode bei der initialen Kommunikation eine Kennung für den Anwender generiert wird. Diese ist jedoch eindeutig dem Anwender zugeordnet. Sie wird dem Webbrowser mitgeteilt, welcher sie temporär speichert und bei jeder erneuten Anfrage mitsendet. Die Eindeutigkeit kommt erst dadurch zu Stande, dass eine möglichst kurze Timeout Zeit gewählt wird. Bekommt das System innerhalb dieser definierten Zeit keine neue Anfrage mit einer Kennung, wird die Zuordnung aufgehoben. Dies bedeutet jedoch auch, dass dieses Verfahren nur für kontinuierliche Interaktionen funktioniert und keine dauerhafte Profilverwaltung möglich ist [Gau+07]. Innerhalb einer Sitzung können jedoch Informationen gesammelt und Methoden zur Personalisierung angewandt werden [Gau+07]. Sitzungs-IDs sind für den Benutzer ebenfalls transparent. Dadurch, dass keine Daten längerfristig auf Endgeräten gespeichert werden, eignen sie sich für die temporäre Anwendung. Hierbei werden sie üblicherweise mit anderen Methoden, vorwiegend Anmeldungen am System, kombiniert. Dies ermöglicht eine ungestörte Interaktion eines Anwenders mit der Webapplikation. Zusätzlich besteht die Möglichkeit, dass die Daten durch Registrierung des Benutzers längerfristig erhalten werden können.

3.2.1.4 Proxy Server

Bei der Verwendung von Proxy Servern läuft die gesamte Kommunikation zwischen Anwender und Webapplikation über eine Zwischenstelle [Gau+07]. Diese identifiziert den Anwender beziehungsweise das Endgerät und teilt die Benutzerkennung der Webapplikation mit. Diese Methode erfordert es, dass sich Personen am Proxy Server registrieren [Gau+07]. Üblicherweise identifiziert sich ein Be-

nutzer mit Benutzername und Passwort. Diese Anmeldeinformationen müssen danach in allen Webbrowsern eingetragen werden, welche vom Anwender benutzt werden. Falls der Proxy Server keine Anmeldung erfordert, wird die Zuordnung von Anwendern basierend auf IP Adressen der Endgeräte vorgenommen. Dies zieht die gleichen Probleme nach sich, welche auch bei IP basierter Identifikation am Websystem vorhanden sind. Weiters können am Proxy Server jegliche übertragenen Informationen abgegriffen und für andere Zwecke gespeichert oder an andere Geräte übermittelt werden. Eine sichere Kommunikation zwischen Client und Server ist somit nicht mehr möglich. Weiters muss der Proxy stets erreichbar sein. Ist dieser nicht funktionsfähig, oder so konfiguriert, dass er nicht von jeder beliebigen IP Adresse angesprochen werden kann, ist es dem Endgerät nicht möglich, mit dem Informationssystem zu kommunizieren. Solche Proxy Server werden öfters etwa für das kontrollierte Sperren von Webinhalten in Unternehmen als für Benutzeridentifizierung an Websystemen verwendet.

3.2.1.5 Software auf den Endgeräten der Anwender

Dedizierte Software auf Clientseite kann ebenfalls verwendet werden, um Anwender zu identifizieren. Solche Programme kommunizieren meist über ein eigenes Protokoll mit dem Server [Gau+07]. Dabei ist es dem System überlassen, in welcher Form die Identifizierung stattfindet [Gau+07]. In der Regel besteht hierbei wie bei Cookies das Problem, dass mehrere Anwender einem Benutzerprofil zugeordnet werden, falls diese das selbe Profil auf dem Endgerät verwenden. Um dies zu vermeiden, müsste es in der Software die Möglichkeit geben, zwischen Anwendern zu wechseln oder eine separate Anmeldung der Anwender im Programm vorhanden sein. Weiters entsteht hier Aufwand, weil diese Programme, auch Softwareagenten genannt, installiert werden müssen [Gau+07]. Der gravierendste Nachteil jedoch ist jener der Einschränkung der Privatsphäre. Dadurch, dass die Software auf den Anwendergeräten ausgeführt wird, können in der Regel auch jegliche andere Informationen an den Server übermittelt werden. Üblicherweise werden solche Agenten nicht nur zur Identifizierung, sondern auch zur Sammlung von Daten über den Anwender verwendet [Gau+07]. Einerseits ermöglicht es diese Methode, einen größeren Umfang an persönlichen Informationen für das Benutzerprofil zu sammeln. Andererseits hat der Anwender darauf meist keinen Einfluss und kann diesen Informationsbeschaffungsprozess nicht einschränken. Aus diesem Grund wird diese Art der Benutzeridentifizierung selten für web-basierte Systeme eingesetzt.

3.2.1.6 Anmeldung an der Webapplikation

Die derzeit wohl am weitesten verbreitete Methode zur Benutzerkennung ist die Verwendung von Anmeldeinformationen in Form von Benutzername oder Email-Adresse und Passwort. Auf diese Weise müssen keine Daten auf den Endgeräten gespeichert werden. Nach einmaliger Registrierung ist ein Benutzerprofil angelegt [Gau+07], welches danach eindeutig einem Anwender zugeordnet werden kann. Bei jeder neuen Verbindung zwischen Client und Server muss sich der Anwender am System anmelden [Gau+07]. In Kombination mit Sitzungen kann dadurch innerhalb dieser ohne erneute Eingabe von Benutzername und Passwort auf die Webseiten des Systems zugegriffen werden. Nach festgelegter Timeout Zeit wird die Sitzung inaktiv und es wird eine erneute Anmeldung nötig. Ein Nachteil ist die benötigte Benutzerinteraktion bei der Registrierung beziehungsweise Anmeldung [Gau+07]. Der große Vorteil liegt jedoch in der Datensicherheit. Es können lediglich Informationen vom System für das Benutzerprofil verwendet werden, welche die Applikation betreffen. Dies sind etwa Log-Dateien oder freiwillige Angaben der Anwender. Lediglich durch explizite Zustimmung eines Benutzers können von anderen Webapplikationen wie sozialen Netzwerken zusätzliche Daten gesammelt werden.

3.2.2 Informationsbeschaffung

Für das Sammeln von Daten über Anwender gibt es unterschiedlichste Methoden, welche verschieden klassifiziert werden können. Dabei differenzieren sich die Verfahren in Eigenschaften wie der mögliche Umfang an Daten, welche für Benutzerprofile gesammelt werden oder ob explizite Angaben von Anwendern notwendig sind. Meistens ist es zudem sinnvoll, einen hybriden Ansatz zu wählen, damit ab dem Erstellungszeitpunkt eines Benutzerprofiles genügend Informationen für Personalisierungszwecke vorhanden sind. Folgende Dimensionen können unterschieden werden.

3.2.2.1 Aktiv (explizit) oder passiv (implizit)

Die Unterscheidung zwischen aktiver (expliziter) und passiver (impliziter) Beschaffungstechnik liegt in der Interaktion mit und Einbeziehung der Systemanwender [Chi93]. Werden diese etwa durch Befragungen mittels Formularen zu

ihren Vorlieben oder anderen Informationen gefragt, handelt es sich um aktive Erhebung von Daten [Chi93]. Dies benötigt die Interaktion des Benutzers. Dadurch können Informationen gewonnen werden, welche anders nicht zur Verfügung stehen würden. Jedoch kann es sein, dass inkorrekte Daten angegeben werden. Passive Methoden hingegen sammeln relevante Informationen, ohne den Benutzer um Eingaben zu bitten [Chi93]. Dieser wird sozusagen beobachtet und aus Quellen wie den Webserver Log-Dateien oder besuchten Inhaltsseiten werden Daten für das Benutzerprofil extrahiert [Koc01]. Dies ist unaufdringlich, jedoch können nur Informationen beschafft werden, welche durch die Interaktion mit dem System entstehen [Chi93].

Diese beiden Arten der Informationsbeschaffung werden in vielen Fällen gemeinsam in einem Informationssystem verwendet. Um ein grundlegendes Benutzerprofil von einem Anwender zu erstellen, wenn dieser zum ersten Mal mit der Applikation kommuniziert, werden Informationen aktiv abgefragt [Chi93]. In einem System mit Bezug auf den Tourismus könnten dem Benutzer etwa unterschiedliche Unterkünfte präsentiert werden, welche sich grundlegend voneinander unterscheiden. Der Anwender wählt eine gewisse Anzahl an Objekten aus, die er präferiert. Weiters könnten Freizeitaktivitäten oder Informationen zum Familienstatus erfragt werden. Dies ermöglicht dem System, erste Daten zur Person zu speichern und diese eventuell einem Stereotyp zuzuordnen. Hiermit sind grundsätzliche Personalisierungsmethoden anwendbar. Nach dieser initialen Phase kann auf passive Techniken gewechselt werden, um den Anwender nicht unnötig bei der Interaktion mit dem System zu behindern. Durch die kontinuierliche Gewinnung von Informationen kann ein individuelles Benutzerprofil aufgebaut und die Adaption besser zugeschnitten werden.

3.2.2.2 Automatisch oder benutzerinitiiert

Bei der automatischen Methode zur Benutzerprofilverwaltung entscheidet die Applikation darüber, ob Daten über Anwender gesammelt und für ein Profil verwendet werden [Chi93]. Der Anwender hat keinen Einfluss darauf, ob er beobachtet wird und wann neu gewonnene Information eine Änderung des entsprechenden Benutzerprofiles zur Folge hat [Koc01]. Im Gegensatz dazu bestimmt dieser bei der benutzerinitiierten Technik selbst, zu welchem Zeitpunkt sein Profil aktualisiert wird [Koc01]. Weiters besteht die Möglichkeit, das Profil direkt selbst zu

verändern und somit erheblich zu beeinflussen, wie sich der Personalisierungsprozess auf die jeweilige Person auswirkt [Chi93].

Das Problem mit benutzerinitierten Mechanismen ist, dass Anwender sich um ihr Profil im Informationssystem kümmern müssen [Chi93]. Dies bedeutet, dass Zeit aufgewendet werden muss, um aktuelle Gegebenheiten aus der realen Welt im System widerzuspiegeln [Chi93]. Da die meisten Anwender nicht gewillt sind, diesen Aufwand auf sich zu nehmen, werden in den meisten Applikationen automatische Methoden eingesetzt. Diese sind zwar nicht so einfach zu implementieren und erfordern bessere Integration mit dem restlichen System (etwa das Aufzeichnen von besuchten Seiten, um anschließend daraus Information für das Profil zu extrahieren), erfordern jedoch kein Engagement des Anwenders. Dieses wird vor allem nicht aufgebracht, wenn keine hohe Wahrscheinlichkeit für eine längerfristige Benutzung der Applikation besteht [Chi93].

3.2.2.3 Direkt oder indirekt

Bei direkten Techniken werden gesammelte Benutzerinformationen ohne weitere Zusammenfassung oder andere drastische Weiterverarbeitung direkt für ein individuelles Benutzerprofil verwendet [Koc01]. Dies ermöglicht die präzise Verwaltung von unterschiedlichen Profilen für Anwender. Falls diese etwa eine Unterkunft in einem Tourismussystem betrachten und daraus Stichwörter extrahiert werden, können diese direkt für die jeweiligen Profile verwendet werden. Bei indirekt verwalteten Benutzerprofilen werden die gesammelten Informationen zwar ebenfalls verwendet, aber zuvor noch weiterverarbeitet [Koc01]. Oft werden durch Inferenzen Eigenschaften für ein Benutzerprofil bestimmt [Koc01]. Das Paradebeispiel für diese Form der Profilverwaltung sind Stereotypen. Basierend auf den direkt gesammelten Informationen werden ein oder mehrere Stereotypen aus einer fixen Anzahl von vorhandenen ausgewählt und dem Anwender zugewiesen [Chi93]. Dies bedeutet, dass die Personalisierung für Anwendergruppen und nicht für einzelne Personen erfolgt, basierend auf dem eruierten Stereotyp, welcher zu einer Person passt. Eine weitere Möglichkeit für indirekte Profile sind vorgefertigte Ziele, welche Personen in der Vergangenheit ausgeführt haben [Chi93]. Durch die Interaktion wird von gewonnenen Informationen auf das wahrscheinlichste Ziel aus dem Profil geschlossen, welches der Anwender erreichen möchte. Dadurch kann dieser bei der Bewältigung der Aufgabe unterstützt werden.

3.2.2.4 Logisch oder plausibel

Durch logische Informationsbeschaffung können nur Gegebenheiten in Benutzerprofilen dargestellt werden, welche durch die Befragung des Anwenders oder die Interaktion mit diesem ausgewertet wurden. Um von solchen gegebenen Daten auf andere schließen zu können, müsste die Möglichkeit zur Speicherung von Wahrscheinlichkeit in den Profilen vorhanden sein [Koc01]. Dadurch können dann auch plausible Informationen inferiert werden. Dies benötigt jedoch zusätzliche Systemkomponenten für die Berechnung und Kombination der Plausibilitätswerte sowie zur späteren Auswertung für die Personalisierung. Des Weiteren ist mehr Aufwand von Nöten, um ein Benutzerprofil mit Wahrscheinlichkeitswerten aktuell zu halten. Bei einer Änderung von nur einem oder weniger Werte müssen alle anderen verknüpften oder inferierten Werte neu berechnet werden [Chi93]. Der Vorteil solcher Methoden liegt jedoch darin, dass mehr Daten für die Personalisierung zur Verfügung stehen.

Falls für einen Anwender in einem Tourismussystem die logisch gewonnen Informationen vorliegen, dass dieser gerne schwimmt sowie sauniert, kann mit einem gewissen Grad inferiert werden, dass auch Massagen zu den Interessen zählen. Durch diese Information können bei der Suche oder Empfehlung Unterkünfte mit Massageeinrichtungen oder Thermen höher priorisiert werden.

3.2.2.5 Online oder offline

Bei Methoden, die als online klassifiziert sind, werden Daten direkt bei der Interaktion von Anwendern mit dem System gewonnen und für Benutzerprofile verwendet [Chi93]. Dies entspricht der Profilverwaltung im ursprünglichen und gedachten Sinn und bedeutet, dass die Person, über die Informationen beschafft werden, zum Zeitpunkt der Erfassung eine Sitzung mit der Applikation aufgebaut hat und Inhalte von dieser konsumiert [Chi93]. Techniken, die offline funktionieren, generieren Daten für Profile, wenn Anwender gerade nicht mit dem System verbunden sind [Koc01]. In diesem Fall werden meist gespeicherte Informationen von Datenbanken verarbeitet, um die Benutzerrepräsentationen im System zu aktualisieren [Koc01]. Dabei können die Datenquellen mit dem Programm selbst zusammenhängen oder aus externen Informationssystemen bestehen. Dadurch wird es ermöglicht, Profile mit Informationen anzureichern, welche nicht direkt durch die Webapplikation gewonnen werden können [Chi93]. Weiters besteht da-

durch die Möglichkeit, rechenintensive Vorgänge auf leistungsstarken Verbunden von Computern durchzuführen. Dies ist vor allem für Systeme interessant, welche aufgrund der enormen Anwenderzahlen und Datenmengen keine zeitnahen Ergebnisse für benötigte Berechnungen zur Verfügung stellen können. Ein Nachteil von solchen Mechanismen ist andererseits, dass die aufbereiteten Daten erst nach gewisser Zeit für die Personalisierung zur Verfügung stehen. In der Regel betrifft dies einen Anwender erst bei seinem nächsten Besuch des Websystems [Chi93].

Falls sich Anwender in Webapplikationen, welche Informationen zu Unterkünften zur Verfügung stellen, mit einem Authentifizierungsmechanismus von sozialen Medien registrieren und anmelden, erlaubt dies dem Informationssystem, etwa auf Freizeitaktivitäten zuzugreifen, welche von dem jeweiligen sozialen Netzwerk über die Person gesammelt wurden. In der Regel setzt dies eine einmalige Zustimmung des Anwenders voraus. Durch diesen Zugriff auf zusätzliche Informationen kann das Benutzerprofil angereichert beziehungsweise aktualisiert werden, sobald sich Daten in der externen Quelle ändern.

3.2.2.6 Serverseitig oder clientseitig

Durch serverseitige Informationsbeschaffung können nur Informationen für das Benutzerprofil gewonnen werden, welche vom System eruiert werden können oder explizit vom Benutzer zur Verfügung gestellt werden [Gau+07]. Darunter fallen auch externe Quellen wie soziale Medien. Falls ein Agent auf der Clientseite für die Identifikation eines Anwenders verwendet wird, ist dieser in der Regel auch für die Sammlung von Daten zuständig [Gau+07]. Dies benötigt Interaktion mit dem Anwender, weil dieser Software installieren muss. Weiters können auf Informationen des jeweiligen Computers zugegriffen und der Applikation zur Verfügung gestellt werden [Gau+07]. Dies wird von Anwendern nicht immer toleriert, weil die Möglichkeit besteht, sensitive Daten zu verwenden. Der Umstand, dass für solche Programme meist ein eigenes Protokoll verwendet wird, bedeutet auch, dass die Funktionsweise möglicherweise nicht gegeben ist [Gau+07]. Diese kann etwa durch eine Firewall unterbunden werden. In diesem Fall können keine Informationen für die Profilverwaltung gewonnen werden.

Aufgrund der gravierenden Nachteile von Agenten auf Seite der Clients werden fast ausschließlich serverseitige Techniken zur Beschaffung von Benutzerinformationen verwendet [Gau+07]. In der Regel sind die Daten, welche damit erfasst

werden können, für maßgeschneiderte Adaption ausreichend [Gau+07]. Vor allem durch die Einbindung externer Datenquellen gibt es keine Vorteile mehr für die Installation von Software durch den Anwender.

3.2.3 Profilverwaltung mit gesammelten Informationen

Um eine bestmögliche Adaption eines Informationssystems für Anwender zu ermöglichen, müssen Benutzerprofile nicht nur einmalig angelegt und mit Daten befüllt werden. Es erfordert weiters eine stetige Aktualisierung dieser Informationen über die Anwendereigenschaften und Präferenzen [Gau+07]. Ein auf diese Art verwaltetes Profil wird dynamisch genannt [Koc01]. Die gespeicherten Daten ändern sich ständig, basierend auf der Interaktion von Anwender und Applikation. Die Informationsgewinnung muss demnach kontinuierlich durchgeführt werden [Gau+07]. Die gesammelten Daten werden dazu verwendet, um ungenaue Eigenschaften in den Benutzerprofilen zu erkennen. Dies ist der Fall, sobald sich ein Systemanwender anders als vom System vorhergesagt verhält [Koc01]. Zu diesem Zeitpunkt müssen die Profildaten geändert werden, um die Person, welche das System verwendet, wieder akkurat zu repräsentieren [Gau+07]. Im Normalfall ist diese Erkennung und Korrektur der Informationen im Profil durch den Verwaltungsprozess implizit gegeben. Die verwendeten Techniken funktionieren in einer Art und Weise, sodass durch die Gewinnung und Speicherung neuer Information das Profil automatisch verändert wird [Gau+07].

Bei impliziter Sammlung von Stichwörtern, welche auf besuchten Seiten vorhanden sind, werden neue Stichwörter zum Beispiel automatisch zum Profil hinzugefügt. Durch das Ergänzen der Daten um Zeitstempel kann weiters implementiert werden, dass Wörter aus dem Profil entfernt werden, welche über einen gewissen Zeitraum nicht auf besuchten Seiten vorgekommen sind.

Die Auswahl der verwendeten Methoden zur Erstellung und Aktualisierung von Benutzerprofilen hängt stark von der verwendeten Profilart ab. Üblicherweise werden hierfür Techniken aus den Bereichen Informationsgewinnung und maschinelles Lernen eingesetzt [Gau+07]. Weiters können die Prozesse zur Verwaltung automatisch vom System vorgenommen oder dem Anwender beziehungsweise einem Domänenexperten überlassen werden. In den meisten Systemen wird die erstere Variante bevorzugt, weil sie keinen Aufwand für den Systemanwender mit sich bringt [Gau+07]. Dies erspart zeitintensive Anpassung des eigenen Profils.

Außerdem könnte diese Arbeit für manche Benutzer eine zu große technische Hürde bedeuten [Gau+07]. Diese würden demnach nicht von personalisierten Angeboten Gebrauch machen können. Die Profilaktualisierung kann noch immer über Feedback Mechanismen vom Anwender beeinflusst werden, sofern die Applikation diese implementiert [Gau+07]. Ein Beispiel für solch ein Feedback ist etwa die Bewertung von Unterkünften im System. Dadurch wird auf einfache Weise die Möglichkeit geschaffen, Präferenzen in die Profilierung einzubringen.

3.2.4 Zugriff auf Benutzerprofile zur Personalisierung

Um personalisierte Dienstleistungen innerhalb eines Informationssystems anbieten zu können, müssen die im Benutzerprofil gespeicherten Daten ausgewertet werden und basierend auf dem Ergebnis können unterschiedliche Maßnahmen zur Adaption getroffen werden [Gau+07]. Die einfachste Möglichkeit besteht darin, Einstellungen des Anwenders zu verwenden, um Entscheidungen über Seitendesign oder Ähnliches zu treffen. In diesem Fall besteht für die Applikation kein zusätzlicher Rechenaufwand. Werden jedoch die Daten aus dem Profil zur Adaption herangezogen, müssen diese mit einer klassifizierten Darstellung von Inhalten des Systems vergleichbar sein [Gau+07]. Durch unterschiedliche Methoden kann danach entschieden werden, ob gewisse Objekte für einen Anwender von Interesse sein könnten oder nicht. Um solch einen Vergleich anstellen zu können, müssen Profileigenschaften und Inhaltseigenschaften in einem vergleichbaren Format vorliegen [Gau+07]. Wenn etwa das Benutzerprofil aus Stichwörtern besteht, bietet es sich an, auch aus Inhaltsseiten Stichwörter zu extrahieren und dadurch das Dokument zu repräsentieren.

Die Relevanz von Objekten für Benutzer kann ebenfalls durch diverse Methoden bestimmt werden. Da in web-basierten Informationssystemen in der Regel unstrukturierte Textkonstrukte zur Informationsvermittlung verwendet werden, besteht die klassifizierte Darstellung oft aus Stichwörter oder zusammengefassten Konzepten [MSM07]. Der Prozess zur Repräsentation und Bestimmung des Übereinstimmungsgrades werden aufgrund der Wichtigkeit in Kapitel 4 separat beschrieben.

3.3 Profile zur Darstellung und Verwaltung von Benutzerinformation

Informationen über den Benutzer können unterschiedlich im System gespeichert werden. Die Art des Benutzerprofils unterscheidet sich auch in den Möglichkeiten, die Daten für die Personalisierung zu verwenden. Weiters ist nicht jede Darstellungsmöglichkeit für alle möglichen Applikationen optimal. Das Overlay Modell etwa wurde vorwiegend für adaptive Lernsysteme entwickelt und benötigt Domäneninformation, um danach von Anwendern zu speichern, inwiefern ihr Wissen mit dem gesamt verfügbaren Wissen übereinstimmt. Basierend darauf können dem Benutzer passende Inhalte präsentiert werden. Stichwörter hingegen eignen sich besser zur Darstellung von Anwenderinteressen in Systemen, bei denen keine Art von Fortschritt beziehungsweise keine Unterscheidung von bereits bekannten und noch unbekannten Inhalten notwendig ist. In diesem Abschnitt werden sechs unterschiedliche Möglichkeiten zur Speicherung und Verwaltung von Profilinformationen erläutert.

3.3.1 Das Overlay-Modell

In einem Overlay-Modell werden die Daten des Benutzerprofils als Teilmenge der gesamten zur Verfügung stehenden Information betrachtet (Overlay bedeutet zu Deutsch Überlagerung) [Koc01]. Diese Art der Benutzerprofile werden vor allem im Bereich der adaptiven Lernsysteme eingesetzt, weil sich damit gut Lernfortschritte einzelner Anwender abbilden lassen und die Domäneninformation bekannt ist [BM07]. Vorwiegend wird mit dieser Profilart also Wissen gespeichert.

3.3.1.1 Repräsentation von Information

Die Grundlage in einem Overlay-Modell bildet das Domänenwissen. Dieses wird in einzelne Elemente aufgeteilt, wobei hierfür etliche Namen gebräuchlich sind (etwa Gegenstände, Themen, Wissenselemente, Lernziele) [BM07]. Unabhängig von der Benennung bilden diese einzelnen Elemente einen Teil der gesamten im System verfügbaren Information, welche für Benutzer relevant ist [BM07]. Diese Teilstücke werden meist als Fragmente bezeichnet [BM07]. Die Aufteilung in Fragmente erfolgt üblicherweise von Experten des jeweiligen Gebietes. Dabei kann die

Granularität der einzelnen Fragmente frei bestimmt und somit an die Applikationsbedürfnisse angepasst werden [BM07]. In Lernsystemen etwa können einzelne Aufgaben zu Themenbereichen zusammengefasst werden. Dies ermöglicht die Zusammenfassung der Information in Konzepte [BM07].

Eine Möglichkeit zur Darstellung aller Fragmente in einem Vektor [BM07]. Dadurch ist es möglich, für einen Anwender zu bestimmen, welche Teilinformationen ihm bereits bekannt sind und welche noch erlernt werden müssen. Je nach Größe der Konzepte oder Fragmente ergibt sich ein mehr oder weniger fein-granulares Profil zur Bestimmung des Wissenstandes eines Benutzers. Diese Methode hat jedoch den Nachteil, dass keine Beziehungen zwischen den einzelnen Wissenselementen festgehalten werden können [BM07]. Bei einer großen Anzahl an Konzepten besteht bei dieser Technik nur noch eine sehr generelle Möglichkeit zur Personalisierung, weil nur zwischen bekannten und unbekannten Elementen unterschieden werden kann.

Durch entweder hierarchisch angeordnete oder beliebig verknüpfte Konzepte lassen sich Beziehungen zwischen diesen realisieren [BM07]. Im ersten Fall werden Fragmente mit gleichen generellem Informationsgehalt am selben Vaterfragment zusammengefasst [BM07]. In einem Lernsystem über Pflanzen beispielsweise können die Baumarten Fichte und Kiefer unter dem Fragment Nadelbäume, welches selbst ein Subelement von Bäumen ist, eingefügt werden. Dadurch entsteht eine Aufteilung in zusammengehörige Konzepte. Dieser Umstand kann dann bei der Adaption berücksichtigt werden, indem etwa zuerst nicht erlernte Konzepte aus einem Bereich präsentiert werden, in dem der Anwender bereits Wissen vorweisen kann. Die zweite Art, ein komplexes Netzwerk mit beliebigen Beziehungen zwischen Fragmenten aufzubauen, ist gerade für adaptive Lernsysteme besser geeignet [BM07]. Diese Darstellung wird vorwiegend verwendet, um eine sequenzielle Abarbeitung des Lernmaterials zu verwirklichen [BM07]. Durch solche Links können Voraussetzungen bestimmt werden, welche dem Anwender bekannt sein müssen, bevor ein neues Teilgebiet erforscht wird. Durch eine Kombination der beiden Möglichkeiten können zusätzlich Beziehungen zwischen zusammengefassten Konzepten hergestellt werden [BM07].

3.3.1.2 Darstellung des Anwenderwissens

Um nun im Benutzerprofil darzustellen, dass einem Anwender gewisse Konzepte geläufig sind, wird die gleiche Struktur als jene der Domäneninformation im Pro-

fil abgelegt. Durch unterschiedliche Kennzeichnung von bereits Erlerntem oder Bekanntem und noch nicht angeeignetem Wissen kann das Informationssystem passende Konzepte für die nächste Interaktion mit dem Benutzer auswählen [BM07].

Die einfachste und älteste Methode zur Unterscheidung zwischen bekannter und unbekannter Information ist die binäre Darstellung [BM07]. Hierbei wird keine Differenzierung hinsichtlich des Grades der Bekanntheit gemacht. Ein Fragment ist dem Anwender entweder geläufig oder nicht. Es besteht etwa keine Möglichkeit zum Rückschluss auf den Fortschritt beim Lernen eines Konzeptes [BM07]. Falls etwa Fragen im Lernsystem falsch beantwortet werden, kann ein Konzept nur von gelernt auf nicht gelernt gesetzt werden.

Durch eine gewichtete Darstellung des Bekanntheitsgrades von Informationen kann die Funktionalität erweitert werden. Dabei bestehen unterschiedliche Möglichkeiten für den Ausdruck, wie gut ein Konzept dem Anwender geläufig ist [BM07]. Die bekanntesten Methoden sind qualitative, numerische Einteilung sowie Speicherung eines Wahrscheinlichkeitswertes. Bei der qualitativen Bewertung gibt es festgelegte Werte, welche die Bekanntheit eines Fragments beschreiben (etwa wenig erlernt, mittelmässig erlernt, gut erlernt) [BM07]. Numerische Werte sind Zahlen in einem festgelegten Bereich [BM07]. Der Vorteil dieser Repräsentationsmethoden ist, dass bei Änderung des Lernstandes dieser Umstand mehr oder weniger gut im Benutzerprofil vermerkt werden kann. Weiters sind speziell bei der Verwendung von Wahrscheinlichkeitswerten Berechnungen des Wissens von verknüpften Konzepten möglich [BM07].

Um Aussagen über den Wissensstand eines Anwenders aus unterschiedlichen Quellen nicht zu vermischen, besteht die Möglichkeit, diese in separaten Repräsentationen der Domäneninformation zu speichern (in einem Schichtenmodell) [BM07]. Dies erlaubt die getrennte Verarbeitung von Daten aus unterschiedlichen Quellen (etwa Einschätzung durch den Benutzer selbst und Beobachtung durch Interaktion). Lediglich für die Personalisierung werden die einzelnen Schichten zusammen für Berechnungen verwendet [BM07].

Unabhängig davon, mit welcher Methode die Wissensinformation im Benutzerprofil festgehalten wird, technisch gesehen werden überwiegend Schlüssel-Wert Paare verwendet [BM07]. In diesen wird das jeweilige Konzept als Schlüssel verwendet und der Wert gibt den Grad der Bekanntheit an. Sollen einzelne Quellen

in Schichten ergänzt werden, können Schlüssel-Aspekt-Wert Triple angewendet werden [BM07].

3.3.1.3 Aktualiserung des Benutzerprofiles

In adaptiven Lernsystemen ist üblicherweise nur eine Änderung des Profils nötig, wenn eine Lerneinheit abgeschlossen ist. Zu diesem Zeitpunkt können die Wissensstände für die einzelnen Konzepte angepasst werden, um den Fortschritt des Benutzers zu vermerken. Hier kann etwa ein neues Konzept mit einem Wert belegt werden, falls dieses zum ersten Mal bearbeitet wurde [BM07]. Falls genauere Unterscheidung des Wissensgrades möglich sind, kann basierend auf der Interaktion der Wert erhöht oder erniedrigt beziehungsweise zu einem anderen qualitativen Wert geändert werden.

3.3.1.4 Verwendung zur Personalisierung

Je nach Komplexität des Profils können dem Anwender unterschiedliche Vorschläge zu Themen gemacht werden, welche als nächstes abgearbeitet werden sollen. Falls keine Beziehungen zwischen Konzepten vorhanden sind, kann ein noch nicht oder bisher unzureichendes Konzept präsentiert werden [BM07]. Bei der Verwendung von Hierarchien können zuerst Fragmente aus Bereichen vorgeschlagen werden, in denen der Anwender bereits tätigt war und Wissen gesammelt hat. Durch die Verwendung von Beziehungen zur Verwirklichung von Voraussetzungen kann die Auswahl noch weiter eingeschränkt werden und es besteht die Möglichkeit, gewissen Konzepte für den Benutzer unzugänglich zu machen, weil dieser damit eventuell dem Wissensstand zu urteilen überfordert wäre [BM07].

Dadurch, dass die verfügbare Information im System bekannt sein muss und das Benutzermodell als Teil dieses gesamten Domänenwissens ist, empfiehlt sich die Verwendung eines Overlay-Modells vor allem für adaptive Lernsysteme oder andere geschlossene Systeme, in denen Inhalte bekannt sind [BM07]. Diese Art von Profil lässt sich zwar auf Interessen von Anwender übertragen, jedoch bleibt die Einschränkung, dass die Informationen des Systems zu Konzepten zugeordnet werden müssen [BM07]. Da es sich bei Informationssystemen, welche auf Interessen der Benutzer basieren, zumeist um offene Informationsquellen handelt, müsste die Vielzahl an neu hinzukommenden Inhaltselemente entweder händisch oder

automatisch zu Fragmenten zugeteilt werden. Weiters müssten die Interessen der Anwender ebenfalls in Konzepte eingeteilt werden. Dies hat den Nachteil, dass keine beliebigen Eigenschaften ausgedrückt werden können, woran der Grad der Personalisierung leidet.

3.3.2 Bayessche Netze

In Benutzerprofilen für adaptive Informationssysteme besteht oft die Gegebenheit, dass Daten gespeichert werden sollen, welche Unwahrheiten beziehungsweise Wahrscheinlichkeiten entsprechen. Zusammen mit der Möglichkeit von Inferenzen kann auf diese Weise der beste Inhalt für Anwender basierend auf aktuell bekannten sowie berechneten Informationen zur Verfügung gestellt werden. Bayessche Netze ist die am meisten angewandte Technik zur Darstellung solcher Wahrscheinlichkeitswerte [Koc01]. Am Besten eignen sich solche Netze für geschlossene, adaptive Lernsysteme, aus dem selben Grund wie jener der auch für die Verwendung eines Overlay-Modells spricht [BM07]. Es müssen Informationen über die Daten der Zieldomäne vorliegen, welche von der Webapplikation verwendet werden [Koc01].

3.3.2.1 Repräsentation von Wahrscheinlichkeiten und deren Beziehungen

Bayessche Netze sind gerichtete, azyklische Graphen [Koc01]. Die Knoten in diesen entsprechen verschiedenen Benutzereigenschaften [Koc01]. In einem adaptiven Lernsystem sind dies wie beim Overlay-Modell Fragmente von Lerninformation in beliebiger Granularität. Diese Variablen können unterschiedliche Werte annehmen. Jedem Wert wird dabei eine Wahrscheinlichkeit zugewiesen [Koc01]. Im einfachsten Fall kann ein Knoten einen der beiden binären Zustände wahr oder falsch besitzen. Demnach muss für jeden dieser beider Zustände eine Wahrscheinlichkeit vorliegen, mit welcher er eintrifft [BM07]. Die Anzahl der unterschiedlichen Werte ist jedoch nicht auf zwei begrenzt. In Lernsystemen mit qualitativen Bewertungen von Lernthemen könnten etwa die drei Zustände schlecht erlernt, mittelmäßig erlernt und gut erlernt verwendet werden. Weiters sind Knoten mittels Kanten verbunden, falls diese in Beziehung stehen [Koc01]. Unabhängige Knoten sind nicht verbunden. Durch diese Möglichkeit der Beziehung können mittels Inferenzen Wahrscheinlichkeitswerte und somit Vermutungen über den Anwender berechnet werden beziehungsweise für eine gegebene Menge an Wer-

ten die Wahrscheinlichkeit, dass ein Benutzerprofil genau diesen Zustand besitzt [BM07].

Im Normalfall wird der Graph so aufgebaut, dass eine Ursache-Effekt Beziehung modelliert wird. Durch Gesetze der Wahrscheinlichkeitsrechnung ergibt sich, dass die Berechnung der Wahrscheinlichkeitsverteilung in diesem Fall durch alle Vaterknoten eindeutig bestimmt ist. Ein sehr einfaches Beispiel wäre etwa, dass durch die Beantwortung der Frage *Ist eine Fichte ein Nadelbaum?* auf das Erlernen des Themas Bäume geschlossen wird. In diesem Fall besteht das Netz aus den Knoten Bäume und Frage mit einer Kante, die von Bäume zu Frage verläuft. Diese Graphenstruktur wird auch als qualitatives Modell bezeichnet.

Um das Netz zu vervollständigen, müssen noch quantitative Informationen hinzugefügt werden. Für jeden Knoten werden die verschiedenen Zustände mit den zugehörigen Wahrscheinlichkeitswerten benötigt [BM07]. Eine passende, einfache Annahme ist, dass für beide Knoten ein binäres Zustandsmodell ausreicht. Das Thema Bäume kann entweder erlernt sein oder nicht. Die Frage, ob die Fichte ein Nadelbaum ist, kann richtig oder falsch beantwortet sein. Zur Vereinfachung werden die Werte erlernt und richtig mit durch 1 gekennzeichnet und nicht erlernt sowie falsch durch 0. Nun werden die folgenden Wahrscheinlichkeitswerte benötigt.

- $P(\text{Bäume} = 1) = 0.3$ gibt an, dass ein zufällig gewählter Anwender mit einer Wahrscheinlichkeit von 30% mit dem Thema Bäume vertraut ist. Demnach hat die Mehrheit von 70% kein Wissen über Bäume.

- $P(\text{Frage} = 1|\text{Bäume} = 1) = 0.98$ hält fest, dass Anwender die Frage zu 98% richtig beantworten, wenn sie mit Bäumen vertraut sind. Dies lässt einen Rest von 2%, dass obwohl das Thema Bäume erlernt ist, eine falsche Antwort auf die Frage gegeben wird.

- $P(\text{Frage} = 1|\text{Bäume} = 0) = 0.03$ ist die Wahrscheinlichkeit von 3%, die Frage, ohne Kenntnisse über Bäume zu besitzen, richtig zu beantworten. Dies bedeutet, dass ein Anwender die Frage in diesem Fall zu 97% falsch beantworten wird.

Für einen Knoten mit n Vaterknoten mit jeweils k Zuständen müssen für diesen Knoten k^n Wahrscheinlichkeitswerte angegeben werden. Diese beschreiben die Wahrscheinlichkeit für eine Zustand bei einer gegebenen Kombination von

Zuständen der Elternknoten [BM07]. Bei einer hohen Anzahl an Zuständen sowie einem stark verknüpfen Netz ist die Verwendung von Bayesschen Netzen nicht mehr vertretbar. Es müssen eine Unmenge an Werten zur Verfügung stehen und das System muss unzählige Werte neu berechnen, falls sich Zustände ändern [Koc01].

Nachdem das Netz vollständig definiert ist, können Inferenzen ausgeführt werden. Dabei gibt es zwei Richtungen, in welche Berechnungen durchgeführt werden können. Durch Diagnose kann von vorhandenen Beweisen auf mögliche Ursachen geschlossen werden [BM07]. Im Beispiel wäre dies der Fall, wenn berechnet wird, zu welcher Wahrscheinlichkeit das Thema Bäume geläufig ist, falls die Frage richtig beantwortet wird. Zum anderen können Vorhersagen getroffen werden [BM07]. Falls das Thema Bäume bekannt ist, besteht eine hohe Wahrscheinlichkeit, dass die Frage richtig beantwortet werden kann. Hierbei werden die Wahrscheinlichkeiten für eine gegebene Konfiguration von Variablenbelegungen basierend auf einer gegebenen Menge von Beweisen berechnet.

Der Aufbau des Bayesschen Netzes wird üblicherweise durch Experten im jeweiligen Informationsgebiet durchgeführt [Koc01]. Diese müssen die vorhandenen Domänendaten in sinnvolle Teile zerlegen, diese durch Beziehungen miteinander verknüpfen und anschließend die entsprechenden Wahrscheinlichkeiten festlegen. Alternative können vor allem diese konditionierten Wahrscheinlichkeitswerte aus empirischen Daten ermittelt werden oder auf generellen Theorien beruhen [Koc01]. Nichtsdestotrotz empfiehlt sich der Einsatz solcher Netze in erster Linie für geschlossene Systeme, in denen Domäneninformation bekannt ist und nicht zu häufig verändert wird.

3.3.2.2 Aktualisierung des Profils

Falls sich Eigenschaften von Anwendern ändern, wie beispielsweise in adaptiven Lernsystemen dem Benutzer ein Thema geläufiger ist als in der Vergangenheit, ändert sich der Zustand des entsprechenden Knotens im Benutzerprofil [BM07]. Dadurch ergibt sich eine mögliche Änderung der Zustände aller Kinderknoten [BM07]. Durch deren Neuberechnung ändern sich Wahrscheinlichkeitswerte, welche für Inferenzschritte herangezogen werden können. Dadurch entstehen eventuell andere Inhalte, welche für die jeweilige Person als nächstes von Interesse sind.

3.3.2.3 Adaption mit Bayesschen Netzen

Im Gegensatz zu adaptiven Lernsystemen, welche ein Overlay-Modell ohne Wahrscheinlichkeitswerten verwenden, können durch Bayessche Netze erweitere Auswahlverfahren für den nächsten zu präsentierenden Inhalt angewendet werden. Unter der Annahme, dass dem Anwender die Möglichkeit gegeben wird, einen Schwierigkeitsgrad für die zu erlernenden Inhalte festzulegen, können neben der üblichen Auswahl von Inhalten, die noch nicht vollständig oder gut genug erlernt sind, zusätzlich basierend auf den Wahrscheinlichkeiten, dass Fragen richtig oder falsch beantwortet werden, jene passend zum gewählten Schwierigkeitsgrad präsentiert werden [BM07]. Desto größer das Selbstvertrauen des Anwenders ist, desto kleiner kann die Wahrscheinlichkeit einer richtigen Beantwortung gewählt werden [BM07]. Hierbei sind keine zusätzlichen Daten im Profil zu speichern, denn die Unsicherheiten müssen dem System für Inferenzen zur Verfügung stehen.

Weiters lassen sich Kinderelemente basierend auf dem aktuellen Wissensstand des Anwenders in eine Abarbeitungsreihenfolge sortieren. Ein Beispiel wäre, die Knoten nach dem wahrscheinlichen Bekanntheitsgrad absteigend zu präsentieren [BM07]. Dies gibt dem Benutzern die Möglichkeit, zuerst Material zu lernen, in welchem er laut System schon zu einem gewissen Grad Wissen besitzt, wodurch eventuell die Motivation gesteigert wird.

Für die Verwendung in anderen adaptiven Websystemen gelten die selben Einschränkungen wie für ein Overlay-Modell [BM07]. Anwenderinteressen müssen etwa Konzepten zugeordnet werden. Dies verringert den Grad der Personalisierung. Weiters können nur schwer neue Interessen in das System eingefügt werden. Diese müssten händisch oder automatisch in Konzepte eingeteilt werden. Für geschlossene Systeme, welche auf Interessen basieren, in denen sich jene auch gut zu Konzepten zuordnen lassen, kann diese Profilierungstechnik sehr hilfreich sein. Es besteht dann die Möglichkeit, von bekannten Interessen eines Anwenders auf nicht bekannte zu schließen und entsprechende Inhalte zur Verfügung stellen.

3.3.3 Stichwort-basierte Profile

Eine der einfachsten und ältesten Formen von Benutzerprofilen ist jene basierend auf Stichwörtern. Eigenschaften eines Anwenders werden durch einzelne Schlagwörter repräsentiert und im Profil abgelegt [SD10]. Dies hat den Vorteil, dass

eine Verwaltung eines solchen Benutzerprofils relativ einfach ist und die benötigten Daten durch konventionelle Informationsgewinnung aus unstrukturierten Inhaltstexten extrahiert werden können [Gau+07]. Nachteile ergeben sich jedoch durch Besonderheiten vom Wortgebrauch in Sprachen sowie den Umstand, dass viele Wörter gesammelt werden müssen, um das Profil akkurat zu verwalten.

3.3.3.1 Speichern von Anwenderinteressen im Profil

Stichwörter werden als Vektoren zusammengefasst in einem Benutzerprofil abgelegt. Die einzelnen Einträge im Vektor entsprechen den Wörtern, welche Interessen repräsentieren [Gau+07]. Diese erhalten zusätzlich eine numerische Gewichtung. Dadurch lässt sich die Wichtigkeit für den jeweiligen Anwender festhalten [Gau+07]. Implizite Quellen für diese Information sind üblicherweise Webseiten, die der Anwender in der Vergangenheit besucht hat oder die Extraktion wird zum Zeitpunkt von Seitenaufrufen durchgeführt [Gau+07]. Alternativ können Benutzer explizit Stichwörter angeben, welche das Informationssystem in das Profil integriert. Durch explizites Feedback auf Webseiten, etwa durch Bewertung des Inhaltes mit Sternen, kann die Gewichtung der extrahierten Stichwörter aktiv beeinflusst werden [Gau+07]. Dadurch stehen dem Anwender Möglichkeiten offen, um die gesammelten Daten an seine Vorlieben anzupassen.

Bei der Sammlung von Stichwörtern basierend auf dem Inhalt von besuchten Seiten muss der unstrukturierte Text aufbereitet werden [Gau+07]. Dadurch werden die wichtigsten Wörter des Textes gefiltert sowie unwichtige ignoriert. Aus diesen übrig gebliebenen Wörtern werden Stammformen gebildet, wovon anschließend Duplikate entfernt werden. Am Schluss dieser Aufbereitung stehen eindeutig unterscheidbare, den Text repräsentierende Wörter zur Verfügung. Diese müssen gewichtet werden, bevor sie in den gewichteten Vektor im Benutzerprofil eingefügt werden. Die häufigste Methode hierzu ist die Verwendung der $df * idf$ Gewichtung [Gau+07]. Diese multipliziert das Vorkommen eines Stichworts im Dokument mit der Inverse der Anzahl der Dokumente, in deren Repräsentation dieses Wort vorkommt. Die Kollektion der Dokumente entspricht in diesem Fall allen Seiten, aus denen Daten für das Profil extrahiert werden [Gau+07]. Es besteht hier weiters die Möglichkeit, Wörter unterschiedlich stark zu gewichten. Zum Beispiel können Teile von Überschriften eine größere Bedeutung für den Anwender haben, weil eine solche Textpassage aussagekräftig für den gesamten Inhaltstext einer Seite ist [MSM07]. Oftmals wird die Einschränkung getroffen,

nur eine bestimmte Anzahl an höchst-gewichteten Termen in den Vektor des Benutzerprofils aufzunehmen [Gau+07]. Eine genauere Beschreibung des Prozesses für die konventionelle Informationsgewinnung aus strukturiertem Text findet sich in Kapitel 4.

Ein Beispiel basierend auf einem Tourismussystem ist folgender Vektor mit Interessen des Benutzers: *{Suite:* 0.8*, Doppelbett:* 0.76*, Golf:* 0.54*, Wandern:* 0.32*}*. Dieser Vektor beschreibt also, dass für die Person, welche dem Profil zugeordnet ist, eine Übernachtung in einer Suite von hohem Stellenwert ist, sowie ein vorhandenes Doppelbett im Zimmer. Weiters unternimmt sie gerne sportliche Freizeitaktivitäten wie Golf und, wenn auch mit niedrigerem Interesse, Wandern. Obwohl die Speicherung in einem Vektor den geringsten Aufwand bedeutet, birgt er jedoch das Problem, dass Interessengruppen des Anwenders vermischt werden [Gau+07]. Das bedeutet, dass bei der Auswertung des gesamten Vektors eine Charakteristik des Anwenders zwischen Zimmereigenschaften und sportlichen Aktivitäten liegt. Dies kann dazu führen, dass etwa bei der Suche nur Unterkünfte angezeigt werden, welche alle Kriterien erfüllen, die im Benutzerprofil festgelegt sind. Falls der Anwender jedoch dringend eine Suite für eine Nacht benötigt, ist ihm möglicherweise egal, dass kein Golfplatz in der Nähe liegt.

Um die Vermischung von Interessen im Benutzerprofil zu vermeiden, können Stichwörter thematisch zusammengefasst werden. Dies ergibt mehrere gewichtete Vektoren im Profil, wobei jeder die interessanten Eigenschaften für eine Interessengruppe speichert [Gau+07]. Durch Dienste wie WordNet kann die Zuordnung von Wörtern zu Kategorien erfolgen [Gau+07]. Die Aufteilung in multiple Vektoren kann weiters für zeitliche Unterscheidungen verwendet werden. In diesem Fall werden etwa gesammelte Informationen der aktuellen Sitzung in einem eigenen Vektor gehalten [Gau+07]. Daten gelangen in einen Vektor, welche langzeitliche Interessen darstellt, indem gewisse Voraussetzungen erfüllt werden, wie etwa explizites Feedback auf einer Seite. Das bereits genannte Beispiel lässt sich folgendermaßen in die Kategorien Unterkunftseigenschaften und Sportaktivitäten einteilen: *Unterkunftseigenschaften: {Suite:* 0.8*, Doppelbett:* 0.76*}, Sportaktivitäten: {Golf:* 0.54*, Wandern:* 0.32*}*.

Eine andere Variante, wichtige Teile von Inhaltsseiten für Profile zu speichern, sind n-Gramme. Diese extrahieren nicht nur einzelne Stichwörter aus Texten, sondern Phrasen von n Wörtern [Gau+07]. In diesem Fall werden neben der Wichtigkeit im Profil auch noch die Wahrscheinlichkeit gespeichert, zu welcher die Wörter in einem n-Gramm gemeinsam in einem Dokument vorkommen [Gau+07]. Bei der

Verwendung einzelner Wörter besteht das Problem, dass diese mehrere Bedeutungen haben können. Falls diese Ausprägungen in einem gemeinsamen Kontext vorkommen, nennt man dies Polysemie [Gau+07]. Ein besonders prominentes Beispiel hierfür ist der Begriff Bank. Dieser steht für ein Kreditunternehmen und dessen Gebäude. Durch die Erfassung von n-Grammen wird die Bedeutung eines Wortes durch umliegende Wörter eingeschränkt und sprachliche Probleme werden minimiert [Gau+07]. Ob die Speicherung von Wortsequenzen effektiver für die spätere Adaption ist, hängt stark vom verwendeten Vokabular im Informationssystem ab.

3.3.3.2 Aktualisierung von gewichteten Vektoren mit Stichwörtern

Wird für die Darstellung von Anwendereigenschaften ein einzelner Vektor mit Stichwörtern verwendet, können noch nicht vorhandene Wörter einfach eingefügt werden [Gau+07]. Bei bereits vorhandenen wird die Gewichtung erhöht, um zu vermitteln, dass der Benutzer an diesem mehr Interesse hat [Gau+07]. Das Problem hierbei ist, dass in diesem Vektor in kurzer Zeit ein sehr umfangreicher Inhalt angesammelt wird. Dadurch können keine genauen Unterscheidungen zwischen den einzelnen Charakteristiken vorgenommen werden. Werden multiple Vektoren eingesetzt, etwa getrennt nach Konzepten, wird die Aktualisierung gleichermaßen vorgenommen [Gau+07]. Es muss lediglich vor dem Einfügen oder Ändern eines Wortes ermittelt werden, zu welchem Vektor dieses zugehörig ist.

Um die Menge an unterschiedlichen Stichwörtern überschaubar zu halten, kann weiters überprüft werden, ob sich bereits ein ähnliches Wort im Benutzerprofil befindet. In diesem Fall kann dann der entsprechende Wert einfach verändert werden, anstatt weitere Wörter einzufügen, welche sich nur geringfügig voneinander unterscheiden [Gau+07]. Hierbei ist jedoch darauf zu achten, dass auch die gleichen Konzepte repräsentiert werden, weil sonst das Interesse an einem möglicherweise verloren geht, falls entsprechende Stichwörter zusammengefasst werden.

3.3.3.3 Adaption durch stichwort-basierte Profile

Um einem Anwender Objekte oder Seiten basierend auf seinen Interessen zur Verfügung stellen zu können, müssen diese auf die gleiche Weise repräsentiert werden,

wie bereits besuchte Inhalte im Profil gespeichert sind [Gau+07]. Dies bedeutet, dass ebenfalls ein Vektor von Stichwörtern erstellt werden muss. Im Fall einer Suche durch den Anwender können alle Ergebnisse, welche auf die Anfrage als Antwort in Frage kommen, weiterbearbeitet werden. Durch die Extraktion von Stichwörtern und das Erstellen eines gewichteten Vektors kann der Inhalt von Seiten mit dem Inhalt des Benutzerprofils verglichen werden [Gau+07]. Hierzu gibt es unterschiedliche Möglichkeiten, welche im Kapitel 4 genauer beschrieben sind. Eine der am häufigsten verwendeten Vergleichstechniken ist die Anwendung des Vektorraummodells. In diesem werden Vektoren in einem Raum dargestellt und die Ähnlichkeit durch die Lage berechnet. Durch die berechneten Ähnlichkeiten von Suchresultaten mit den Anwenderinteressen können die Ergebnisse entsprechend sortiert werden beziehungsweise durch eine Grenze festgelegt werden, welche Seiten dem Anwender überhaupt präsentiert werden [Mic+07]. Diese Vorgehensweise eignet sich auch für die Empfehlung von Objekten.

Bei der Verwendung von multiplen Vektoren im Benutzerprofil kann das System mit Einschränkungen erweitert werden [Gau+07]. Etwa ist es wichtiger, dass Interessen aus dem Bereich Unterkunftseigenschaften erfüllt sind. Sportaktivitäten hingegen sind zwar eine nette Zugabe für die Übernachtung, speziell bei kurzfristiger Buchung jedoch eher nebensächlich.

3.3.4 Generalisierung von Interessen mittels Konzepten

Durch die Verwendung von Konzepten kann die sehr spezifische Ausprägung von Benutzerprofilen basierend auf Stichworten generalisiert werden. Dies bedeutet, dass durch eine Eigenschaft im Profil ein größerer Bereich von Interessen abgedeckt ist und dadurch einem Anwender mehr Inhalte zur Verfügung gestellt werden können [Gau+07]. Durch eine hierarchische Aufteilung der Konzepte können die Möglichkeiten zur Personalisierung nochmals erweitert werden. Weiters entfallen die typischen Probleme von Vektoren mit Stichwörtern.

3.3.4.1 Repräsentation von Konzepten im Benutzerprofil

Konzepte sind Zusammenfassungen von Interessen zu Gruppen. Dabei ermöglicht eine Hierarchie, solche Eigenschaften in unterschiedlicher Granularität darzustellen [Gau+07]. Ein Beispiel dafür ist die Generalisierung von Laufen und Wandern

auf der untersten Ebene zu Sport, welches wiederum ein Kindkonzept von Freizeitaktivität sein kann. Die einfachste Variante zur Speicherung von Konzepten in Benutzerprofilen ist jedoch ein simpler Vektor. Dieser beinhaltet Konzepte von Interessen, ohne Beziehung zueinander, welche für eine Person von Bedeutung sind [Gau+07]. Es erfolgt eine Gewichtung entsprechend ihrer Relevanz. Dies entspricht im Wesentlichen dem gleichen Ansatz wie er auch bei Profilen, welche aus Stichwörtern basieren, verfolgt wird. Der Vorteil ist jedoch, dass die Probleme mit gleichen Wortbedeutungen eliminiert ist [Gau+07].

Eine bessere Verwendung von Konzepten ist deren hierarchische Darstellung in Bäumen. Hierbei werden interessante Konzepte für Anwender in entweder einer fixen oder dynamisch veränderbaren Anzahl an Ebenen im Profil abgelegt [Gau+07]. Die unterschiedlichen Knoten sind durch Vater-Kind Beziehungen verbunden. Den einzelnen Konzepten wird wiederum eine Gewichtung zugeteilt, welche den Grad der Wichtigkeit für den Benutzer darstellt. Diese Technik hat den Vorteil, dass unterschiedliche Stufen der Generalisierung gespeichert werden [Gau+07]. Dies ermöglichen eine umfangreichere Personalisierung. Um ein hierarchisches Profil an Konzepten aufbauen zu können, muss eine Referenzhierarchie vorliegen [Gau+07]. Die Daten im Profil entsprechen einer Teilmenge dieser Referenzhierarchie. Eine Möglichkeit besteht in der manuellen Erstellung der Hierarchie [Gau+07], was jedoch auf Grund der Vielzahl an unterschiedlichen Konzepte sehr aufwändig ist. Deshalb wird oft eine fertige Hierarchie als Referenz verwendet [Gau+07]. Beispiele sind das Yahoo! Directory oder das Open Directory Project [Gau+07]. Die Anzahl der Ebenen, welche aus der Referenz verwendet werden, hat starken Einfluss auf die Möglichkeiten zur Personalisierung. Durch weniger Ebenen entsteht ein generelles Benutzerprofil, welchem viele Inhaltsseiten zugeordnet werden können [Gau+07]. Dies hat jedoch den Nachteil, dass eventuell welche dabei sind, die für den Anwender nicht relevant sind. Mehr Ebenen bewirken ein spezifischeres Profil [Gau+07]. In diesem Fall können generelle, breitere Interessen verloren gehen.

Um ein Benutzerprofil aufbauen und später geeignete Inhalte zuordnen zu können, müssen Dokumente durch Konzepte repräsentiert werden [Gau+07]. Eine simple Lösung ist, die Dokumente zu klassifizieren bevor Profile angelegt werden. Dies bedeutet für den Profilierungsprozess, dass lediglich bei einem Seitenaufruf die zugeordneten Konzepte im Profil mit Werten versehen werden müssen [Gau+07]. Jedes neue Objekt im Informationssystem muss demnach mit Konzepten verknüpft werden [Gau+07]. Dadurch, dass die Konzeptzuordnung bereits

gegeben ist, kann das Benutzerprofil die gesamte Hierarchie beinhalten [Gau+07]. Alternativ kann die Klassifizierung in Konzepte automatisch zum Zeitpunkt des Seitenaufrufs durchgeführt werden [Gau+07]. In diesem Fall wird der unstrukturierte Text durch Konzepte dargestellt. Diese können dann in das Benutzerprofil aufgenommen werden. Um solche eine automatische Zuordnung zu ermöglichen, wird die Klassifizierungskomponente mit Trainingsdaten versorgt [Gau+07]. Das Open Dictionary Project bietet etwa Beispieltexte für erfasste Konzepte. Mithilfe dieser oder andere Daten kann das System trainiert werden. Auf Basis dieser Zuordnungsbeispiele werden die relevanten Objekte später mit Konzepten verknüpft [Gau+07]. Während bereits klassifizierte Dokumente sehr starr sind, kann die automatische Klassifizierung sehr leicht von Inhalten auf andere Texte übertragen werden, etwa Suchanfragen. Diese sind dem System nicht bekannt und können erst Einfluss auf das Profil nehmen, sobald der Anwender die Suchfunktion verwendet.

3.3.4.2 Aktualisierung der Konzepte in einem Profil

Bei der Verwendung eines gewichteten Vektors kann die Relevanz von bereits vorhandenen Konzepten einfach erhöht werden, sobald ein aufgerufenes Dokument mit diesem verknüpft ist. Neue Konzepte werden in den Vektor eingefügt. [Gau+07]

Falls eine Hierarchie im Profil repräsentiert wird, stehen mehr Möglichkeiten zur Anpassung der Wichtigkeit einzelner Konzepte zur Verfügung. Bei bereits klassifizierten Dokumenten, in denen der gesamte Baum bereits bekannt ist, müssen keine neuen Knoten eingefügt werden. Lediglich eine Aktualisierung der Gewichte muss vorgenommen werden [Gau+07]. Dies kann einfach realisiert werden, indem der Besuch von Seiten mit einem Konzept gespeichert wird. Dies ist auch für Systeme der Fall, bei denen die Klassifizierung automatisch beim Seitenaufruf erfolgt. Zusätzlich müssen jedoch Konzepte in die Hierarchie aufgenommen werden, falls diese im Profil noch nicht vorhanden sind. Je nach Implementierung können bei beiden Varianten weiters Vaterkonzepte von geänderten Konzeptknoten angepasst werden, um auch generellen Interessen eine höhere Gewichtung basierend auf der Bewertung speziellerer Vorlieben zukommen zu lassen [Gau+07]. Bei bereits indizierten Objekten kann dies auch schon durch die Zuordnung der Konzepte verwirklicht werden.

3.3.4.3 Personalisierung mittels Konzepten

Der Vorteil von Profilen basierend auf Konzepten ist, dass bei Empfehlungen oder Suche die Ergebnismenge sehr einfach erweitert werden kann, falls etwa zu wenig Resultate vorhanden sind. Im Unterschied zu Stichwörtern, wo man durch weglassen einer oder mehrerer relativ große Einbussen an Personalisierung hinnehmen muss, können durch die Beziehung von Konzepten in der Hierarchie ähnliche Konzepte für den Personalisierungsprozess herangezogen werden [Gau+07]. Je nach Granularität der Konzepthierarchie kann dadurch ein entsprechend gutes Ersatzergebnis erzielt werden. Falls etwa keine Unterkünfte auf eine Suchanfrage passen, in welcher das Kriterium Laufen für Sport in der Kategorie Freizeitaktivität angegeben ist, können ähnliche Unterkünfte im System gesucht werden, welche eventuell anderen Konzepten unterhalb von Sport zugeordnet sind. Vielleicht finden sich welche, in denen Wandern angegeben ist. Das Ergebnis würde sich durch ein weiteres Konzept zwischen Sport und den Kindern Laufen sowie Wandern verbessern lassen, weil sonst auch andere Sportarten als Kinder von Sport für die Erweiterung herangezogen werden können. Durch die Navigation über mehrere Väter kann die Suche so lange auf Kosten des Grades der Personalisierung erweitert werden, bis zumindest Ergebnisse auf eine Suchanfrage geliefert werden können.

3.3.5 Stereotypen als Abstraktion von Merkmalen

Stereotypen sind eine Zusammenfassung von Benutzermerkmalen, um eine Gruppe von Anwendern zu repräsentieren. Dies ermöglicht die Anwendung gleicher Personalisierungsmaßnahmen von Inhalten für mehrere Personen. Ein Benutzer ist dabei nur einem Stereotyp zugeordnet. Hierdurch werden keine individuellen Charaktere berücksichtigt. Diese Art von Benutzerprofil eignet sich vor allem in Zusammenhang mit anderen Varianten, um bereits unmittelbar nach Erstellung eines Profils personalisierte Angebote vermitteln zu können.

3.3.5.1 Informationsdarstellung für Stereotypen

Wie auch bei anderen Profilarten werden Merkmale, wie etwa Interessen, im Profil gespeichert, mit dem Unterschied, dass damit ein Stereotyp darstellt wird. Dies bedeutet, dass keine individuellen Benutzerprofile für Anwender verwaltet

werden [Koc01], sondern lediglich eine begrenzte Menge an Profilen zur Repräsentation der Stereotypen benötigt wird. Ein Anwender wird lediglich mit einem Typen verknüpft und erhält durch die Ausprägung des Gemeinschaftsprofils personalisierte Information zur Verfügung gestellt [Koc01]. Dadurch, dass es nur eine begrenzte Anzahl an Profilen gibt und Anwender zugeordnet werden, ist der Grad der Adaption eingeschränkt und alle Mitglieder der Gruppe eines Typs werden auf die gleiche Weise vom System behandelt [Gau+07]. In der Regel bedeutet die Anwendung dieser Profilierungstechnik weniger Aufwand für das Informationssystem sowie einen geringeren Speicherplatzbedarf für Profildaten.

Die Profile für Stereotypen müssen dem System bekannt sein, das bedeutet, sie müssen von Experten erstellt werden. Sie sind sehr stark von der verfügbaren Domäneninformation abhängig und auf ein Informationssystem zugeschnitten. Weiters müssen Trigger oder Regeln definiert werden, nach denen Anwender zu den Stereotypen zugeordnet werden [SD10]. Diese beziehen sich auf die Merkmale der Personen oder deren Interaktion mit der Applikation [SD10]. Falls etwa ein 35 Jahre alter Mann späte Eincheck- und frühe Auscheckzeiten bei Unterkünften bevorzugt sowie in den Präferenzen Einzelzimmer als primäre Vorliebe definiert hat, könnte das System auf einen Geschäftsreisenden schließen und den entsprechenden Stereotyp aktivieren. In diesem sind noch weitere typische Eigenschaften für diese Personengruppe hinterliegt, wie etwa ein vorhandener Internetzugang auf dem Zimmer, um arbeiten zu können. Auch wenn diese Eigenschaft für die Person nicht essentiell ist, wird ihm diese Charakteristik durch den Stereotyp zugewiesen und für die Personalisierung verwendet. Daraus ergibt sich das Problem, dass keine akkurate, individuelle Adaptierung der zur Verfügung stehenden Inhalte möglich ist [Gau+07].

3.3.5.2 Änderung des Stereotyps für einen Anwender

Bei Änderungen von Anwendereigenschaften werden diese neu evaluiert und entsprechend ein anderer Stereotyp zugewiesen [SD10]. Durch diesen Umstand müssen Regeln für die Änderung durchdacht sein [SD10]. Stereotypen sind in der Regel so unterteilt, dass sich diese sehr stark voneinander unterscheiden, so dass man mit einer geringen Menge ein möglichst großes Spektrum an Anwender abdecken kann. Deshalb muss die Gesamtheit der Präferenzen eines Anwenders sowie seine Charakteristik in die Entscheidung einbezogen werden. Wenn etwa eine Person, welche dem Stereotyp Geschäftsreisender zugeordnet ist, seine Präferenz

von Einzelzimmer zu Doppelzimmer ändert, rechtfertigt diese alleinige Umstellung keine Zuordnung zu einem anderen Stereotypen. Möglicherweise bevorzugt die Person ein größeres Bett. In welcher Weise sich die Vorzüge oder Eigenschaften des Anwenders ändern müssen, wird wie bei der initialen Zuweisung durch Trigger festgelegt [SD10]. Eine mögliche Vorgehensweise ist die gewichtete Zuordnung eines Anwenders zu allen möglichen Stereotypen. Trigger, welche mit einem Stereotyp verbunden sind, erhöhen den Grad der Zuordnung zu diesem und verringern jene zu anderen [SD10]. Der Benutzer wird zu jedem Zeitpunkt mit dem Stereotyp der höchsten Gewichtung verbunden. Nur wenn eine andere Ausprägung den aktuellen Wert des verwendeten Typs überschreitet, wird dieser zugewiesen [SD10]. Durch dieses System wird die Zuweisung eines unpassenden Typs minimiert, weil der Benutzer ständig auf alle möglichen Ausprägungen evaluiert wird [SD10].

3.3.5.3 Verwendung zur Adaption

Stereotypen eigenen sich in erster Linie für Informationssysteme, in welchen keine individuelle Adaption von Nöten ist [Gau+07]. Ein Beispiel hierfür ist die Anpassung der Darstellung basierend auf dem verwendeten Endgerät. Diese werden je nach Bildschirmeigenschaften in Kategorien eingeteilt. Es muss also nicht jedes Gerät separat behandelt werden. Weiters werden Stereotypen oft in der Anfangsphase der Verwaltung von neuen Profilen verwendet [BM07]. Der Vorteil hierbei ist die rasche Möglichkeit einer Personalisierung, wenn auch nicht sehr spezialisiert, bis genügend Daten für ein individuelles Profil gesammelt sind.

3.3.5.4 Mögliche Stereotypen für den Tourismus

Folgende Stereotypen sind für die Adaption in Tourismussystemen geeignet. Im Speziellen beschreiben sie typische Benutzerklassen für das System Tiscover[1], im Besitz von Tiscover GmbH, welche ein Tochterunternehmen von HRS[2] ist.

- Stark festgelegte Benutzer sind jene, bei denen bereits klare Vorstellungen über die Destination sowie den Typ der Unterkunft vorhanden sind. Sie buchen in der Regel sehr früh, mit vielen inkludierten Extras. Interesse be-

[1] http://www.tiscover.com/
[2] http://www.hrs.de/

steht an Unterstützung bei der Reiseplanung. Sie gehören der Altersgruppe ab 40 Jahren, sowie beiden Geschlechtern an.

- Unterkunfts-orientierte Personen legen großen Wert auf den Typ sowie die Lage der Unterkunft. Hierbei ist eine größere Region mehr bedeutend als ein einzelner Ort. Weiters haben mögliche Attraktionen und Angebote in der Nähe der Destination hohen Stellungswert. Der Preis spielt erst in einer späten Phase der Entscheidung eine Rolle. Die Anwender sind überwiegend weiblich und ebenfalls ab 40 Jahre alt.

- Empfehlungs-orientierte Anwender haben keine klaren Vorstellungen von einer Unterkunft oder dem Ziel. Sie buchen gerne spät und individuell. Empfehlungen sind für sie ausschlaggebend. Bei diesen werden viele Optionen offen gelassen. Sie sind überwiegend männlich und im Durchschnitt 39 Jahre alt.

- Geografisch-orientierte Personen legen Wert auf ein bestimmtes Land. Dabei ist der Ort sowie Typ der Unterkunft von geringerer Bedeutung. Sie bevorzugen fertig gebündelte Reisen und wählen innerhalb einer Destination die beste zur Verfügung stehende Unterkunft. Überwiegend sind diese Anwender weiblich sowie 39 Jahre alt oder jünger.

- Preis-orientierte Anwender haben in der Regel ein festgelegtes Budget. Das Ziel und die Art der Unterkunft ist weniger wichtig. Angebote sind diesen Personen sehr wichtig. Bei der Destination sind Freizeitaktivitäten bedeutend. Der Geschlechteranteil ist ausgeglichen mit leichter Mehrheit an Männern, das Alter 35 Jahre oder jünger.

- Aktivitäts-orientierte, unabhängige Urlauber sind an einem Land interessiert. Typ der Unterkunft sowie Preis folgen in der Liste der Prioritäten. Die Details der Unterkunft sind eher unwichtig. Sie legen großen Wert auf spezifizierte Aktivitäten. Sie sind überwiegend männlich. Zum Alter gibt es keine Angabe.

Anhand dieser Einteilung ist erkennbar, dass eine klare Abgrenzung von Anwendern in Klassen nicht möglich ist. Es gibt zum Beispiel zwei Klassen in denen das Alter vorwiegend über 40 Jahre ist. Eine Änderung diese Merkmales alleine kann also nicht ausschlaggebend für die Zuweisung des einen oder anderen Stereotyps sein. Deshalb ist es umso wichtiger, bei der Profilverwaltung alle möglichen Aus-

prägungen ständig in Betracht zu ziehen und falls möglich, so rasch wie möglich zu individuellen Profilen zu wechseln.

Weiters lässt sich durch die Eigenschaften in der Aufteilung erkennen, dass viele dieser Merkmale am Besten mittels Befragung nach der Registrierung vom Anwender ermittelt werden. Es ist etwa nur sehr schwer oder gar nicht möglich, durch die Interaktion mit dem System festzustellen, dass jemand erst sehr spät in der Entscheidungsfindung am Preis interessiert ist, vor allem wenn er diese Auswahl nicht bei den Suchen einschränkt.

3.3.6 Profile basierend auf semantischen Netzen

Semantische Netze sind nützlich, um die grundlegenden Techniken von stichwort- und konzeptbasierten Benutzerprofilen um die Möglichkeit zu erweitern, beliebige Beziehungen zwischen den einzelnen Eigenschaften im Profil zu erstellen [Gau+07]. Dadurch wird der Zusammenhang dieser innerhalb eines repräsentierten Dokuments verdeutlicht. Dies verringert das Problem der Unterscheidbarkeit von einzelnen Wörtern. Weiters können nicht nur Hierarchien, sondern auch zusammenhängende Konzepte im Profil abgelegt werden.

3.3.6.1 Verschiedene Darstellungen von Information in Profilen

Ein semantisches Netz ist ein ungerichteter Graph bestehend aus Knoten und Kanten. Die Knoten repräsentieren dabei Interessen der Anwender [Gau+07]. Diese können beliebig durch Kanten verknüpft sein [Gau+07]. Durch die Gewichtung von Interessen sowie deren Beziehungen werden die für eine Person wichtigen Elemente von anderen unterschieden [Gau+07]. Beziehungen entstehen zwischen extrahierten Stichwörtern oder Konzepten innerhalb eines Dokumentes. Durch diese Information soll ein Kontext für die gespeicherten Daten geschaffen werden, um etwa Bedeutungen von Wörtern eindeutig zuordnen zu können [Gau+07].

Die einfachste Variante, um repräsentative Information für ein Benutzerprofil zu speichern, ist die Extraktion von Stichwörtern aus Inhaltsseiten. Diese werden als Knoten in das Netz eingefügt und miteinander verbunden [Gau+07], was sozusagen einem erweiterten Vektor von Stichwörtern entspricht. Falls ein semantisches Netz einen Anwender im Profil darstellt, ist das mit einem einzigen Vektor an

Stichwörtern gleichzusetzen [Gau+07]. Auf die gleiche Weise wie bei Profilen basierend auf Konzepten, können Dokumente durch diese klassifiziert werden. Die Konzepte werden in Knoten des semantischen Netzes eingefügt und ebenfalls miteinander verbunden, falls sie im selben Dokument vorkommen [Gau+07]. Die Information, welche Wörter zu welchen Konzepten zugeordnet sind, kann durch Dienste wie WordNet eruiert werden, wodurch die Unterscheidbarkeit von Interessen weiter gesteigert werden soll [Gau+07]. Im Normalfall ist die Verwendung von Konzepten effektiver als jene von einfachen Wörtern. Jeder Knoten wird gewichtet, um den Grad des Interesses an dem jeweiligen Wort oder Konzept festzuhalten [Gau+07]. Die Beziehungen erhalten ebenfalls ein Gewicht, welches im weiteren Verlauf angibt, wie oft die beiden Wörter oder Konzepte gemeinsam in Dokumenten vorhanden sind [Gau+07].

Das Problem mit einem einzigen semantischen Netz bestehend aus Stichwörtern ist, dass wie bei einem Vektor mit Stichwörtern, keine akkuraten Aussagen über die Interessen des jeweiligen Anwenders gemacht werden können. Dies hat den Grund, weil sich verschiedene Bereiche von Interessen vermischen und keine Unterscheidung getätigt werden kann [Gau+07].

Es besteht weiters die Möglichkeit, Konzepte und Stichwörter gemeinsam zu verwenden. Stichwörter werden einem Konzept zugeordnet, indem die jeweiligen Knoten verbunden werden [Gau+07]. Weiters gibt es Beziehungen zwischen den Konzeptknoten, um deren Vorkommen in einem Dokument darzustellen [Gau+07]. Durch diese Kombination werden unterschiedliche Ebenen an Detailgehalt erreicht. Konzepte verwirklichen generellere Interessen, während die Stichwörter spezifische Vorlieben der Anwender aufzeigen. Auch in diesem Fall werden Knoten sowie Kanten gewichtet.

3.3.6.2 Aktualisierung eines Profils

Beim Einsatz von Knoten bestehend aus Stichwörtern werden repräsentative Wörter aus besuchten Inhaltsseiten extrahiert [Gau+07]. Diese werden anschließend in das Netz eingefügt, falls sie noch nicht vorhanden sind und mit den Knoten von anderen vorkommenden Wörtern verbunden [Gau+07]. Ist ein entsprechender Knoten bereits vorhanden, wird die Gewichtung erhöht. Weiters werden die Kanten zu ebenfalls im Dokument vorhandenen Wörtern stärker gewichtet [Gau+07]. Das gleiche Verfahren wird für Knoten bestehend aus Konzepten angewendet, mit dem Unterschied, dass Dokumente zuerst durch Konzepte repräsentiert werden

müssen [Gau+07]. Die Anpassung des Netzes geschieht auf die gleiche Weise. Diese Änderungen der Gewichte haben zur Folge, dass öfters extrahierte und zusammen vorkommende Stichwörter beziehungsweise Konzepte von größerer Bedeutung für die Personalisierung sind [Gau+07]. Andererseits werden Wörter oder Konzepte, welche nicht so oft aus Dokumenten gewonnen werden, mit der Zeit immer unwichtiger. Dies kann weiter verstärkt werden, indem zeit-basiert seit dem letzten Auftreten eine Verringerung der jeweiligen Werte erfolgt.

3.3.6.3 Verwendung zur Personalisierung

Je nach Aufbau des semantischen Netzes kann ein darauf basierendes Benutzerprofil auf die gleiche Weise verwendet werden, wie ein Profil bestehend aus Stichwörtern oder Konzepten. Beim Einsatz von Stichwörtern können Profile mit den repräsentativen Wörtern von Dokumenten verglichen werden. Die Ähnlichkeit ist ausschlaggebend für die Relevanz des Inhaltes für einen Benutzer. Bei der Darstellung von Inhalten durch Konzepte können die auf einer Webseite vorhandenen Konzepte ermittelt werden und ebenfalls mit den im Profil vorhandenen auf Ähnlichkeit überprüft werden. In diesen Fällen werden die Gewichte der Knoten für die Berechnung der Relevanz herangezogen. Es besteht kein Unterschied zur direkten Verwendung von Stichwörtern oder Konzepten in Benutzerprofilen.

Anders als bei den beiden anderen Varianten zur Darstellung von Informationen im Profil können bei semantischen Netzen jedoch auch die Gewichte der Kanten für die Ermittlung der Bedeutung eines Dokuments für den jeweiligen Anwender verwendet werden. Diese geben an, wie oft verknüpfte Stichwörter oder Konzepte miteinander in einem Dokument auftauchen [Gau+07]. Falls nun auf einer Webseite mehrere dieser miteinander verbundenen Wörter oder Konzepte vorhanden sind, welche im Profil mit hohen Gewichten miteinander verbunden sind, ist das entsprechende Dokument von großer Relevanz. Andere Webseiten, welche vielleicht durch reine Betrachtung der repräsentativen Wörter oder Konzepte eine hohe Ähnlichkeit aufweisen, sind nur von geringerer Bedeutung, falls die entsprechenden Objekte im Profil nicht durch stark gewichtete Kanten miteinander verbunden sind. Dies würde bedeuten, dass die einzelnen Inhalte eines Dokumentes nur getrennt voneinander, jedoch nicht gemeinsam als Inhalt für einen Benutzer relevant sind.

3.3.7 Kombinierte Verwendung zur Vermeidung von typischen Problemen

In vielen Fällen bietet sich eine gemeinsame Benutzung mehrere Profilierungstechniken an, um sofort nach der Registrierung einen gewissen Grad an Personalisierung bereitstellen zu können. Methoden, welche individuelle Benutzerprofile für Anwender erstellen, haben meist das Problem, dass nach der Registrierung eines neuen Benutzers eine gewisse Menge an Daten über die verknüpfte Person zur Verfügung stehen muss [BM07]. Nur mit ausreichender Information lassen sich akkurate Adaptionen durchführen. Dies bedeutet, dass bis zur Verfügbarkeit dieser Daten der Anwender selbst die angebotenen Inhalte erforschen muss. Erst durch den Besuch von Seiten oder die explizite Bewertung von Inhalten werden nützliche Informationen gesammelt.

Um bereits direkt nach der Registrierung grundlegende Adaption anwenden zu können, bietet sich die Zuordnung des Anwenders zu einem vordefiniertem Stereotyp [BM07]. Basierend auf diesem können Inhalte gefiltert oder empfohlen werden. Mit der Zeit kann dann das individuelle Profil aufgebaut werden. Nachdem genug gesammelte Information verfügbar ist, kann vom Stereotyp zum individuellen Benutzerprofil gewechselt werden [BM07].

Die Verwendung von Stereotypen ist deswegen optimal, weil sich sehr rasch Personen basierend auf deren Einstellungen und Merkmalen zuordnen lassen [BM07]. Eine gängige Methode ist die Verwendung von Daten wie dem Geschlecht sowie Alter. Diese sind in nahezu jedem System erfasst und stehen sofort zur Verfügung. Zusätzlich können durch eine kurze Befragung weitere nützliche Informationen gesammelt werden. Im Beispiel Tourismus könnten dem Benutzer unterschiedliche Unterkünfte präsentiert werden, aus denen der Benutzer das für ihn passende auswählt. Die Vorauswahl berücksichtigt die typischen Eigenschaften des entsprechenden Stereotyps. Durch diese Auswahl sollten genügend Daten vorliegen, um den neuen Benutzer einer Klasse zuordnen zu können.

Diese Kombination bedeutet für den Anwender sehr geringen Aufwand. Eine solcher Fragenkatalog ist in ein wenigen Minuten abgearbeitet. Danach kann ihm bereits Inhalt präsentiert werden, welcher für ihn relevant sein könnte. Durch die weitere Interaktion mit dem System werden individuelle Vorlieben sehr schnell erfassbar, weil Benutzer bei Empfehlungen oder Suchergebnissen sich Objekte meiste nur im Detail ansehen, falls diese ihr Interesse geweckt haben. Ab die-

sem Zeitpunkt können implizit Informationen über besuchte Seiten gesammelt werden.

3.3.8 Vergleich von Benutzerprofilen für den Einsatz in Tourismussystemen

In web-basierten Informationssystemen, welche im Tourismusbereich angesiedelt sind, können sich die darin präsentierten Inhalte sehr häufig ändern. Es kann zu jeder Zeit vorkommen, dass neue Unterkünfte zur Vermittlung an Anwender zur Verfügung stehen oder andere Übernachtungsmöglichkeiten nicht mehr zur Vermietung bestimmt sind. Dadurch sind Methoden zur Profilierung, welche eine stabile Informationsmenge voraussetzen, für die Verwendung in solchen Systemen ungeeignet. Dies würde bedeuten, dass bei jeder Änderung die verfügbaren Inhalte, speziell die neu hinzu gekommenen, manuell als Domänendaten aufbereitet werden müssen. Weiters müssen die Interessen und Eigenschaften der Anwender ebenfalls durch verfügbare Elemente aus dieser Informationsmenge repräsentiert werden. Dies funktioniert in adaptiven Lernsystemen gut, weil dort Lernziele und -einheiten definiert werden können. Die inhaltlichen Daten ändern sich nicht allzu häufig und der Fortschritt von Anwendern kann ebenfalls mit den selben Informationen ausgedrückt werden. Overlay-Modelle sowie bayessche Netze sind für diese Art des Einsatzes besser geeignet.

Alle anderen Arten für die Verwaltung von Benutzerprofilen eigenen sich für Informationssysteme mit mehr oder weniger stabiler Informationsmenge. Diese können für Webapplikationen zur Vermittlung von Unterkünften verwendet werden. Hierbei sind Profile basierend auf Stereotypen am wenigsten sinnvoll, weil sie nur ganze Gruppen von Anwendern repräsentieren können. Individuelle Eigenschaften und Bedürfnisse der einzelnen Benutzer gehen dabei verloren. Alle Personen, die einem Stereotyp zugeordnet sind, werden gleich behandelt. Der Vorteil dieser Technik ist jedoch, dass ein Anwender sehr schnell mit einer geringen zur Verfügung stehenden Menge an Informationen über ihn in eine der möglichen Gruppen eingeordnet werden kann. Aus diesem Grund eignen sich Stereotypen sehr gut für neue Anwender in einem System. Aufgrund mangelnder bekannter Eigenschaften können keine individuellen Profile angelegt werden. Nachdem jedoch genügend Daten über den Benutzer gesammelt wurden, kann dieser durch eine andere Methode im System individuell dargestellt werden.

Profile basierend auf Stichwörtern sind sehr einfach zu realisieren und für den Einsatz im Tourismusbereich gut geeignet. Hierzu sollten mehrere Vektoren von Wörtern basierend auf Konzepten verwendet werden, damit keine ungewollte Mischung von Interessen entsteht. Der Nachteil von sprachabhängigen Eigenschaften der Wörter und möglicher Polysemie kann je nach Anwendungsbereich als unbedeutend eingestuft werden. Die Wörter im Vokabular, welches in entsprechenden Inhalten von Webapplikationen zur Vermittlung von Unterkünften verwendet wird, sind ausreichend voneinander unterscheidbar. Bei Vektoren von Stichwörtern besteht jedoch nicht die Möglichkeit, auf verwandte Inhalte zurückzugreifen, von denen bekannt ist, dass sie ähnlich zu den Interessen eines Anwenders sind. Es kann also vorkommen, dass unpassende Inhalte präsentiert werden, falls keine Dokumente vorhanden sind, welche gut zu den Daten im Profil passen.

Dieser Nachteil kann durch die Verwendung von Konzepten, am besten in hierarchischer Struktur, vermieden werden. In diesem Fall können aus einer Subkategorie andere Elemente als ebenfalls relevant eingestuft werden (wie etwa Wandern statt Laufen in sportlichen Freizeitaktivitäten). Dies bedeutet, dass auch Inhalte für den Anwender als bedeutend angesehen werden, welche nicht exakt gleich zu dessen Interessen sind. Es werden jedoch Alternativen (etwa in Empfehlungssystemen oder bei Suchen) angeboten, welche ähnlich sind und nicht willkürliche andere Eigenschaften besitzen. Bei Profilen basierend auf Stichwörtern können so etwa alle möglichen Freizeitaktivitäten empfohlen werden, falls die anderen Voraussetzungen erfüllt sind. Bei Konzepten kann eingegrenzt werden, dass zumindest das übergeordnete Konzept gleich sein muss. Die Effektivität dieser Art der Profilverwaltung hängt natürlich stark von der Granularität der Hierarchie ab. Weiters müssen aus den Stichwörter, welche eine Dokument repräsentieren, Konzepte ermittelt werden. Dies bedeutet zusätzlichen Aufwand für die Applikation. Der Einsatz dieser Technik eignet sich zumindest ebenso gut wie jene basierend auf Stichwörtern.

Semantische Netze eigenen sich gleichermaßen wie die beiden zuvor genannten Arten zur Darstellung von Benutzerprofilen, je nachdem welche Inhalte darin gespeichert werden (Stichwörter oder Konzepte). Der Vorteil ist jedoch, dass die Relevanz von Inhalten besser bestimmt werden kann, indem auch die Kanten des Netzes in Betracht gezogen werden. Die Interessen und Eigenschaften von Benutzern können genauer dargestellt werden, indem zwischen gemeinsam auftretenden Stichwörtern oder Konzepten eine hohe Gewichtung verwendet wird. Dieser Umstand lässt sich bei der Bestimmung von Ähnlichkeiten zwischen dem

Benutzerprofil und Dokumenten ausnutzen. Vor allem bei sehr umfangreichen Informationsmengen kann der Aufwand für erforderliche Berechnungen jedoch höher sein, als bei anderen Methoden.

Kapitel 4

Klassifizierung von Dokumenten für personalisierte Angebote

Dieses Kapitel beschäftigt sich mit der repräsentativen Darstellung von Dokumenten (Webseiten) sowie dem Vergleich dieser mit Daten in einem Benutzerprofil oder einer Suchanfrage. Da es sich bei Webseiten in der Regel um unstrukturierten Text handelt, können Methoden aus dem Bereich der klassischen Informationsgewinnung eingesetzt werden [MSM07]. Erweiterungen ermöglichen eine bessere Verarbeitung von Eigenschaften, welche für HTML Dokumente spezifisch sind. Durch diese wird die semantische Bedeutung der einzelnen Komponenten in die Klassifizierung miteinbezogen. Unterschiedliche Modelle existieren für den Vergleich von stichwortbasierten Darstellungen von Dokumenten mit etwaigen Suchanfragen oder dem Benutzerprofil. Weiters gibt es spezielle Methoden für web-basierte Inhalte.

Für die Möglichkeit von Empfehlungen oder der Suche nach relevanten Inhaltsseiten basierend auf Suchanfragen müssen inhaltliche Seiten mit dem Benutzerprofil eines Anwenders oder den Suchtermen vergleichbar sein. Für bestmögliche Vergleiche werden Inhalte repräsentativ dargestellt. Dies kann zum Beispiel durch Stichwörter oder Konzepte erfolgen [MSM07]. Eine Webseite enthält überwiegend unstrukturierten Text, bei dem sehr viele unnötige Informationen vorhanden sind. Zu diesen zählen Stoppwörter. Weiters sind Wörter in unterschiedlichen Zeiten und Formen vorhanden, die jedoch im Grunde für die Informationsdarstellung das selbe Konzept repräsentieren. Aus diesen Gründen werden Dokumente aufbereitet und durch eine genormte Form repräsentiert [MSM07]. Dies erlaubt den Vergleich von Objekten, sowie deren Reihung nach der Relevanz basierend auf einer Suchanfrage.

Nachdem Dokumente erfolgreich klassifiziert wurden, können unterschiedliche Ansätze für die Informationgewinnung verwendet werden. Diese reichen von simplen boolschen Vergleichen bis hin zu komplexeren Modellen, welche genauere Aussagen über die Relevanz einzelner Dokumente treffen können. Bei diesem Schritt werden die Repräsentationen von Dokumenten mit Daten aus dem Benutzerprofil beziehungsweise Suchanfragen verglichen [MSM07]. Die berechnete Ähnlichkeit ist ausschlaggebend für die Relevanz des jeweiligen Objektes für die Einbeziehung in die Ergebnismenge der aktuellen Operation [MSM07].

4.1 Klassische Informationsgewinnung mittels Stichwörter

Die Informationsgewinnung basierend auf Stichwörtern ist bereits vor den Zeiten des Internets angewandt worden. Sie eignet sich jedoch gleichermaßen für die Klassifizierung und Bestimmung von Relevanz von Webseiten. Diese bestehen nämlich, wie normaler Text auch, zum Großteil aus unstrukturierten Wortkonstrukten. Spezielle Eigenschaften von HTML, wie Tags, können gesondert im Prozess behandelt werden [MSM07]. Bei dieser Methode repräsentieren Stichwörter ein gesamtes Dokument. Diese sind aus dem Text extrahiert und werden für den späteren Vergleich mit anderen Wörtern (etwa aus einer Suchanfrage) verwendet.

Im Wesentlichen besteht diese Art der Informationsbeschaffung aus zwei Schritten [MSM07].

- Dokumente mit Inhalten müssen durch Stichwörter dargestellt werden. Dazu müssen die Daten vorbereitet werden, so dass nur wichtige Wörter übrig bleiben. Weiters werden diesen Wörtern Gewichte zugewiesen, welche der Wichtigkeit dieser für die gesamte Dokumentensammlung entsprechen.

- Diese gewichteten Stichwörter werden mit gleichartig repräsentierten Stichwörtern verglichen. Diese können aus dem Benutzerprofil eines Anwenders oder aus den Termen einer Suchanfrage stammen. Dadurch kann die Relevanz des Dokuments bestimmt und entsprechend dem Anwender zur Verfügung gestellt werden.

4.1.1 Vorbereitung eines Dokumentes

In der ersten Phase werden Dokumente aufbereitet, so dass am Ende nur noch gewichtete Stichwörter als Darstellung übrig bleiben [MSM07]. Zu den Schritten zählen die Entfernung von unwichtigen Textelementen sowie die Reduzierung von unterschiedlichen Formen und Zeiten auf die Stammform. Speziell bei Webseiten müssen spezifische Konstrukte ausgenommen werden, weil diese keinerlei Relevanz für den Dokumenteninhalt haben. Dies sind zum Beispiel Tags zum Einbinden von CSS-Dokumenten[1] oder Script-Dateien [MSM07]. Anschließend werden die übrigen Wörter gewichtet, damit sie für Vergleiche verwendet werden können.

Im Folgenden werden die einzelnen Schritte zur Aufbereitung eines Dokuments erläutert. Für eine bessere Veranschaulichung wird folgender Beispieltext verwendet (siehe 4.1).

Listing 4.1: Beispiel zur Aufbereitung von Dokumenten

```
1 <html>
2   <head>
3    <link rel="stylesheet" href="style/layout.css" type="text/
        css">
4    <title>Unterkunft Alpen</title>
5   </head>
6
7   <body>
8     <p>Diese Unterkunft mit Einzelbett bietet bezaubernde
         Aussichten auf die direkt benachbarten Alpen.
         Ebenfalls enthalten: Dusche, Fernseher, Internet.</p>
9   </body>
10 </html>
```

4.1.1.1 Extraktion von Wörtern

In der ersten Phase wird eine Textsequenz in einzelne Wörter zerlegt. Dabei werden Großbuchstaben in Kleinbuchstaben umgewandelt [CMS09]. Satzzeichen werden aus dem Text entfernt [CMS09]. Weiters werden unwichtige Teile der Dokumentenstruktur entfernt [CMS09]. In HTML-Seiten sind dies zum Beispiel alle Tag-Elemente. Abgesehen davon soll nur Text innerhalb von Tags erhalten bleiben, welcher auch inhaltlich relevant ist [CMS09]. Durch diesen Schritt kann semantische Information verloren gehen, etwa durch die Umwandlung von Groß-

[1] Cascading Style Sheet

in Kleinbuchstaben. Aus diesem Grund muss dieser Prozess an die jeweilig Sprache und andere Gegebenheiten angepasst werden [CMS09]. Ein Beispiel für einen Tag mit unerwünschtem Inhalt ist *link*. Nach dem Anwenden dieses Schrittes bleibt im Beispiel nur noch folgender Text enthalten.

unterkunft alpen diese unterkunft mit einzelbett bietet bezaubernde aussichten auf die direkt benachbarten alpen ebenfalls enthalten dusche fernseher internet

4.1.1.2 Entfernung von Stoppwörtern

In der nächsten Phase werden nicht relevante Wörter aus dem Text entfernt. Diese bieten keinerlei Information bezüglich des Inhaltes und werden auch Stoppwörter genannt [MSM07]. Weiters werden solche Wörter häufig in Dokumenten verwendet. Aus diesem Grund wird durch die Speicherung unnötig Speicherplatz verschwendet. Beispiele hierfür sind Artikel und Pronomen [MSM07]. Die zu entfernenden Wörter variieren je nach Sprache. Durch die Verwendung einer Liste können diese sehr einfach definiert, sowie aus dem Text entfert werden [MSM07]. Im Beispiel ergeben sich folgende Wörter.

unterkunft alpen unterkunft einzelbett bietet bezaubernde aussichten direkt benachbarten alpen ebenfalls enthalten dusche fernseher internet

4.1.1.3 Stammformreduktion

Im nächsten Schritt werden alle verbleibenden Wörter auf ihre Stammform reduziert. Damit werden unterschiedliche Zeiten und Formen von Wörtern eliminiert, sowie Wörter mit unterschiedlicher Ausprägung, jedoch gleichem Stamm vereinheitlicht [MSM07]. Dieser Prozess ist ebenfalls sprachabhängig. Porter's Stemmer ist ein Beispiel für einen Algorithmus zur Stammformreduktion [MSM07]. Dieser kommt ohne Wortlisten aus und entfernt so lange bekannte Teile am Beginn sowie Ende eines Wortes, bis die Stammform erreicht ist [MSM07]. Dies passiert in mehreren Zyklen. Dabei können auch Stammformen entstehen, die nicht Teil der Sprache sind (etwa computer und computing werden zu comput) [MSM07]. Nach der Anwendung auf den Beispieltext ergibt sich folgender verbleibender Text.

unterkunft alp unterkunft einzelbett bietet bezaubernd aussicht direkt benachbart alp ebenfall enthalt dusch fernseh internet

Die folgenden Ersetzungen wurden vorgenommen.

- alpen zu alp
- bezaubernde zu bezaubernd
- aussichten zu aussicht
- benachbarten zu benachbart
- ebenfalls zu ebenfall
- enthalten zu enthalt
- dusche zu dusch
- fernesher zu fernseh

4.1.1.4 Gewichtung der Terme

Nach der Stammformreduktion kann das Dokument als Wortmenge dargestellt werden. Die Wörter entsprechen den Termen t_k des Dokuemntes d [MSM07]. Im Beispieltext verbleiben 13 unterscheidbare Terme. Dies bedeutet eine Repräsentation durch $d = t_1, t_2, \ldots, t_13$. Die Ausprägung für das Beispiel lautet wie folgt (4.1).

$$\begin{aligned} d = \{&unterkunft, alp, einzelbett, bietet, bezaubernd, aussicht, \\ &direkt, benachbart, ebenfall, enthalt, dusch, fernseh, internet\} \end{aligned} \tag{4.1}$$

Diesen Termen kann nun ein Gewicht w_k zugewiesen werden, welches die Relevanz eines solchen Terms aus der vorhandenen Dokumente D für ein einzelnes Dokument repräsentiert [MSM07]. Eine sehr einfache Gewichtung ist die Anzahl der Vorkommnisse eines Terms im Dokument [MSM07]. Dies wird als sogenannten Termfrequenz $TF(t_k, d)$ bezeichnet. Ein Vektor enthält alle repräsentativen Terme des Dokuments mit der Anzahl in der Form $TF(t_k, d) =$

$(t_1, w_1), (t_2.w_w), \ldots, (t_k, w_k)$. Die Termfrequenzen für den Beispieltext sind folgende (4.2).

$$\begin{aligned} TF(t_k, d) = \{&(unterkunft, 2), (alp, 2), (einzelbett, 1), (bietet, 1), (bezaubernd, 1), \\ &(aussicht, 1), (direkt, 1), (benachbart, 1), (ebenfall, 1), (enthalt, 1), \\ &(dusch, 1), (fernseh, 1), (internet, 1)\} \end{aligned} \tag{4.2}$$

Es ist zu sehen, dass die Terme *unterkunft* und *alp* stärker gewichtet werden, weil diese öfters im Dokument vorkommen. Weiters werden alle Terme aus der Dokumentensammlung in eine sogenannte Index-Term-Datenbank eingetragen [MSM07]. Diese lässt sich gut als Matrix darstellen. Hierbei werden alle verfügbaren Terme in den Zeilen sowie alle Dokumente in den Spalten festgehalten. Die Werte stellen die Gewichte des jeweiligen Terms in einem spezifischen Dokument dar [MSM07]. Es ist zu beachten, dass jedoch nicht jeder Term aus der Kollektion in einem Dokument vorkommt. Durch die Verwendung einer Matrix kann außerdem schnell auf alle Dokumente geschlossen werden, in denen ein gesuchter Term enthalten ist. Dieser Umstand wird auch als invertierter Index bezeichnet [CMS09]. Das Problem mit der einfachen Variante der Termfrequenz ist, dass folgende Eigenschaften bezüglich der Gewichtung gelten sollen [MSM07].

- Umso öfter ein Term in einem Dokument vorkommt, desto wichtiger ist dieser für die Charakterisierung des jeweilige Dokuments.

- In je mehr Dokumenten der Kollektion ein Term vorkommt, desto geringer ist die Wichtigkeit für das Vorkommen in einem einzelnen Dokument.

Die erste Aussage kann durch die Termfrequenz erfüllt werden. Je öfter ein Term im Dokument vorkommt, desto größer ist dessen Gewichtung. Für die Einhaltung der zweiten Grundlage wird jedoch eine weitere Information benötigt. Diese beschreibt, in wie vielen Dokumenten ein Term auftaucht. Basierend darauf kann die Gewichtung für einzelne Dokumente entsprechend gestärkt oder geschwächt werden. Diese Eigenschaft wird auch Dokumentenfrequenz genannt und mit $DF(t)$ gekennzeichnet [MSM07] [CMS09]. Hierbei spielt die Anzahl der Vorkommen in einzelnen Dokumenten keine Rolle. Sobald der Term mindestens einmal in einem Dokument vorkommt, wird dieser Wert inkrementiert [MSM07]. Ohne die Ver-

wendung dieses Wertes ist jeder Term in einer Dokumentenkollektion von gleich großer Bedeutung, egal in wie vielen Dokumenten er vorkommt [MSM07]. Vor allem bei Inhalten, welche ähnliche Konzepte beschreiben, scheinen einige Wörter wahrscheinlich in jedem Dokument auf, weil diese beschreibend für diese Konzepte sind. Solch ein Term ist etwa *Unterkunft* für Seiten, die unterschiedliche Übernachtungsmöglichkeiten beschreiben. Diese Wörter tragen weniger stark zur Relevanz eines Dokumentes bei [MSM07].

Basierend auf diesen beiden Informationen lassen sich folgende wichtige Gewichtungsverfahren spezifizieren.

- In der booleschen Gewichtung wird lediglich angegeben, ob ein Term in einem Dokument vorkommt oder nicht [MSM07]. Weder die Termfrequenz noch die Dokumentenfrequenz haben einen Einfluss auf das Gewicht eines Terms in einem Dokument [CMS09]. Ist der Term im Dokument vorhanden, wird ihm der boolesche Wert 1 in der Index-Term-Datenbank zugewiesen, andernfalls 0 (siehe 4.3) [MSM07] [CMS09].

$$\begin{aligned} w_i = 1 \Leftrightarrow t_i \in d_i \\ w_i = 0 \Leftrightarrow t_i \notin d_i \end{aligned} \tag{4.3}$$

- Bei der sogenannten $tf * idf$ Gewichtung ist diese proportional zur Termfrequenz eines Terms im Dokument und umgekehrt proportional zur Dokumentenfrequenz [MSM07] [CMS09]. Dies zieht beide genannten Grundsätze in Betracht. Je öfter ein Term im Dokument vorkommt, desto wichtiger ist er in diesem. Je öfter er jedoch in der gesamten Kollektion vorkommt, desto unwichtiger wird er für alle Dokumente, in denen er vorkommt. Das Gewicht wird folgendermaßen berechnet (siehe 4.4) [MSM07]. Hierbei entspricht $TF(t_i, d)$ der Termfrequenz in einem Dokument, $DF(t_i, d)$ der Dokumentenfrequenz und $|D|$ der Anzahl der Dokumente in der Sammlung.

$$w_i = TF(t_i, d) * log(\frac{|D|}{DF(t_i)}) \tag{4.4}$$

  Diese Art der Gewichtung ist aufgrund der einfachen, aber doch effektiven Berechnung weit verbreitet. Sie wird etwa im Vektorraummodell verwendet [MSM07].

- Weitere Beispiele für die Gewichtung von Termen sind Okapi BM25 sowie die Gewichtung basierend auf Entropien [MSM07]. Erstere Methode be-

rechnet die Relevanz eines Terms für ein Dokument basierend auf der Wahrscheinlichkeit, dass ein Term in einem relevanten sowie nicht-relevanten Dokument der Sammlung vorkommt. Zweitere Technik basiert auf Entropien von Wörtern in Dokumenten, basierend auf Ideen aus Informationstheorien.

4.1.2 Besondere Gewichtung für HTML-Dokumente

Dadurch, dass HTML-Dokumente eine gewisse Struktur aufgrund der Verwendung von Tags besitzen, kann dieser Umstand bei der Gewichtung miteinbezogen werden. Ein Tag-Parser kann verwendet werden, um nicht alle Tags aus dem Text einfach zu entfernen, sondern Inhalte von spezfischen Tags gesondert zu behandeln [MSM07]. Das beste Beispiel hierfür ist das *title* Tag. Im Titel eines HTML-Dokuments befinden sich üblicherweise einzelne Wörter, welche den gesamten Inhalt des Dokumentes kurz und prägnant beschreiben. Aus diesem Grund können diese Terme stärker gewichtet werden [MSM07]. Diese Verfahren kann auch für andere Tag-Typen verwendet werden. Diese können zum Beispiel in unterschiedliche Klassen eingeteilt werden. Jeder Klasse wird danach ein Faktor für die Gewichtung zugewiesen [MSM07]. Inhalte der entsprechenden Tags werden mit diesem Faktor multipliziert. Im oben angeführten Beispiel würden also die Terme *unterkunft* sowie *alp* noch stärker gewichtet werden, weil diese im Titel vorkommen.

4.2 Methoden zur Datengewinnung

Genauso wie es unterschiedliche Techniken zur Ermittlung der Relevanz von Termen für Dokumente gibt, existieren unterschiedliche Modelle für den Vergleich von Dokumenten mit Benutzerprofilen oder Suchanfragen. Diese verwenden für die Gewichtung eine passende Methode. Zum Beispiel wird im Fall vom Booleschen Modell auch die boolesche Gewichtung angewendet. Im Folgenden werden die gebräuchlichsten Arten für die Informationsgewinnung basierend auf der Darstellung von Dokumenten durch Stichwörter vorgestellt.

4.2.1 Boolesches Modell

Bei dieser Form der Informationsgewinnung werden Dokumente entsprechend der booleschen Gewichtung klassifiziert. Terme, welche in einem Dokument enthalten sind, werden für diese mit 1 bewertet, anderenfalls mit 0 [MSM07]. Eine Anfrage wird in Form eines booleschen Ausdrucks formuliert. Die einzelnen Terme können dabei etwa mit logischen *und* beziehungsweise *oder* verknüpft werden [CMS09]. Im Ergebnis der Abfrage finden sich danach nur jene Dokumente, die der Anfrage entsprechen. Dies bedeutet, dass die entsprechenden Terme der Abfrage auch in diesen Dokumenten vorhanden sein müssen [MSM07].

Als Beispiel sollen folgende Dokumente dienen, welche Inhalte zu unterschiedlichen Unterkünften beinhalten (siehe 4.5).

$$\begin{aligned}
\vec{d_1} &= \{(alp,1),(einzelzimmer,1),(doppelzimmer,0),(lauf,1),(wand,0)\} \\
\vec{d_2} &= \{(alp,1),(einzelzimmer,0),(doppelzimmer,1),(lauf,0),(wand,1)\} \\
\vec{d_3} &= \{(alp,1),(einzelzimmer,0),(doppelzimmer,1),(lauf,0),(wand,0)\} \\
\vec{d_4} &= \{(alp,1),(einzelzimmer,0),(doppelzimmer,1),(lauf,1),(wand,1)\}
\end{aligned} \tag{4.5}$$

Angenommen, ein Anwender ist an Unterkünften in den Alpen interessiert, wobei er gerne in einem Doppelzimmer übernachten würde und mindestens eine der Sportarten Laufen oder Wandern in der Nähe ausüben möchte. Diese Eigenschaften können etwa aus seinem Profil extrahiert werden oder durch eine Suchanfrage festgestellt werden. Als booleschen Ausdruck lässt sich die Abfrage q folgendermaßen beschreiben (siehe 4.6).

$$q = alp \wedge doppelzimmer \wedge (lauf \vee wand) \tag{4.6}$$

In Dokumenten der Resultate müssen also die Terme *alp* und *doppelbett* sowie mindestens eines der beiden Terme *lauf* und *wand* vorkommen. Durch die Evaluierung der Dokumentenrepräsentation ergibt sich, dass die Ergebnismenge die Dokumente d_2, d_4 enthält. Beide enthalten die Terme *alp* sowie *doppelbett*. Dokument 2 beschreibt weiters eine Unterkunft, in deren Nähe man Laufen kann, wohingegen in der Umgebung der Unterkunft aus Dokument 4 zusätzlich auch noch Wandermöglichkeiten bestehen. Dokument 1 beschreibt ein Einzelzimmer.

Dies entspricht nicht dem Ausdruck in der Abfrage. In Dokument 3 ist keine Beschreibung bezüglich Laufen oder Wandern vorhanden. Deshalb ist dieses ebenfalls nicht in der Ergebnismenge.

Der Nachteil bei diesem Modell der Informationsgewinnung ist, dass keine Reihung der Dokumente im Resultat erfolgen kann. Durch die Art der Klassifizierung sowie dem Vergleich mit einer Anfrage ergibt sich lediglich, ob ein Dokument in die Ergebnismenge aufgenommen wird oder nicht [MSM07]. Im gezeigten Beispiel erfüllt Dokument 4 beide Anforderungen an Freizeitaktivitäten. Dies wäre beispielsweise ein Grund, diesem Objekt eine stärkere Relevanz zuzuweisen. Weiters ist diese Methode nur schwer für die Erstellung von Abfragen aus Eingaben einer Suchmaske zu verwenden. Dies hat den Grund, dass Anwender üblicherweise nicht mit boolescher Algebra vertraut sind [MSM07]. Dies bedeutet im Endeffekt, dass entweder zu wenige oder zu viele Dokumente im Ergebnis vorhanden sind, je nachdem wie die Wörter miteinander für die Abfrage verknüpft werden [MSM07]. Werden diese etwa mit einem logischen *und* verbunden, würde nur d_4 im Ergebnis vorhanden sein. Bei der Verwendung von *oder* jedoch würden alle Dokumente vorkommen, weil jedes einzelne den Term *alp* enthält. Eine weitere Einschränkung ist, dass nur Dokumente zurückgeliefert werden, welche exakt den Vorgaben der Abfrage entsprechen. Dies hat etwa zur Folge, dass auch leere Ergebnismengen entstehen können [MSM07]. Bei der Verwendung von anderen Techniken können zumindest Objekte zur Verfügung gestellt werden, auch wenn diese nicht exakt zur Anfrage passen.

4.2.2 Das Vektorraum-Modell

Die Gewichtung der Terme in einem Dokument basiert beim Vektorraum-Modell auf der $tf * idf$ Formel [CMS09]. Anhand der berechneten Werte wird jedes Dokument als Vektor in einem m-dimensionalen Raum dargestellt, wobei m die Anzahl von Termen in allen Dokumenten der Kollektion beziffert [MSM07]. In diesem Raum entspricht also jeder Term einer Dimension. Die Gewichtung w_{ij} eines Terms t_i für ein Dokument d_j entspricht der Koordinate in der entsprechenden Dimension [MSM07]. Aufgrund der Verwendung der $tf * idf$ Formel repräsentiert ein Gewicht größer 0, dass der entsprechende Term im Dokument vorkommt, aber

nicht in allen Dokumenten der Sammlung enthalten ist [MSM07]. Ein Dokument wird also folgendermaßen dargestellt (siehe 4.7) [MSM07].

$$\vec{d_j} = w_{1j}, w_{2j}, \ldots, w_{mj} \tag{4.7}$$

Die Terme in einer Abfrage werden auf die gleiche Weise repräsentiert. Es entsteht also ebenfalls ein Vektor mit m Elementen [MSM07]. Eine Anfrage wird also wie ein Dokument aus der Kollektion behandelt und im selben Raum dargestellt [CMS09]. Dies ist durch folgende Gleichung (4.8) veranschaulicht [MSM07].

$$\vec{q} = w_{1q}, w_{2q}, \ldots, w_{mq} \tag{4.8}$$

Aufgrund dess, dass Anfrage sowie Dokumente der Kollektion als Vektoren im selben Raum dargestellt sind, kann relativ einfach auf Ähnlichkeit geprüft werden. Hierzu können unterschiedliche Formeln zur Berechnung verwendet werden. Eine der gebräuchlichsten ist die Kosinus-Ähnlichkeit. Durch diese wird der Kosinus des Winkels zwischen einem Dokumentenvektor $\vec{d_j}$ und dem Abfragevektor $\vec{q}$ berechnet [CMS09]. Umso kleiner der Winkel ist, desto größer ist die Übereinstimmung. Im folgenden wird diese Berechnung veranschaulicht (siehe 4.9) [MSM07].

$$sim(d_j, q) = cos(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| * |\vec{q}|} = \frac{\sum_{i=1}^{m} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{m}(w_{ij})^2 * \sum_{i=1}^{m}(w_{iq})^2}} \tag{4.9}$$

Das Ergebnis, also die Ähnlichkeit zwischen der Abfrage und dem Dokument, ergibt einen Wert zwischen 0 und 1. Um zu entscheiden, welche Dokumente in die Ergebnismenge aufgenommen werden, wird üblicherweise ein Grenzwert definiert [CMS09]. Alle Seiten mit einem geringeren Ähnlichkeitswert können aus dem Resultat weggelassen werden. Jene mit einem Wert über der Grenze werden in dieses eingefügt.

Ein Vorteil dieser Art der Informationsgewinnung ist, dass eine Reihung der zurückgelieferten Dokumente erfolgen kann. Dies ist der Fall, weil die Kosinus-Ähnlichkeit numerische Werte liefert [MSM07]. Weiters wird durch das Gewichtungsverfahren darauf geachtet, dass ein gewisses Maß an Qualität vorhanden ist.

Diese wird durch die Diskriminierung von Gewichten für Terme, welche in vielen oder allen Dokumenten enthalten sind, erreicht [MSM07]. Es müssen auch nicht alle Wörter der Abfrage in einem zurückgeliefertem Dokument vorhanden sein. Jene Inhalte, die besser auf die Anfrage passen, weisen eine größere Ähnlichkeit auf [MSM07]. Dies ermöglicht die Erweiterung der Resultatmenge im Fall von wenigen Dokumenten, die über dem Grenzwert liegen. Durch die Verringerung dieses wird eine größere Menge an Objekten zurückgeliefert. Diese entsprechen nur noch in geringerem Ausmaß den Anforderungen einer Suche oder eines Benutzerprofils. Im Allgemeinen ist es jedoch noch immer wichtiger, dem Anwender einige Objekte zur Auswahl anzuzeigen, anstatt im schlimmsten Fall einen Fehler auszugeben, wonach keine passenden Inhalte gefunden wurden.

Die folgende Berechnung bezieht sich auf das Beispiel aus dem booleschen Modell. Die zur Verfügung stehenden Dokumente haben die nachstehenden Termfrequenzen $TF(t_i, d_j)$ für die Terme $t_k = alp, einzelzimmer, doppelzimmer, lauf, wand$ (siehe Gleichung 4.10). Ein einfacher Schritt zur Normalisierung ist, die Termfrequenz-Werte durch die Anzahl an verfügbaren Termen im jeweiligen Dokument zu dividieren $(3, 3, 2, 4)$. Die normalisierten Werte sind in der Gleichung 4.11 aufgeführt.

$$\begin{aligned} \vec{d_1} &= \{1, 1, 0, 2, 0\} \\ \vec{d_2} &= \{3, 0, 1, 0, 3\} \\ \vec{d_3} &= \{1, 0, 1, 0, 0\} \\ \vec{d_4} &= \{1, 0, 1, 2, 1\} \end{aligned} \tag{4.10}$$

$$\begin{aligned} \vec{d_1} &= \{0.25, 0.25, 0.00, 0.50, 0.00\} \\ \vec{d_2} &= \{0.43, 0.00, 0.14, 0.00, 0.43\} \\ \vec{d_3} &= \{0.50, 0.00, 0.50, 0.00, 0.00\} \\ \vec{d_4} &= \{0.20, 0.00, 0.20, 0.40, 0.20\} \end{aligned} \tag{4.11}$$

Die Gewichtung erfolgt nach der $tf * idf$ Formel. Um etwa die Relevanz des Terms *wand* für das Dokument 2 zu ermitteln, werden folgende Werte eingesetzt

(4.12). Der Term kommt dreimal im Dokument vor. Weiters sind insgesamt vier Dokumente vorhanden, wobei in zwei davon der Term zu finden ist.

$$w_{52} = TF(t_5, d_2) * log(\frac{|D|}{DF(t_5)}) = 0.43 * log(\frac{4}{2}) = 0.13 \tag{4.12}$$

Daraus ergeben sich folgende Gewichtungen der Terme für die einzelnen Dokumente (4.13).

$$\begin{aligned} \vec{d_1} &= \{0.00, 0.15, 0.00, 0.15, 0.00\} \\ \vec{d_2} &= \{0.00, 0.00, 0.02, 0.00, 0.13\} \\ \vec{d_3} &= \{0.00, 0.00, 0.06, 0.00, 0.00\} \\ \vec{d_4} &= \{0.00, 0.00, 0.02, 0.12, 0.06\} \end{aligned} \tag{4.13}$$

Hier lässt sich sehr gut die Wirkung der $tf * idf$ Formel betrachten. Dadurch, dass der Term *alp* in allen Dokumenten vorkommt, hat dieser keine Relevanz bei der Gewichtung. Auf der anderen Seite kommt das Wort *Einzelbett* nur im Dokument 1 vor. Deshalb hat dieser Term in diesem Dokument eine sehr hohe Gewichtung.

Angenommen, ein Anwender ist an alpinen Unterkünften mit Doppelbettzimmer mit der Möglichkeit zum Wandern in der Nähe interessiert. Die dazugehörige Abfrage würde dann folgende Terme enthalten: *alp, doppelbett, wand*. Auch hier kann eine Gewichtung durch Termfrequenzen erfolgen. Durch die Annahme, dass alle Terme gleich wichtig sind, ergibt sich ein Wert von 1 für vorhandene und ein Wert von 0 für nicht vorhandene Terme aus der Gesamtmenge der Terme (siehe 4.14). Dieser Vektor wird ebenfalls normalisiert (mittels einer Division durch 3, weil die Anfrage aus 3 Termen besteht, siehe 4.15).

$$\vec{q} = \{1, 0, 1, 0, 1\} \tag{4.14}$$

$$\vec{q} = \{0.33, 0, 0.33, 0, 0.33\} \tag{4.15}$$

Nun kann die gewichtete Repräsentation der Abfrage mit jenen der Dokumente verglichen werden. Hierzu wird die Kosinus-Ähnlichkeit als Metrik verwendet. Es ergeben sich folgende Ähnlichkeitswerte (siehe 4.16).

$$\begin{aligned} sim(d_1, q) &= 0.00 \\ sim(d_2, q) &= 0.80 \\ sim(d_3, q) &= 0.71 \\ sim(d_4, q) &= 0.44 \end{aligned} \tag{4.16}$$

Wenn ein Grenzwert bei 0.5 festgelegt wird und eine Reihung nach Ähnlichkeit vorgenommen wird, erscheinen die Dokumente 3 und 2 in dieser Reihenfolge im Ergebnis. Dokument 2 hat die höchste Übereinstimmung mit der Abfrage. Alle abgefragten Terme sind in diesem Dokument enthalten. Weiters sind keine Terme darin enthalten, die nicht in der Anfrage vorkommen. Dies entspricht einer exakten Übereinstimmung. Eine hundertprozentige Übereinstimmung ist hierbei deshalb nicht gegeben, weil die Gewichte auf den normalisierten Termfrequenzen basieren. Diese Normalisierung lässt die Anzahl der vorhandenen Terme im jeweiligen Dokument in die Gewichtung einfließen. Dokument 1 erhält eine Ähnlichkeit von 0, weil lediglich das Wort *alp* mit der Abfrage übereinstimmt. Dieses kommt jedoch in jedem Dokument vor und hat somit keine Auswirkung auf die Berechnung.

4.2.3 Modelle mit besonderer Berücksichtigung für Dokumente im Internet

Auf Webseiten existieren üblicherweise Links zu anderen Seiten. Die Anzahl der ein- sowie ausgehenden Links kann hierbei von Seite zu Seite variieren [MRS+08]. Aufgrund des Grades der Vernetzung kann einer Seite eine Relevanz zugewiesen werden, welcher zusammen mit anders ermittelten Werten für die Bestimmung der Wichtigkeit verwendet werden kann [MRS+08]. Die Bewertung der Seiten aufgrund von vorhandenen Links hat keine direkte Beziehung zu Abfragen. Es ist lediglich ein Faktor, der bei der Reihung einen Einfluss nimmt. PageRank ist ein Verfahren, welches die Struktur von Links zwischen Inhalten im Web als Grundlage hat [MRS+08]. Dabei wird ein Anwender simuliert, welcher bei einer Startseite

beginnt, mit dem Internet zu interagieren. Er hat von diesem Ausgangspunkt nun die folgenden Möglichkeiten [MRS+08].

- Er klickt auf einen Link, welcher auf dieser Seite vorhanden ist. Die Wahrscheinlichkeit teilt sich gleichmäßig auf alle ausgehenden Links. Wird auf drei andere Seiten verwiesen, gelangt der Anwender wahrscheinlich zu einem Drittel zu einer dieser anderen Seiten. Während der Anwender weiteren Links folgt, werden einzelne Seiten öfters besucht als andere. Dies richtet sich nach der Anzahl der eingehenden Links auf einer Seite.

- Alternativ kann der Benutzer willkürlich zu einer anderen Seite springen, etwa durch Eingabe einer URL in die Adresszeile des Browsers. Dies ist auch der Fall, wenn eine Seite keine ausgehenden Links enthält. Die Wahrscheinlichkeit für eine solche Operation richtet sich nach der Anzahl an vorhandenen Seiten N und wird mit $1/N$ berechnet.

Zur Gewichtung der beiden unterschiedlichen Möglichkeiten gibt es einen weiteren Parameter a. Dieser bestimmt die Wahrscheinlichkeit, zu welcher der Benutzer zu einer beliebigen anderen Seite springt [MRS+08]. Der Wert $1 - a$ entspricht demnach der Wahrscheinlichkeit, dass er einem Link folgt. Dieser Wert wird im Vorhinein festgelegt [MRS+08].

Anschließend wird solange iterative berechnet, wie relevant einzelne Knoten im Netz sind, bis die Werte konvergieren und keine bedeutsame Veränderung mehr festgestellt werden kann [MRS+08]. Die dabei verwendete Formel findet sich in 4.17 [MSM07]. Zum Beginn können beliebige Werte für den PageRank angenommen werden, üblicherweise der gleiche für jede Seite [MRS+08]. Das Endergebnis ist unabhängig von diesen.

$$rank(p) = (1 - a) * \sum_{q:(q,p)\in E} \frac{rank(q)}{N_q} + \frac{a}{N} \tag{4.17}$$

Der erste Term beschreibt, dass alle eingehenden Links zur Berechnung herangezogen werden [MSM07]. Dabei wird der PageRank Wert der Quelle auf die aktuelle Seite übertragen. Dieser wird jedoch je nach vorhandenen ausgehenden Links an der Quelle geschwächt. Dies bedeutet, dass der Wert einer Quelle gleichmäßig auf alle Seiten aufgeteilt wird, welche von dieser zu erreichen sind. Die Summe dieser Einflüsse wird mit der Wahrscheinlichkeit multipliziert, dass die

aktuelle Seite tatsächlich über einen Link erreicht wurde [CMS09]. Der zweite Teil in der Formel gibt an, zu welchem Grad die Seite zufällig aufgerufen wird, ohne dabei von einem Link zu kommen [MSM07].

Der Algorithmus benötigt eine ständige Neuberechnung, um strukturelle Änderungen im Netz einzubeziehen. Mit steigender Anzahl an Knoten und Links wird diese Aufgabe sehr schnell zu zeitintensiv für die Ausführung in Echtzeit [MSM07].

4.2.4 Andere Modelle

Im erweiterten booleschen Modell besteht die Möglichkeit, Terme in Dokumenten sowie der Abfrage zu gewichten [Fox+92]. Dies geschieht üblicherweise durch die $tf * idf$ Formel. Die Abfrage bleibt dabei wie beim normalen booleschen Modell ein boolescher Ausdruck [Fox+92]. Im Unterschied dazu gibt es jedoch für jeden Operator eine Formel zur Berechnung der Ähnlichkeit zwischen der Anfrage und einem Dokument [Fox+92]. Durch diesen Umstand können die Dokumente im Ergebnis entsprechend deren Relevanz sortiert werden. Weiters bleiben die Möglichkeiten der Abfragegestaltung mittels unterschiedlicher Operatoren erhalten.

Im Wahrscheinlichkeitsmodell wird ermittelt, wie wahrscheinlich ein Dokument für eine Abfrage relevant ist. Die Ähnlichkeit ist hierbei so definiert, dass die Wahrscheinlichkeit für Relevanz umgekehrt proportional zu jener ist, die angibt, dass ein Dokument nicht für die Abfrage von Bedeutung ist [MSM07]. Dies Erfordert die Aufteilung der Dokumentensammlung in relevante und nicht relevante Objekte. Weiters werden Dokumente sowie Anfragen als boolesche Vektoren dargestellt [MSM07]. Es entsteht also keine Gewichtung der einzelnen Terme. Es besteht die Möglichkeit zur Sortierung nach Relevanz.

Semantische Netze, wie etwa WordNet, können verwendet werden, um extrahierte Stichwörter aus Dokumenten durch Konzepte zur repräsentieren. Diese werden danach gewichtet und zur Darstellung der Dokumente verwendet [MSM07]. In WordNet sind Wörter zu Konzepte zusammengefasst. Diese können wiederum miteinander in Beziehung stehen. Dadurch wird auch eine Hierarchie an Konzepten unterstützt. Durch solche können Dokumente beliebig nach Granularität klassifiziert werden beziehungsweise bei der Abfrage eine Generalisierung erfolgen, falls für spezifische Konzepte keine Resultate vorhanden sind [MSM07].

4.2.5 Optimierung der Performance

Um passende Inhalte zu einem Benutzerprofil oder einer Suchanfrage zu finden, müssen alle in Frage kommenden Seiten mit der Anfrage auf Ähnlichkeit geprüft werden. Entsprechend dem Ähnlichkeitsgrad ergibt sich, welche Webseiten in das Resultat aufgenommen werden. Mit steigender Anzahl an Inhalten steigt also auch die Anzahl der nötigen Vergleiche. Dies führt zu mehr Berechnungen und ist ab einer gewissen Größe der Dokumentenkollektion nicht mehr in Echtzeit zu erledigen [CMS09]. Aus diesem Grund können die Dokumente basierend auf ihren Ähnlichkeiten zu Clustern zusammengefasst werden.

Durch Clustering werden Dokumente, die sich inhaltlich ähnlich sind, zusammengefasst und durch ein einzelnes Dokument stellvertretend repräsentiert [CMS09]. Welche Dokumente zusammen gruppiert werden, ist üblicherweise durch deren Abstände zueinander definiert. Das stellvertretende Dokument wird auch Centroid genannt [CMS09]. Die Annahme bei der Verwendung von Clustern ist, dass alle Dokumente in einer Gruppe für eine Abfrage relevant sind, falls eines davon als relevant bestimmt wird [CMS09].

Dadurch, dass der Centroid geringsten Abstand zu allen anderen Dokumenten in der Gruppe hat, kann dieser für Vergleiche verwendet werden. Das bedeutet, dass lediglich die stellvertretenden Dokumente aus allen Clustern mit der Abfrage auf Ähnlichkeit geprüft werden [CMS09]. Aus den Ergebnissen kann eine bestimmte Anzahl der ähnlichsten Dokumenten ausgewählt werden. Nun ist die Ergebnismenge bereits auf wenige Cluster eingeschränkt. In einem zweiten Schritt wird die Anfrage mit allen Dokumenten aus den gewählten Gruppen verglichen, um die endgültige Ergebnismenge an relevanten Dokumenten festzustellen [CMS09]. Durch diesen Prozess lassen sich sehr schnell relevante Cluster finden, weil die Anzahl dieser weit geringer ist als die gesamte Menge an verfügbaren Dokumenten [CMS09]. Im Gegensatz dazu muss ohne Clustering jedes einzelne Dokument aus der Kollektion überprüft werden.

Kapitel 5

Methoden zur Personalisierung

In diesem Kapitel werden die unterschiedlichen Klassen der Personalisierung von web-basierten Informationssystemen erläutert. Dabei werden die unterschiedlichen Arten erklärt. Anschließend werden die wichtigsten Methoden zur Adaption genauer vorgestellt. Hierbei sind Techniken aus allen Klassen vertreten. Im speziellen werden personalisierte Suchen, Empfehlungssysteme, Adaptionen der Navigation sowie Möglichkeiten zur unterschiedlichen Darstellung von Inhalten behandelt.

5.1 Klassen der Adaptierung

Die Anpassung von Webapplikationen auf die persönlichen Bedürfnisse der Anwender kann in unterschiedliche Klassen eingeteilt werden. Diese entsprechen den unterschiedlichen Ebenen, in denen die Information aufbereitet und dem Benutzer präsentiert wird. Grundsätzlich können Techniken in die folgenden drei Bereiche eingeteilt werden [KKP01].

- Adaption des Inhaltes, welcher einem Anwender zur Verfügung gestellt wird. Dies erfordert in der Regel eine Änderung oder Anreicherung der zu übermittelnden Information.

- Änderung der Darstellung von Inhalten. Diese Art der Adaption beschäftigt sich mit verschiedenen Möglichkeiten, Informationen einer Person zu vermitteln, ohne dabei den Inhalt dieser Informationen selbst zu ändern.

- Personalisierung der Struktur in der Webanwendung. Durch solche Techniken wird einem Anwender erlaubt, effizient zwischen Seiten zu navigieren oder einen Überblick über wichtige Informationen gesammelt auf einer Webseite zu erhalten.

Hierbei stellt die erste Kategorie, also die Darstellung sowie Anreicherung des eigentlichen Inhaltes, wohl die wichtigste Klasse der Personalisierung dar. Durch diese wird ein auf den Anwender zugeschnittenes Angebot zur Verfügung gestellt. Die dabei gefilterten Informationen werden vom System für den Benutzer als relevant empfunden. Seitdem jedoch immer häufiger mobile Geräte zur Darstellung von Webinhalten verwendet werden, gewinnt auch die geänderte Darstellung basierend auf Kriterien wie Bildschirmgrößen an Bedeutung. Dadurch wird ein optimales Erlebnis bei der Interaktion mit der Webapplikation gewährleistet.

5.1.1 Personalisierung des Inhaltes

Unter der Adaption des Inhaltes versteht man all jene Methoden, welche direkt mit den zur Verfügung gestellten Informationen eines Informationssystems zusammenhängen [KKP01]. Dabei wird entschieden, welche Daten für einen Benutzer von Relevanz sind. Somit wird versucht, von der Masse an verfügbaren Informationen nur eine Teilmenge zu präsentieren [KKP01]. Weiters fallen unter diese Kategorie Techniken, welche den eigentlichen Inhalt mit zusätzlichen Daten anreichern, um etwa dem Benutzer ein besseres Verständnis über gewisse Konzepte zu verschaffen [KKP01]. Durch die Verwendung von Adaptionen dieser Gruppe sollen Anwender gezielt gesuchte Information in kürzester Zeit finden, ohne dabei möglichst lange mit Seiten interagieren zu müssen, welche sie gar nicht interessieren. Dadurch soll einerseits sichergestellt werden, dass Personen mit dem Angebot einer Webapplikation zufrieden sind und diese in Zukunft wieder in Anspruch genommen wird. Auf der Seite des Betreibers führt dies andererseits dazu, dass potentiell mehr geschäftliche Transaktionen abgeschlossen werden.

Als Beispiel für den Tourismus kann die Empfehlung von Unterkünften in entsprechenden Informationssystemen genannt werden. Dabei werden einem Anwender Hotels oder Ferienwohnungen vorgeschlagen, an denen er interessiert sein könnte. Die Entscheidung, welche Objekte präsentiert werden, erfolgt durch den Abgleich der klassifizierten Inhalte mit dem Benutzerprofil des Anwenders. Dadurch werden passende Inhalte zur Verfügung gestellt, ohne dass danach gesucht werden

muss. Von speziellem Interesse ist hierbei auch der Kontext einer aktuellen Sitzung, vor allem der Aufenthaltsort. Durch diesen lassen sich Ergebnisse auf einen gewissen Umkreis einschränken.

Eine Art der Informationsanreicherung ist beispielsweise die Darstellung eines Unterkunftsortes auf einer Karte. In dieser können dann auch für den Benutzer relevante Standorte angezeigt werden. Dabei können an diesen etwa Freizeitaktivitäten ausgeübt werden, welche im Profil des Anwenders als Vorlieben gekennzeichnet sind. Damit wird signalisiert, dass die Unterkunft für die Bedürfnisse der jeweiligen Person optimal ist, weil viele Möglichkeiten zur Verfügung stehen. Weiters wird visuell dargestellt, dass all diese Aktivitäten in unmittelbarer Nähe vorhanden sind.

Folgende Techniken zählen zur Personalisierung des Inhaltes [KKP01].

- Suchen können personalisiert werden, indem die Ergebnisse auf das Benutzerprofil von Anwendern zugeschnitten werden. Hierdurch werden relevante Resultate in der Sortierung an bessere Positionen gereiht.

- Empfehlungssysteme können verwendet werden, um dem Anwender relevante Inhalte zur Verfügung zu stellen, ohne dass dieser explizit danach suchen muss.

- Die Inhalte können einem Anwender in optimierter Form präsentiert werden, so dass dieser genau jene Informationen erhält, welche für ihn wichtig sind. Dies wird durch Weglassen von unnützen Daten beziehungsweise Anreicherung durch zusätzlichen Informationen erreicht. Weiters kann die Form der Darstellung so geändert werden, dass bestimmte Bereiche einer Seite dominanter erscheinen.

5.1.2 Anpassung der Darstellung von Inhalten

Durch die Anpassung der Darstellung von Inhalten lassen sich diese an den jeweiligen Kontext anpassen beziehungsweise kann damit Rücksicht auf die individuellen Charakteristiken von Anwendern genommen werden. Im Unterschied zur zuvor beschriebenen Anpassung der Präsentation soll bei diesen Techniken weder eine Anreicherung noch ein Verlust von übermittelter Informationen erfolgen [KKP01]. Dies bedeutet, dass lediglich das Medium geändert wird, in welchem

die Daten dem Benutzer vermittelt werden. Eine textuelle Präsentation einer Grafik fällt etwa in diese Gruppe der Anpassungen. Das Weglassen des Bildes ohne Ersetzung durch eine andere Form ist hingegen der Adaption des Inhaltes zuzuordnen. In vielen Informationssystemen werden Anpassungen dieser Klasse zusammen mit inhaltlichen Adaptionen verwendet, um ein optimales Erlebnis für den Anwender zu gewährleisten.

Die Entscheidung über die Personalisierung der Darstellungsform basiert meist auf Einstellungen von Anwendern [KKP01]. Diese können üblicherweise direkt nach der Registrierung festgelegt werden und müssen über die Zeit nicht verändert werden. Dies ist der Fall, weil sich die Eigenschaften von Personen, auf denen diese Einstellungen basieren, nicht oder nur über sehr lange Zeit verändern. Visuelle Personen, welche eine bildliche Darstellung von Information bevorzugen, werden diese Charakteristik auch in einigen Jahren noch besitzen. Ein anderer wichtiger Faktor für diese Art der Personalisierung sind die Ausprägung von speziellen Bedürfnissen in Personen [KKP01]. Für eine blinde Person ist es zum Beispiel essentiell, dass an der Stelle von Bildern eine textuelle Präsentation des Inhaltes vorhanden ist. Diese kann durch spezielle Software in Sprache umgesetzt und somit dem Anwender vermittelt werden.

Der aktuelle Kontext einer Sitzung zwischen Endanwender und Webapplikation kann außerdem zur Anwendung der Adaption von Darstellungsformen führen [KKP01]. Diese Eigenschaften werden nicht fix im Benutzerprofil abgelegt, sondern für jede Sitzung einzeln evaluiert. Unter Kontext fallen etwa das verwendete Endgerät oder die zur Verfügung stehende Bandbreite. Videos können beispielsweise durch repräsentative Bilder ersetzt werden, falls eine zu geringe Übertragungsrate festgestellt wird. Die gesamte Webseite kann in ein freundliches Format für Mobilgeräte transformiert werden, wenn der Anwender auf einem solchen Gerät mit entsprechend kleinem Bildschirm interagiert.

Die bereits genannten Beispiele lassen sich auch in Applikationen für den Tourismus einsetzen. So lässt sich etwa ein Video einer Unterkunft durch eine Serie von Bildern oder einen geeigneten beschreibenden Text ersetzen.

5.1.3 Struktur-basierte Adaptierung

Die Klasse der struktur-basierten Personalisierung fasst alle Methoden zusammen, mit welchen die Vernetzung einzelner Seiten durch Links verändert wird

[KKP01]. Weiters zählen hierzu unterschiedliche Darstellung der Navigationselemente dem Anwender gegenüber [KKP01]. Durch diese Anpassungen werden für den Benutzer relevante Seiten, welche von der aktuellen erreichbar sind, hervorgehoben. Weiters kann er davon abgehalten werden, Inhalte zu betrachten, die ihn nicht interessieren. Die Bedienung, vor allem die effiziente Navigation, kann durch das gezielte verknüpfen von häufig aufgerufenen Seiten vereinfacht werden.

Eine implizite Verwendung von Techniken dieser Klasse findet sich in webbasierten Systemen vor allem auch bei der Anwendung von inhaltlicher Personalisierung. Durch die Filterung von Inhalten basiert auf Suchen oder Empfehlungssystemen entsteht eine adaptierte Sortierung der jeweiligen Objekte nach Relevanz für die entsprechende Operation. Bei der Darstellung dieser Inhalte wird zumeist mittels eines Links auf die Detailseite verwiesen, auf welcher mehr Informationen zum jeweiligen Objekt zu finden sind. Dadurch, dass sich die Reihung bei jeder Suchanfrage ändern kann und nicht immer die selben Ergebnisse erreicht werden, impliziert dies auch eine Veränderung der Linkstruktur in solchen Situationen [KKP01].

Nicht alle Methoden dieser Gruppe sind unproblematisch für den Eindruck auf den Anwender. Vor allem die Änderung von wichtigen Navigationselementen, wie etwa ein Hauptmenü, kann für Verwirrung sorgen. Eine solche Adaption sollte, wenn überhaupt, nur in sehr großen zeitlichen Abständen erfolgen, so dass der Benutzer nicht bei jedem Seitenaufruf eine geänderte Menüstruktur vorfindet. Die implizierte Änderung in Suchen oder bei Empfehlungen hingegen hat keine Auswirkung auf die Benutzerinteraktion, weil in diesen Fällen eine Änderung der Objektreihenfolge erwartet wird.

Die einzig wirklich relevante Anpassung von Links in Tourismussysteme ist die implizite Sortierung durch Such- und Empfehlungsprozesse. Diese sind essentiell, um dem Benutzer relevante Inhalte zu präsentieren. Alle anderen Techniken sind eher nebensächlich und sind in der Regel den Aufwand für die Umsetzung nicht wert.

Unter die Kategorie der strukturellen Anpassung fallen folgende Methoden [KKP01].

- Kollaterale Adaption von Links entsteht durch Anpassungen von Inhalten, wobei beispielsweise durch Weglassen von Fragmenten einer Seite auch Links entfernt werden.

- Die Sortierung von Links kann je nach Relevanz der Ziele für die Anwender adaptiert werden. Dies beinhaltet auch die impliziten Varianten durch Suchen und Empfehlungssysteme

- Wichtige Links können hervorgehoben werden, um mehr Aufmerksamkeit zu erzeugen.

- Durch das Ein- oder Ausblenden von Links können diese als solche gekennzeichnet oder als normaler Text dargestellt werden.

- Links können aktiviert oder deaktiviert werden. Dabei bleiben diese für einen Anwender sichtbar, bei einem Klick auf einen Link wird er jedoch entsprechend weitergeleitet oder nicht.

- Verknüpfungen können dynamisch in eine Seite aufgenommen oder entfernt werden. Hierbei ist nach der Entfernung kein Hinweis auf den Link mehr vorhanden.

5.2 Personalisiertes Suchen

Es gibt im Wesentlichen zwei unterschiedliche Wege, auf denen ein Anwender zu gesuchter Information in einer Webapplikation kommen kann. Einerseits kann so lange durch die angebotenen Webseiten navigiert werden, bis der gewünschte Inhalt gefunden ist [Mic+07]. In diesem Fall ist ein Benutzer der alleinige Entscheidungsträger. Er wählt aus, auf welche Seite als nächstes gewechselt wird. Der klare Nachteil hierbei ist, dass unter Umständen eine große Anzahl an Zwischenschritten nötig ist, um die gewünschten Daten zu erhalten. Dies hängt auch direkt mit der Größe des Informationssystems zusammen. Weiters kann es möglich sein, dass diese erst gar nicht gefunden werden, weil sie nur unzureichend oder etwa gar nicht verlinkt sind. Andererseits erhält ein Anwender mehr Information, welche ihn eventuell auch interessiert. Dadurch kann er allerdings von seinem eigentlichen Suchziel abgelenkt werden. Die zweite Möglichkeit ist die Verwendung einer Suchmaschine [Mic+07]. Diese suchen gezielt nach relevanten Dokumenten, basierend auf den Vorgaben des Anwenders. Dieser gibt im Normalfall einzelne Wörter oder Phrasen in ein Suchfeld ein. Dokumente, welche zu dieser Eingabe relevant sind, sollen durch die Suche gefiltert werden und dem Benutzer präsentiert werden. Diese Methode basiert auf der klassischen Informationsbeschaffung, in welcher

Dokumente sowie Abfragen durch Stichwörter dargestellt werden [Mic+07]. Dadurch wird ein Vergleich ermöglicht und relevante Inhalte können auf eine Abfrage zurückgeliefert werden.

In web-basierten System besteht immer die Möglichkeit zur Navigation [Mic+07]. Es ist demnach üblich, dass die beiden Techniken zur Suche gemeinsam vorhanden sind und vom Anwender kombiniert werden können [Mic+07]. Durch die Interaktion mit unterschiedlichen Webseiten erhalten Benutzer ein Gefühl für Ausdrücke, welche sie in späteren Suchen verwenden können [Mic+07]. Dies ermöglicht dann ein gezieltes Auffinden von benötigten Informationen. Basierend auf Seiten im Ergebnis können danach wieder durch das Aufsuchen von verknüpften Seiten verwandte Informationen eingeholt werden.

In diesem Abschnitt werden personalisierte Suchen mittels Suchmaschinen behandelt. Ohne den Einfluss von Eigenschaften, wie Interessen, des Anwenders in Suchanfragen würde das Ergebnis für alle Benutzer gleich sein, eine identische Wortfolge in der Eingabemaske vorausgesetzt [Mic+07]. In diesem Fall muss die Ergebnismenge danach noch manuell durchgesehen werden, um für die jeweilige Person passende Objekte zu ermitteln. Indem Daten aus dem Benutzerprofil in den Suchprozess eingebunden werden, soll diese Arbeit minimiert werden. Als Ergebnis sollen entweder nur für den Anwender relevante Inhalte diesem präsentiert oder alternativ jene Seiten mit höchster Ähnlichkeit zu den Daten aus dem Benutzerprofil unter den Resultaten ganz am Anfang angezeigt werden [Mic+07]. Durch diese Änderung der Ergebnismenge soll der Benutzer also möglichst schnell die gewünschten Informationen geliefert bekommen. Weiters besteht die Möglichkeit, Suchanfragen so zu erweitern, dass Vorlieben eines Anwenders automatisch berücksichtigt werden [Mic+07]. Dies bedeutet, dass Resultate auf Suchen auch Kriterien entsprechen, welche nicht explizit in der Suchanfrage vorkommen. Dadurch soll ohne Aufwand von Benutzerseite der präsentierte Inhalt noch besser auf die jeweilige Person zugeschnitten werden.

Abhängig davon, zu welchem Zeitpunkt Informationen aus dem Benutzerprofil in den Suchprozess eingebunden werden, ergeben sich drei unterschiedliche Arten der Adaption [Mic+07]. Diese werden im weiterer Folge genauer erläutert.

- Durch die Berechnung der Ähnlichkeit der Objekte in der Ergebnismenge mit dem Benutzerprofil können diese entsprechend der Relevanz sortiert werden.

- Nach der Durchführung einer generischen Suche können die Inhalte der Resultatmenge durch Vergleich mit dem Profil neu sortiert werden.

- Vor der Verarbeitung der Suchanfrage kann diese um Daten aus dem Profil erweitert werden, um direkt spezifischere Ergebnisse zu erhalten.

Des Weiteren gibt es Möglichkeiten zur Personalisierung basierend auf Techniken, welche in Empfehlungssystemen gängig sind, sowie eine Einbeziehung des Anwenders durch interaktive Navigationsmöglichkeiten in der Ergebnismenge [Mic+07].

5.2.1 Personalisierung der Suchergebnisse durch Sortierung

Die Adaption der Suchergebnisse kann direkt als Teil der Informationsbeschaffung erfolgen. In diesem Fall ist kein extra Schritt notwendig, um für den Benutzer relevante Inhalte an eine bessere Stelle zu reihen [Mic+07]. Die Daten aus dem Benutzerprofil werden verwendet, um für die Suchanfrage relevante Dokumente weiter einzuschränken. Dies geschieht, indem der Grad der Ähnlichkeit zwischen den Seiten und dem Benutzerprofil des Anwenders berechnet wird [Mic+07]. Dementsprechend können Inhalte aus der Ergebnismenge entfernt oder an eine bessere Stelle gereiht werden.

Angenommen, ein Anwender sucht nach einer Unterkunft mit den Stichwörtern *Doppelzimmer*, *Alpen* und *Internet*. Er möchte also ein Doppelzimmer in den Alpen buchen, welches einen Internetzugang bietet. Findet sich nun im Benutzerprofil die Information, dass diese Person in der Freizeit gerne läuft, wird dieser Umstand direkt beim Ähnlichkeitsvergleich miteinbezogen. Falls Unterkünfte keine Möglichkeit zum Laufen in der Nähe anbieten, werden diese gar nicht in die Ergebnismenge aufgenommen oder nur mit niedriger Relevanz angeführt.

Das Problem bei dieser Technik ist, dass potentiell relevante Inhalte dem Benutzer vorenthalten werden. Dies ist der Fall bei der Entfernung von Ergebnissen auf die originale Suchabfrage, welche den Daten im Benutzerprofil nicht ähnlich sind [KL03]. Ist es für einen Anwender bei einer Suche nicht relevant, ob er an den angebotenen Destinationen Laufen kann oder nicht, fallen alle Unterkünfte ohne diese Eigenschaft aus dem Resultat und der Benutzer muss sich mit einer eventuell aus seiner Sicht schlechteren Übernachtungsmöglichkeit zufrieden geben.

5.2.2 Erneute Sortierung der Resultate

Eine weitere Möglichkeit ist die erneute Reihung der Ergebnisse einer Suche [Mic+07]. Dabei wird diese nur mit den vom Anwender eingegebenen Suchkriterien durchgeführt. Die zurückgegebenen Dokumente werden anschließend mit dem Inhalt des jeweiligen Benutzerprofils vergleichen und entsprechend der Ähnlichkeit neu geordnet [Mic+07]. Dadurch sind alle Ergebnisse, welche für die Abfrage relevant sind, nach wie vor für den Anwender sichtbar. Es werden ihm lediglich zum ihm besser passende Objekte früher angezeigt [Mic+07]. Der gesamte Suchprozess wird somit in zwei Schritte aufgeteilt. Die Verarbeitung des Resultates kann hierbei separat erfolgen, beispielsweise auf dem Endgerät des Benutzers [Mic+07]. Um dabei zu vermeiden, dass alle in der Ergebnismenge vorhandenen Elemente auf Ähnlichkeit überprüft werden müssen, kann eine begrenzte Anzahl der am höchsten gereihten Dokumente zur erneuten Sortierung evaluiert werden [Mic+07].

Im selben Beispiel wie zuvor werden bei der Abfrage alle Dokumente, welche durch die Stichwörter *Doppelzimmer*, *Alpen* und *Internet* repräsentiert werden, in die Ergebnismenge aufgenommen. Angenommen, das Dokument fünfter Stelle basierend auf der Sortierung enthält in seiner Klassifizierung weiterhin das Stichwort *Laufen*, wohingegen alle besser gereihten Elemente dieses nicht besitzen, dann hätte dies zur Folge, dass dieses Dokument an die vorderste Position gereiht wird, weil es für den Anwender aufgrund der Daten in dessen Profil als relevanter eingestuft wird.

Der Nachteil hierbei ist, dass auf die ursprüngliche Suchanfrage bereits hochqualitative Dokumente in die Ergebnismenge aufgenommen werden müssen [KL03]. Nur dadurch kann durch eine erneute Sortierung eine Verbesserung erzielt werden. Fall keine relevanten Objekte im Resultat vorhanden sind, ergibt eine neue Reihung keinerlei Unterschied.

5.2.3 Modifizierung der Anfrage

Die Suchanfrage kann weiters direkt vor dem Start des Suchprozesses geändert werden [Mic+07]. In diesem Fall wird diese um Stichwörter oder Phrasen aus dem Benutzerprofil erweitert [Mic+07]. Erst danach wird die Anfrage abgesetzt. Dies bedeutet praktisch keinen Mehraufwand für das System bei der Suche selbst

[Mic+07]. Weiters benötigt es auch nur geringe Ressourcen, um Daten aus dem Benutzerprofil zu extrahieren. Durch diese Expansion sind keine weiteren Schritte zur Personalisierung notwendig. Die Information ist bereits in der Abfrage enthalten und wird dadurch bei der Auswahl von passenden Dokumenten aus dem Informationssystem berücksichtigt [Mic+07]. Dadurch, dass bereits vor der Ausführung eine Anpassung erfolgt, können keine Ergebnisse, welche nur der vom Benutzer festgelegten Anfrage entsprechen, ins Ergebnis aufgenommen werden [Mic+07].

Weiters kann diese Art der Suchadaption dafür verwendet werden, um die Anfrage basierend auf relevantem Feedback des Anwenders zu ändern und eine erneute Suche abzusetzen [KL03]. Diese Feedback kann entweder explizit erfolgen, indem der Benutzer ein Ergebnis positiv oder negativ markiert [KL03]. Alternativ dazu kann implizites Feedback durch die besuchten Seite aus der Ergebnisliste generiert werden [KL03]. Diese Information kann bei der Rückkehr zur Suche für eine weitere Abfrage genutzt werden. Dabei können Stichwörter in die Suche eingefügt oder entfernt werden [KL03]. Weiters besteht die Möglichkeit, bereits vorhandene Terme stärker zu gewichten [KL03]. Üblicherweise ist die Verwendung dieser Methode an einen iterativen Prozess gebunden. Je mehr Iterationen durchgeführt werden, desto genauer wird die Suche [KL03].

Im bereits erwähnten Beispiel kann die Anfrage des Benutzers (*Doppelzimmer*, *Alpen* und *Internet*) um das Wort *Laufen* erweitert werden, bevor die Suche gestartet wird. Dadurch wird auch nach der Möglichkeit zum Laufen in der Umgebung einer Unterkunft gefiltert. Falls der Benutzer danach aus der Ergebnisliste ein Objekt auswählt und auf die Detailseite wechselt, kann aus diesem Dokument implizites Feedback generiert werden. Ist in dessen Repräsentation etwa auch das Stichwort *Fernseher* vorhanden, kann die Suche um diesen Begriff erweitert werden. Bei der Rückkehr des Anwenders auf die Suchseite wurde bereits eine neue Anfrage gestartet und es werden weitere Unterkünfte mit der Möglichkeit des Fernsehens präsentiert.

Bei dieser Technik besteht die Möglichkeit, dass irrelevante Inhalte in der Ergebnismenge auftauchen [KL03]. Vor allem durch die Modifizierung der Anfrage basierend auf Feedback kann rasch dazu führen, dass Schlagwörter in die Suche mit aufgenommen werden, welche vom Anwender gar nicht erwünscht sind [KL03]. Bietet eine Unterkunft zum Beispiel inkludiertes Frühstück und der Benutzer sieht sich die Seite im Detail an, kann das System die Abfrage um diese Eigenschaft erweitern. Vielleicht war diese Zusatzleistung auf der Übersichtsseite

der Suchergebnisse jedoch nicht zu finden und die entsprechende Person möchte nur übernachten, ohne ein Frühstück zu konsumieren, wird die Suche in eine falsche Richtung gelenkt.

5.2.4 Verwendung von historischen Suchdaten

Suchanfragen, welche ein Anwender bereits früher gestellt hat, können für aktuelle Suchen nützlich sein. Diese werden vom System gespeichert. Dazu werden auch Ergebnisse sowie jene Objekte aufgezeichnet, an denen der Benutzer genauer interessiert war. Diese Informationen können für aktuelle Suchen in unterschiedlicher Form verwendet werden [KL03].

- Die gesamte Suchanfrage kann mit früher verwendeten Suchen verglichen werden [KL03]. Besteht ein gewisser Grad an Ähnlichkeit mit einer historischen Anfrage, können die Ergebnisse von dieser auch für die aktuelle Suche verwendet werden. Hierbei werden ganze Mengen an Resultaten erneut verwendet. Dies ist mitunter problematisch, falls seitdem neue Inhalte in das System eingefügt wurden, welche eventuell besser zu einer Suche passen würden [KL03].

- Die Ergebnisse einer Suche können so sortiert werden, dass früher besuchte Seiten dem Benutzer zuerst vorgeschlagen werden [KL03]. Durch die Verknüpfung dieser Dokumente mit historischen Suchen wird sichergestellt, dass die Inhalte auch für die aktuelle Suche relevant sind. Dies hat den Vorteil, dass auch neu erstellte Webseiten in die Ergebnismenge miteinbezogen werden können [KL03].

Diese Art von Personalisierung ist etwa sinnvoll, falls Anwender öfters nach gleichen oder ähnlichen Inhalten suchen [KL03]. Zum Beispiel könnte ein Benutzer jeden Winter nach Unterkünften in den Alpen suchen, weil er gerne Ski fährt und üblicherweise jedes Jahr zur selben Zeit Urlaub hat. Jedoch wird er nicht jedes Jahr die selben Stichwörter in die Suchmaske eingeben. Durch die Verwendung von historischen Ergebnissen wird ihm die Möglichkeit geboten, bereits betrachtete Unterkünfte erneut zu finden. Wenn er jedoch lieber woanders verweilen möchte, kann er dies ignorieren und neue Resultate ansehen.

5.2.5 Suchergebnisse basierend auf ähnlichen Benutzern

Bei Suchen können andere Benutzer einer Webapplikation in Betracht gezogen werden. Soziale Methoden sind bereits in Empfehlungssystemen weit verbreitet [Mic+07]. Ein Benutzer wird als Teil einer Gesellschaft angesehen. Es wird angenommen, dass eine Person an Inhalten interessiert ist, welche für ähnliche Anwender von Relevanz sind [Mic+07]. Auch für die Suche können die Daten von anderen verwendet werden. Die Idee ist, dass Ergebnisse auf Suchanfragen mehr Relevanz erhalten, falls diese in der Gesamtheit populär sind [Mic+07]. Im Vergleich dazu können Resultate schlechter gereiht werden, wenn sie nicht sehr attraktiv für die Anwender erscheinen. Hierbei bieten sich die folgenden Möglichkeiten.

- Die Beliebtheit von Webseiten wird entweder durch implizites oder explizites Feedback (Bewertung beziehungsweise Aufrufen der Seite) bestimmt [Mic+07]. Diese Metrik kann danach in Suchergebnissen verwendet werden, um Seiten von großer Beliebtheit als erstes zu präsentieren [Mic+07]. Eine Erweiterung davon ist, dass die Beliebtheit getrennt nach Stereotypen betrachtet wird. Dadurch wird sichergestellt, dass nicht die Gesamtheit an unterschiedlichen Benutzern Einfluss auf die Sortierung hat, sondern nur ähnliche zum aktuellen Anwender.

- Alternative dazu können ähnliche Benutzer des aktuellen ermittelt werden [KL03]. Wenn Dokumente in einem Suchergebnis vorhanden sind, welche von diesen ähnlichen Personen gut bewertet wurden, wird die Relevanz für die aktuelle Operation erhöht [KL03]. Durch die Verbindung mit historischen Suchanfragen wird sichergestellt, dass Seiten auch nur dann einen besseren Stellenwert erhalten, falls diese auch damals im Zusammenhang mit einer ähnlichen Suche für gut empfunden wurden.

5.2.6 Gruppierung von Ergebnissen

Traditionell werden Resultate von Suchen als lange Liste, aufgeteilt auf mehrere Seiten, präsentiert. Die Titel der Ergebnisse werden mit kurzen Inhaltstexten angereichert, um dem Anwender eine Vorschau zu ermöglichen [KL03]. Dieser muss diese Liste auf relevante Einträge untersuchen, um von den ausgewählten Elementen nähere Information zu erlangen, indem er zur entsprechenden Seite wechselt [KL03]. Je nach Qualität der Abfrage und der zurückgelieferten Dokumente kann

deren Abarbeitung eine geraume Zeit in Anspruch nehmen, vor allem wenn keine relevanten Treffer am Beginn der Ergebnisse vorhanden sind. Durch andere Organisationen der Resultate kann ein effizienteres Auffinden von gewünschten Inhalten ermöglicht werden. Weiters wird dem Benutzer die Möglichkeit geboten, sich interaktiv an der Selektion von für ihn wichtigen Themenbereiche zu beteiligen [KL03].

Zu diesem Zweck werden die Ergebnisse in Gruppen unterteilt. Diese Einteilung richtet sich nach Themenzugehörigkeit der Dokumente [KL03]. Dies geschieht zum Zeitpunkt der Suche und die Cluster werden dynamisch generiert. Anschließend wird jeder der Gruppen eine aussagekräftige Bezeichnung oder Beschreibung gegeben. Dem Benutzer werden anfangs nur die unterschiedlichen Cluster präsentiert [KL03]. Dieser hat die Möglichkeit, einen oder mehrere auszuwählen und die darin enthaltenen Dokumente zu betrachten [KL03]. Bei einer gut funktionierenden Methode zur Gruppierung und Benennung dieser kann dies die Interaktion von Anwendern mit einer Suchmaschine deutlich vereinfachen und beschleunigen.

5.3 Empfehlungssysteme

Empfehlungssysteme werden verwendet, um einem Anwender Vorschläge bezüglich des Inhalts eines Informationssystems zu machen [RRS11] [Jan+10]. Dabei sollen die vorgeschlagenen Objekte in gewisser Weise für den Benutzer relevant sein, um seine Aufmerksamkeit zu erlangen [Jan+10]. Dadurch wird dem Benutzer beispielsweise erspart, dass er die mitunter große Menge an Informationen nach interessanten Inhalten durchforstet [RRS11]. Ähnlich wie bei personalisierten Suchen wird das Benutzerprofil als Grundlage für die Adaption verwendet. Einzelne Dokumente werden mit den Daten im Profil verglichen, um relevante Objekte durch Ähnlichkeit zu ermitteln [RRS11]. Alternativ können andere Profile evaluiert werden. Damit werden Nutzer mit ähnlichen Vorlieben eruiert. Unter der Annahme, dass sich ähnliche Anwender für die gleiche beziehungsweise ähnliche Information interessieren, können von diesen Benutzern für relevant empfundene Seiten dem aktuellen Anwender präsentiert werden [RRS11] [Jan+10].

In kommerziellen web-basierten Systemen sollen damit auch die Impulse von Personen angesteuert werden. Diese werden dann möglicherweise dazu verleitet, Produkte zu erwerben, wegen denen sie eigentlich gar nicht mit der Webapplikation

interagieren. Diese führt zur Steigerung des Umsatzes auf Betreiberseite [RRS11]. Ein weiterer Grund für die Verwendung ist, dass Personen im täglichen Leben ebenfalls häufig auf die Empfehlungen von Freunden oder Verwandten zurückgreifen, wenn es um den Kauf eines Produktes geht [RRS11]. Mit Empfehlungssystemen wird versucht, diesen Prozess nachzuahmen. Dem Anwender wird das Gefühl vermittelt, dass seine Vorlieben bekannt sind und er vom System individuell behandelt oder betreut wird [RRS11].

Die folgenden Aktionen zählen zu den Hauptaufgaben eines Empfehlungssystems [RRS11]. Hierbei werden nicht nur Bedürfnisse der Endanwender, sondern auch jene der Betreiber betrachtet.

- Die Anzahl der abgeschlossenen Transaktionen soll, wie bereits angemerkt, aus Sicht des Betreibers erhöht werden.
- Das System soll durch Feedback Informationen über den Anwender lernen, damit dessen Bedürfnisse besser eingeschätzt werden können.
- Inhalte, welche für den Benutzer als relevant betrachtet werden, sollen präsentiert werden. Dies kann unter anderem die Zufriedenheit steigern.
- Dem Anwender sollen Objekte präsentiert werden, die ohne Empfehlungen nur schwer bis gar nicht auffindbar wären.
- Ähnliche zu den bereits angesehenen Inhalten sollen vorgeschlagen werden, da angenommen wird, dass diese ebenfalls interessant sind.

Damit Empfehlungen abgegeben werden können, müssen gewisse Informationen beziehungsweise Elemente in einem Informationssystem vorhanden sein. Neben der Repräsentation von Benutzern durch Profile (ausführlich in Kapitel 3 behandelt), sowie die Darstellung von Inhalten in klassifizierter Form (beschrieben im Abschnitt 4), müssen weiters Transaktionen aufgezeichnet werden [RRS11]. Diese sind Interaktionen von Benutzern mit dem Empfehlungssystem und stellen wichtige Information für spätere Prozesse zur Empfehlung bereit. Die populärste Art solcher Daten sind Bewertungen, die beschreiben, wie relevant bestimmte Inhalte für Anwender sind [RRS11].

Es gibt grundsätzlich zwei große Gruppen von Systemen zur Empfehlung. Inhaltsbasierte Systeme wählen Objekte basierend auf der Ähnlichkeit mit den Daten aus dem Benutzerprofil aus [RRS11] [Jan+10]. Üblicherweise wird das Profil durch

früher abgegebene Bewertungen mit Informationen befüllt [RRS11] [Jan+10]. Umso genauer diese den Eigenschaften der verknüpften Person entsprechen, um so besser können Empfehlungen abgegeben werden. Kollaborative Systeme hingegen empfehlen Inhalte basierend auf ähnlichen Anwendern zum aktuellen oder ähnlichen Inhalten zu jenen, welche der Benutzer früher als relevant empfunden hat [RRS11]. Der Unterschied bei der Betrachtung von ähnlichen Inhalten in kollaboratievn Methoden zur inhalts-basierten Technik ist, dass die Bewertungen für ähnliche Objekte herangezogen werden und keine Ähnlichkeit basierend auf der Repräsentation des Inhaltes berechnet wird [RRS11] [Jan+10].

5.3.1 Bewertungen für Empfehlungssysteme

Wie bereits beschrieben, sind Bewertungen der Ausdruck von Relevanz eines Objektes für einen Anwender. Diese Information ist speziell bei Empfehlungen von Bedeutung, welche auf bereits bewerteten Elementen basieren. Dabei kann zwischen expliziten und impliziten Arten unterschieden werden [RRS11] [Jan+10]. Explizite Bewertungen werden bewusst vom Anwender vorgenommen [Jan+10]. Diese Form ist gebräuchlicher und aussagekräftiger [RRS11]. Beispiele für unterschiedliche Verfahren sind etwa numerische Vergabe von Punkten [RRS11] [Jan+10]. Dabei kann die grafische Darstellung angepasst werden, etwa statt Eingabe einer Zahl, unterschiedlich viele Sterne zu vergeben [RRS11]. Eine weitere Möglichkeit sind fest vorgegebene Begriffe, welche einer Reihung entsprechen [RRS11]. Ein Beispiel ist etwa die Aufteilung in *schlecht*, *passabel*, *gut* und *exzellent*. Die Anzahl der Zustände kann vereinfacht werden, indem ein binäres Modell verwendet wird [Jan+10]. Hierbei hat der Anwender dann nur die Möglichkeiten, anzugeben ob der entsprechende Inhalt für ihn relevant ist oder nicht [RRS11]. Es besteht auch die Variante, festzuhalten, ob etwa eine Webseite vom Benutzer besucht wurde. In diesem Fall ist lediglich ein Zustand notwendig. Diese Art wird vor allem in impliziten Methoden verwendet [RRS11] [Jan+10]. Ein weiteres Beispiel hierfür ist die Aufzeichnung über einen Transaktionsabschluss.

In einem Tourismussystem kann beispielsweise eine Bewertungsskala von eins bis fünf eingesetzt werden. Damit wird einem Anwender eine angemessene Menge an unterscheidbaren Zuständen angeboten. Um das Erlebnis der Interaktion zu steigern, kann die Abgabe der Bewertung über ein grafisches Element erfolgen, je nach Verschieben der Maus über Sterne oder ähnliche Symbole, kann die Zahl erhöht oder erniedrigt werden.

5.3.2 Inhalt-basierte Systeme

In Empfehlungssystemen, welche auf dem Inhalt von Dokumenten basieren, wird das Profil eines Anwenders für Vergleichszwecke herangezogen [PB07] [Jan+10]. Dieses enthält Interessen und Vorlieben des Benutzers. Es stellt eine Zusammenfassung von Eigenschaften der jeweiligen Person dar. Hierbei haben Objekte, welche historisch vom Anwender als relevante bewertet wurden, einen großen Einfluss auf den Inhalt des Profils [PB07]. Webseiten mit Informationen werden durch geeignete Methoden klassifiziert [LGS11]. Diese müssen so ausgewählt werden, dass ein Vergleich dieser Repräsentation mit den Daten aus dem Benutzerprofil möglich ist [LGS11]. Durch diesen wird die Ähnlichkeit von Objekten zu den Interessen des Anwenders berechnet [Jan+10]. Liegt ein bestimmter Grad an Gleichheit vor, wird der entsprechende Inhalt als relevant betrachtet und kann der jeweiligen Person vorgeschlagen werden [LGS11] [Jan+10].

Diese Empfehlungstechnik benützt den selben Prozess wie ihn auch die klassische Informationsbeschaffung einsetzt [LGS11]. Dokumente müssen zuerst durch eine festgelegte Form dargestellt werden. Gebräuchlich ist hierbei die Repräsentation durch Stichwörter. Nach der Vorbereitung (Entfernen unnützer Textteile, Stammformreduktion, Gewichtung der Terme) der Dokumententexte können diese verglichen werden. Die Inhalte des Benutzerprofils müssen auf die gleiche Weise gespeichert sein. Durch geeignete Vergleichstechniken wird die Ähnlichkeit zwischen einzelnen Webseiten und dem Benutzerprofil ermittelt [LGS11] [Jan+10]. Hierbei bietet sich zum Beispiel die Kosinus-Ähnlichkeit an. Diese wird im Vektorraummodell verwendet, wo die Dokumente als Vektor im Raum dargestellt werden und die Ähnlichkeit als Winkel zwischen den entsprechenden Dokumenten repräsentiert wird. Nach dem Vergleich entsteht eine Reihung der verschiedenen Dokumente im Informationssystem [LGS11]. Diese richtet sich nach dem Grad der Ähnlichkeit zu den Daten im Benutzerprofil. Aus diesen kann eine festgelegte Menge an Elementen für die Empfehlung verwendet werden [LGS11]. Eine genauere Beschreibung zum Prozess der klassischen Informationsgewinnung findet sich in Kapitel 4.

Ein Vorteil dieser Methode ist, dass sie nur auf den verfügbaren Informationen in der Webapplikation sowie den Inhalten eines Benutzerprofils aufbaut [LGS11]. Dadurch müssen keine anderen Anwender beziehungsweise deren abgegebenen Bewertungen berücksichtigt werden. Dies bedeutet, dass solche Systeme relativ unabhängig sind [LGS11]. Weiters ist der Prozess hinsichtlich der Auswahl an Ob-

jekten für die Vorschläge transparent. Der Inhalt auf den entsprechenden Webseiten beeinflusst direkt die Ähnlichkeit zum Benutzerprofil und somit die Reihung in der Ergebnismenge [LGS11]. Durch eine Änderung des Textes zu einem Objekt verändert sich dessen Relevanz für einen Anwender. Falls neue Inhalte in ein Informationssystem eingefügt werden, können diese sofort empfohlen werden [LGS11]. Es werden keine Bewertungen benötigt, weil diese nicht zur Entscheidung über Relevanz beitragen. Dies bedeutet, dass sich diese Art von Systemen für Applikationen eignet, in denen kontinuierlich neue Informationen eingefügt werden [LGS11]. Weiters sind sie für den Einsatz von frisch gestarteten Informationssystemen, in denen eine geringe Anzahl von Anwendern vorhanden ist, optimal [LGS11].

Zu den Nachteilen zählt, dass Dokumente aussagekräftig für den Vergleich repräsentiert werden müssen [Jan+10]. Dies setzt voraus, dass die Informationen selbst von guter Qualität sowie in ausreichendem Umfang vorhanden sind [LGS11]. Andernfalls kann es passieren, dass vom System schwer unterschieden werden kann, ob ein entsprechendes Dokument für einen Anwender relevant ist oder nicht [LGS11]. Die Darstellung als Stichwörter zum Beispiel ignoriert multimediale Inhalte. Um spezifische Elemente aus dem Text semantisch anzureichern, müsste Domänenwissen vorhanden sein, in welchem mögliche Ausprägungen unterschiedlicher Konzepte vorhanden sind [LGS11]. Weiters können durch diese Methode für Empfehlungen keine unerwarteten Inhalte vorgeschlagen werden [Jan+10]. Dies hat den Grund, dass nur jene Inhalte empfohlen werden, welche eine hohe Ähnlichkeit zum Benutzerprofil aufweisen [LGS11]. Das führt zur Überspezialisierung [LGS11]. Wenn zum Beispiel ein Anwender immer einen speziellen Typ von Unterkünften bevorzugt, kann ihm keine stark abweichende Übernachtungsmöglichkeit angeboten werden. Als Lösung hierfür können zufällig ausgewählte Inhalte in die Empfehlung aufgenommen werden, auf die Gefahr hin, dass diese für den Anwender nicht relevant sind [LGS11] [Jan+10]. Die Eigenschaften der Ähnlichkeitsberechung zieht auch das Problem nach sich, dass zu ähnliche Objekte in gewissen Situatinen nicht empfohlen werden sollen. Dem kann entgegengewirkt werden, indem solche Elemente vor der Präsentation ausgefiltert werden [LGS11]. Benutzerprofile müssen weiters zu einem gewissen Grad individuell ausgeprägt sein, damit zugeschnittene Vorschläge gemacht werden können. Dies bedeutet, dass für einen neuen Anwender noch keine genauen Empfehlungen abgegeben werden können [LGS11].

5.3.3 Kollaborative Systeme

Kollaborative Empfehlungssysteme basieren nicht direkt auf der Ähnlichkeiten von Dokumenten zu Benutzerprofilen. Stattdessen werden Vorschläge basierend auf früher getätigten Bewertungen abgegeben [Sch+07] [Jan+10]. Um die korrekten Transaktionen auszuwählen, werden in einem ersten Schritt ähnliche Objekte ermittelt [Sch+07]. Hierbei können zum einen Benutzer ausfindig gemacht werden, welche dem aktuellen in dessen Eigenschaften und vorhandenen Bewertungen möglichst gleich sind [KB11] [Jan+10]. Ein anderer Ansatz ist, zu dem Objekt, für welches ein Wert für die Relevanz bestimmt werden soll, ähnliche Objekte zu eruieren [Jan+10]. Dabei wird auf inhaltliche Eigenschaften der jeweiligen Dokumente zurückgegriffen [KB11]. In einem weiteren Schritt werden danach die Bewertungen ähnlicher Anwender oder Inhalte zur Bestimmung für die Empfehlung eines nicht bewerteten Dokumentes verwendet [KB11] [Jan+10].

Um solche Methoden effektiv anwenden zu können, müssen bereits vorhandene Bewertungen in einer geeigneten Form im System hinterlegt sein. Hierzu bietet sich eine Bewertungsmatrix an [Sch+07] [Jan+10]. Die Zeilen einer solchen Matrix enthalten alle Anwender des Systems. In den Spalten hingegen werden alle verfügbaren Dokumente beziehungsweise Objekte dargestellt [Jan+10]. Die Inhalte sind Bewertungen, welche die Benutzer für die Inhalte abgegeben haben [Jan+10]. Durch diese Darstellungsform lassen sich einige für die kollaborative Filterung wichtigen Informationen rasch evaluieren. Alle von einem Benutzer bewerteten Objekte sind in einer Zeile vorhanden. Alle anderen Benutzer, welche ein Objekt ebenfalls bewertet haben, sind in der selben Spalte verfügbar. Ein fehlender Wert in einem Feld bedeutet, dass vom entsprechenden Anwender für das jeweilige Objekt noch keine Bewertung abgegeben wurde [Sch+07]. Folglich können alle noch nicht bewerteten Dokumente für eine Empfehlung herangezogen werden. Ein Beispiel für eine Bewertungsmatrix von Unterkünften ist in der Tabelle 5.1 veranschaulicht. Dabei wird eine numerische Bewertungsskala von eins bis fünf verwendet.

	Unterkunft 1	**Unterkunft 2**	**Unterkunft 3**	**Unterkunft 4**
Anwender 1	1	2	5	
Anwender 2		2	4	5
Anwender 3	5	5	2	1
Anwender 4	5	4	5	5

Tabelle 5.1: Ein Beispiel für eine Bewertungsmatrix in einem Tourismussystem

In diesem Beispiel ist zu erkennen, dass *Unterkunft 4* nicht von *Anwender 1*, sowie *Unterkunft 1* nicht von *Anwender 2* bewertet sind. Für diese Kombinationen kann also von einem Empfehlungssystem ein Grad an Relevanz ermittelt werden. Für einen gegebenen Anwender wird also für ein oder mehrere Objekte eine Bewertung berechnet.

Die in den folgenden Abschnitten vorgestellten Technik zur Abgabe von Empfehlungen basieren auf dem k Nächste-Nachbarn Prinzip. Das bedeutet, dass jeweils die k ähnlichsten Benutzer oder Dokumente ausgewählt werden [KB11] [Jan+10]. Die dadurch ermittelten vorhandenen Bewertungen werden anschließend für die Berechnung des Ergebnisses für die Bewertung eines Benutzers für ein Objekt verwendet [KB11] [Jan+10].

5.3.3.1 Empfehlungen auf Basis von ähnlichen Benutzern

Bei dieser Art von Empfehlungen werden zuerst ähnliche Anwender zu jenem bestimmt, für den die Relevanz eines Dokumentes ermittelt werden soll [KB11] [Jan+10]. Die Bewertungen dieser werden danach zur Berechnung verwendet, zu welchem Grad das Dokument für den entsprechenden Benutzer interessant ist [KB11] [Jan+10]. Hierbei wäre eine naive Methode, alle Benutzer aus der Bewertungsmatrix zu filtern, welche das in Frage stehende Objekt bewertet haben [Sch+07]. Der Durchschnitt aller Bewertungen kann danach für die Empfehlung angenommen werden. Dies ist durch die Funktion in 5.1 veranschaulicht [Sch+07]. Hierbei besteht natürlich das Problem, dass der Faktor der Ähnlichkeit in keiner Weise in die Berechnung einbezogen wird [Sch+07]. Dabei bezeichnet u den aktuellen Benutzer und i das Objekt, für welche eine Bewertung erfolgen soll. Weiters ist n einer jener Anwender aus der Menge $N_i(u)$, welche alle Benutzer beinhaltet, die das Dokument bereits bewertet haben. Die Anzahl dieser Benutzer ist durch $|N_i(u)|$ festgelegt.

$$r_{ui} = \frac{\sum_{n \in N_i(u)} r_{ni}}{|N_i(u)|} \tag{5.1}$$

Um diesen naiven Ansatz zu verbessern, kann eine fixe Menge bestehend aus k Benutzern ermittelt werden, welche dem aktuellen Benutzer am ähnlichsten sind [KB11]. Diese kann durch die Eigenschaften des Benutzerprofils ermittelt werden. Die Funktion zur Berechnung dieser Ähnlichkeit liefert einen numerischen

Wert zurück. Dieser kann in die Berechnung der Bewertung mit einfließen [KB11]. Damit wird erreicht, dass der Grad der Gleichheit eine Auswirkung auf die Bewertung des jeweiligen ähnlichen Anwenders hat [KB11]. In diesem Fall wird die Formel wie folgt verändert (siehe 5.2 [KB11]. Die Metrik der Ähnlichkeit zwischen zwei Benutzern wird mit w_{un} definiert.

$$r_{ui} = \frac{\sum_{n \in N_i(u)} w_{un} * r_{ni}}{\sum_{n \in N_i(u)} |w_{un}|} \tag{5.2}$$

Im bereits vorgestellten Beispiel der Bewertungsmatrix kann für *Anwender 2* eine Empfehlung bezüglich *Unterkunft 1* abgegeben werden. Anhand der naiven Methode berechnet sich das Resultat wie folgt (siehe 5.3). Falls im Gegensatz eine Anzahl von 2 ähnlichsten Benutzern festgelegt ist und in dieser Menge die Benutzer *Anwender 3* sowie *Anwender 4* mit den Werten 0.89 sowie 0.76 vorhanden sind, berechnet sich die Ähnlichkeit nach der Veranschaulichung in Formel 5.4. Es ist deutlich zu erkennen, dass durch die Auswahl einer fixen Anzahl an ähnlichen Benutzern sowie dem Einfluss der Gewichtung ein besseres Ergebnis erzielt wird.

$$r_{21} = \frac{1 + 5 + 5}{3} = 3.66 \tag{5.3}$$

$$r_{21} = \frac{0.89 * 5 + 0.76 * 5}{0.89 + 0.76} = 5 \tag{5.4}$$

Die genannte Methode basierend auf dem k Nächste-Nachbarn Prinzip kann jedoch noch weiter optimiert werden. Es wird nicht berücksichtigt, wie einzelne Anwender bei ihren Bewertungen vorgehen [KB11]. Manche vergeben etwa für alle Objekte den Höchstwert, sobald ihnen dieser in gewisser Weise zusagt und den schlechtesten Wert, wenn sie den entsprechenden Inhalt nicht mögen. Andere vergeben die höchste Bewertung etwa nur für Objekte, welche für sie wirklich von besonderer Bedeutung sind und teilen die bewerteten Dokumente über die gesamte Bewertungsskala auf. Aus diesem Grund können die von anderen Benutzern abgegebenen Bewertungen normalisiert werden [KB11]. Für das Endergebnis muss der berechnete Vorschlagswert wieder an die ursprüngliche Skala angepasst werden. Die einfachste Variante hierbei ist, die Zentrierung um den Mittelwert der vom Benutzer abgegebenen Bewertungen [KB11]. Diese wird wie folgt berechnet

(siehe 5.5) [KB11], wobei $h(r_{ui})$ dem normalisierten Wert entspricht. Die Formel 5.6 zeigt die Anwendung für die Bewertungsberechnung [KB11].

$$h_{ui} = r_{ui} - \overline{r_u} \tag{5.5}$$

$$r_{ui} = h^{-1}(\frac{\sum_{n \in N_i(u)} w_{un} * h(r_{ni})}{\sum_{n \in N_i(u)} |w_{un}|}) = \overline{r_u} + \frac{\sum_{n \in N_i(u)} w_{un} * (r_{ni} - \overline{r_n})}{\sum_{n \in N_i(u)} |w_{un}|} \tag{5.6}$$

Die Verwendung von Empfehlungen basierend auf ähnlichen Anwendern bringt einige Herausforderungen mit sich. Falls nur wenige Benutzer Bewertungen mit dem in Frage gestellten Anwender teilen, kann dies dazu führen, dass diese wenigen Personen dominant in den ähnlichsten Nachbarn vorkommen [Sch+07] [Jan+10]. Aufgrund dieser Spärlichkeit von Bewertungen wird der entsprechende Anwender nur durch eine geringe Menge anderer Benutzer beeinflusst. Dies ist auch der Fall, wenn im System nur eine geringe Anzahl von Anwendern vorhanden ist und es bereits eine Herausforderung ist, genügend ähnliche Benutzer von guter Qualität zu ermitteln [Sch+07]. Weiters wird die Gesamtheit der Bewertungen für ein Objekt nicht berücksichtigt [Sch+07]. So ist es von weit geringerer Bedeutung, dass ähnliche Benutzer ein Objekt gut finden, falls diese Empfindung auch bei den meisten anderen Personen vorherrscht. Andererseits ist des wichtig, diese Meinung zu berücksichtigen, falls nur eine geringe Menge an Benutzern diese teilt. Die Berechnung der Ähnlichkeiten zwischen Benutzern ist weiters sehr aufwändig. Vor allem die Speicheranforderungen steigen linear mit der Anzahl der Anwender und Bewertungen [Sch+07]. Methoden wie etwa Clustering von Benutzern können zur effizienteren Verarbeitung eingesetzt werden [Sch+07].

5.3.3.2 Empfehlungen auf Basis von ähnlichen Objekten

Bei der Variante von Empfehlungen basierend auf ähnlichen Objekten oder Dokumenten werden Bewertungen von jenen Inhalten verwendet, welche dem in Frage stehenden Objekt am ähnlichsten sind [KB11] [Jan+10]. In diesem Fall werden diese Objekte aus der bereits vom Anwender bewerteten Menge ausgewählt [KB11] [Jan+10]. Würde man bei dieser Methode ebenfalls den naiven Ansatz wählen, alle vom Benutzer bewerteten Elemente für eine Durchschnittsberechnung zu verwenden, würden alle Empfehlungen mit dem selben Ergebniswert erfolgen.

Auch in dieser Variante werden die k ähnlichsten Objekte ausgewählt. Die Bewertungen werden durch deren Ähnlichkeit zu dem Dokument, für welches eine Empfehlung abgegeben werden soll, gewichtet [KB11]. Die folgende Formel (5.7) veranschaulicht die Berechnung [KB11]. Hierbei entspricht j einem der ähnlichsten Elemente aus der Menge $N_u(i)$. Die Bewertung des aktuellen Benutzers für dieses Element ist mit r_{uj} bezeichnet. Weiters entspricht w_{ij} dem Grad der Ähnlichkeit zwischen dem aktuellen Objekt und einem aus der angesprochenen Menge.

$$r_{ui} = \frac{\sum_{j \in N_u(i)} w_{ij} * r_{uj}}{\sum_{j \in N_u(i)} |w_{ij}|} \tag{5.7}$$

Gleich wie bei der vorigen Methoden kann unter der Verwendung der beispielhaften Bewertungsmatrix ein Empfehlungswert für *Unterkunft 1* in Relevanz zu *Anwender 2* ermittelt werden. Das Ergebnis ist in der Rechnung 5.8 veranschaulicht. Hierbei wird die Annahme getroffen, dass *Unterkunft 4* sowie *Unterkunft 5* mit Ähnlichkeiten von 0.91 sowie 0.86 in der Menge von ähnlichen Dokumente mit einer Größe von $k = 2$ Elementen enthalten sind.

$$r_{21} = \frac{0.91 * 4 + 0.86 * 5}{0.91 + 0.86} = 4.49 \tag{5.8}$$

Hier ist zu erkennen, dass die stärkere Gewichtung der Ähnlichkeit zwischen *Unterkunft 4* mit *Unterkunft 1* zur Auswirkung hat, dass nicht exakt ein gemitteltes Ergebnis von 4.5 erreicht wird.

Auch bei dieser Methode von Empfehlungssystemen wird die individuelle Bewertungsskala von Anwendern nicht berücksichtigt. Aus diesem Grund ist es auch hier sinnvoll, die Bewertungen zu normalisieren und anschließend das Ergebnis wieder an die verwendete Skala anzupassen [KB11]. Analog zum Verfahren bei ähnlichen Anwendern, kann hier ebenfalls der Mittelwert aller abgegebenen Bewertungen eines Objektes von der Benutzerbewertung abgezogen werden. Die Berechnung der Normalisierung ist in der Formel 5.9 veranschaulicht [KB11]. Die Anwendung in der Bewertung findet sich in der Formel 5.10 [KB11].

$$h(r_{ui}) = r_{ui} - \overline{r_i} \tag{5.9}$$

$$r_{ui} = h^{-1}(\frac{\sum_{j \in N_u(i)} w_{ij} * h(r_{uj})}{\sum_{j \in N_u(i)} |w_{ij}|}) = \overline{r_i} + \frac{\sum_{j \in N_u(i)} w_{ij} * (r_{uj} - \overline{r_j})}{\sum_{j \in N_u(i)} |w_{ij}|} \tag{5.10}$$

Bei dieser Art von Vorschlägen bestehen generell die selben Probleme wie bei jener basierend auf ähnlichen Anwendern [Jan+10]. Falls nur eine geringe Anzahl an Objekten vom in Frage gestellten Benutzer bewertet sind, dominieren diese die Menge der ähnlichsten Objekte [Sch+07]. Dies führt bei der Berechnung dazu, dass alle Empfehlungen den selben Wert erhalten. Des weiteren besteht dieses Problem bei Systemen mit wenigen Inhalten. Dadurch kann es vorkommen, dass keine qualitativ hochwertigen ähnlichen Objekte eruiert werden können, um deren Bewertungen zu verwenden [Sch+07]. Weiters ist die Ermittlung von ähnlichen Inhalten in großen Kollektionen sehr rechen- und speicherintensiv [Sch+07]. Dies kann etwa durch Methoden verbessert werden, die nur Objekte in Betracht ziehen, zwischen denen eine bestimmte Anzahl an Bewertungen von gleichen Anwendern vorhanden sind [Sch+07].

5.3.3.3 Vergleich zwischen den beiden Varianten

Während Eigenschaften, wie etwa die benötigten Ressourcen, ähnlich in beiden Techniken sind, gibt es auch Unterschiede, wonach eines für bessere Resultate sorgt als das andere. Die Genauigkeit der Vorschläge hängt sehr stark vom Charakter des Informationssystems ab. Falls Anwender vorhanden sind, welche gewissenhafte Bewertungen abgeben, eignet sich der Ansatz basierend auf ähnlichen Anwender gut. Für große Systeme mit einer Vielzahl an Benutzern jedoch kann ein System basierend auf ähnlichen Inhalten bessere Ergebnisse erzielen [KB11]. Dies ist vor allem dann der Fall, wenn eine Vielzahl dieser Benutzer die Abgabe von Bewertungen nicht ernst nimmt. Wie stabil ein Empfehlungssystem ist, hängt von der Anzahl und Häufigkeit von Änderungen ab, welche Informationen betreffen. Falls sich laufend die Menge der Anwender ändert, kann mit einem inhalt-basiertem System effizent gearbeitet werden, indem Ähnlichkeiten voraus berechnet werden und nur bei selten vorkommenden Änderungen des Inhaltes erneut evaluiert werden [KB11]. Andererseits lässt sich dieses Verfahren für Systeme basierend auf ähnlichen Anwendern einsetzen, falls die Menge dieser relativ stabil ist und häufig inhaltliche Elemente hinzugefügt oder entfernt werden. Weiters besteht bei der Verwendung des Ansatzes basierend auf ähnlichen Anwendern eine höhere Chance, dass einem Benutzer Vorschläge gemacht werden, die er nicht erwarten würde [KB11]. Dadurch können ihm etwa neue Pro-

dukte vorgestellt werden [KB11]. Dies ist bei Systemen basierend auf dem Inhalt schwer, weil immer ähnliche zu bereits vom Anwender bewerteten Inhalte eruiert werden [KB11].

5.3.4 Hybride Systeme zur Problemminimierung

Wenn ein web-basiertes Informationssystem gerade erst in Betrieb genommen wurde, befinden sich in diesem üblicherweise noch nicht all zu viele Anwender. Weiters kann es sein, dass auch die Menge an inhaltlichen Informationen noch gering ist. In diesem Fall können keine kollaborativen Empfehlungssysteme verwendet werden [Bur02]. Der Vorteil dieser ist jedoch, dass auch Objekte vorgeschlagen werden können, welche ein Benutzer nicht erwartet. Dadurch können diesem neue Inhalte präsentiert werden [Bur02]. Falls ein solches System nun eingesetzt werden soll, besteht die Möglichkeit, diese mit inhalt-basierten Methoden zu kombinieren. Dabei wird zuerst auf ein System gesetzt, welches Inhalte basierend auf dem Vergleich mit dem Benutzerprofil vorschlägt. Dadurch können bereits Empfehlungen gegeben werden, bis genug Inhalt im Informationssystem vorhanden sind. Zu diesem Zeitpunkt kann auf ein kollaboratives System gewechselt werden. Weiters besteht die Möglichkeit, dass wieder auf das inhalts-basierte Verfahren zurückgegriffen wird, falls für eine Anwender und Objekt Kombination zu wenig ähnliche Benutzer oder Dokumente vorhanden sind, um ein aussagekräftiges Ergebnis zu erzielen. Dies erlaubt es, je nach Situation, zumindest Empfehlungen abgeben zu können, ohne eine leere Menge präsentieren zu müssen [Bur07]. Das Problem mit neuen Benutzern im System bleibt jedoch auch bei dieser Variante bestehen [Bur02]. Diese müssen ihr Profil erst individuell ausprägen. Erst dadurch können verlässliche Empfehlungen abgegeben werden. Eine Abhilfe ist die Zuweisung von Anwendern zu Stereotypen, bis genug Informationen für ein eigenständiges Profil gesammelt sind.

5.4 Unterstützung bei der Navigation

Anwender können beim Navigieren durch die einzelnen Seiten eines Informationssystems unterstützt werden, indem ihre Eigenschaften, gespeichert im Benutzerprofil, zur Anpassung der Linkstruktur ausgewertet werden [KKP01]. So kann ein Benutzer auf Inhalte aufmerksam gemacht werden, welche für diesen als in-

teressant erachtet werden. Um die Verlinkung entsprechend zu verändern, gibt es unterschiedliche Methoden. Es muss bei der Verwendung darauf geachtet werden, inwiefern sich die einzelnen Techniken für den gewünschten Einsatz eigenen. So kann ein Anwender beispielsweise sehr schnell verwirrt werden, wenn sich auf einmal andere Verknüpfungen an Positionen finden, wo zuvor ein Link mit gewohnter Funktion vorhanden war. Weiters kann durch verschiedene Techniken bestimmt werden, aufgrund welcher Eigenschaften oder Aktionen Links geändert werden sollen.

5.4.1 Möglichkeiten zur Anpassung von Links

In web-basierten Systemen gibt es verschiedene Arten, um Verknüpfungen zwischen Seiten zu ändern. Während eine Menge an Möglichkeiten etwa unterschiedliche Darstellungsform zur Veränderung nützt, können durch andere Techniken Links dynamisch auf Seiten eingefügt oder von diesen entfernt werden, so dass dadurch direkt die Verknüpfungen betroffen sind [KKP01]. Diese unterschiedlichen Techniken zur Änderung von Verlinkungen werden im folgenden Abschnitt näher erläutert.

5.4.1.1 Direkte Führung durch das System

Die direkte Unterstützung bei der Navigation leitet den Benutzer geführt durch die einzelnen Seiten des Informationssystems, indem der nächste zu besuchende Inhalt vorgeschlagen wird, welcher aus den im Benutzerprofil gespeicherten Daten ermittelt wird. Beispielsweise kann ein bereits auf der Seite vorhandener Link optisch aufgewertet werden, um die Aufmerksamkeit des Anwenders auf diesen Link zu lenken. Andernfalls kann ein Link dynamisch generiert und in die Seite eingefügt werden. [Bru07]

Bei dieser Technik ergibt sich der Nachteil, dass Anwender die Führung annehmen müssen. Wenn ein Benutzer nicht durch ein Informationssystem geführt werden möchte, erhält dieser keinerlei andere Unterstützung, um auf für ihn interessante andere Seiten hingewiesen zu werden [Bru07]. Deswegen ist diese Art der personalisierten Navigation vorrangig für adaptive Lernsysteme von Bedeutung [Bru07]. In diesen ergibt es durchaus Sinn, etwa nach Abschluss eines Themas zu einem verwandten Themenbereich weiterzuleiten. Dadurch kann dem Anwender ein zu-

sammenhängender Lernprozess vermittelt werden, ohne dass dieser zwischen komplett unterschiedlichen Wissensgebieten hin und her navigiert, um komplett willkürliche Inhalte zu lernen.

5.4.1.2 Sortierung von Links

Bei einer adaptiven Sortierung von Verknüpfungen auf einer Seite werden Links in einer Liste präsentiert [KKP01]. Diese wird je nach geänderten Eigenschaften im Benutzerprofil neu sortiert. Hierbei stehen Links zu für den Benutzer relevanten Seiten weiter oben in der Liste, wohingegen weniger wichtige ans Ende gereiht werden [KKP01]. Feedback von Anwendern kann durch manuelle Korrektur der Reihung eingebaut werden. Durch diese werden die Inhalte im Benutzerprofil entsprechend aktualisiert [Bru07]. Durch die geeignete Sortierung kann die Zeit, welche zur Navigation durch das System benötigt wird, verringert werden. Dadurch können dem Benutzer gewünschte Inhalte schneller präsentiert werden [Bru07]. Eine implizierte Variante von Sortierung findet sich bei der Darstellung von Suchergebnissen sowie gefilterten Inhalten von Empfehlungssystemen.

Für relativ unwichtige Navigationselemente mag diese Art der Anpassung eventuell Sinn ergeben. Falls diese Adaption jedoch etwa für das Hauptmenü eines ganzen Informationssystems verwendet wird, kann dies den Anwender sehr rasch verwirren. Hat sich dieser einmal an eine Sortierung der wichtigsten Links gewöhnt, kann die Liste bei der nächsten Interaktion mit der Webapplikation bereits wieder komplett anders aussehen. Deshalb sollte die Sortierung nur für Listen verwendet werden, in denen keine Stabilität vorausgesetzt wird [Bru07]. Ein Beispiel hierfür ist die Auflistung von Artikel in einer Onlinezeitung. Alternativ kann die Anpassung basierend auf Eigenschaften erfolgen, welche sich nie oder nur sehr selten ändern [Bru07]. Dadurch wird einmalig eine personalisierte Struktur ermittelt, welche dann beibehalten wird.

Ein sehr einfaches Beispiel bei Informationssystemen mit einer Suchmöglichkeit für Unterkünfte wäre etwa, dass sich die Eigenschaft eines Anwenders, dass dieser vorwiegend Schnäppchen in letzter Minute bucht, auf das globale Hauptmenü auswirkt. In diesem könnte ein Link zu Angeboten an die erste Stelle gereiht werden. Diese Eigenschaft ist auch relativ stabil über einen längeren Zeitraum. Deshalb ist die Gefahr einer Änderung zwischen Sitzungen relativ gering.

5.4.1.3 Verstecken von Links

Beim Verstecken von Links werden diese mittels verschiedener Techniken vor dem Benutzer verborgen. Dies soll dazu führen, dass Anwender nicht zu Inhalten gelangen, welche für sie ungeeignet sind [KKP01]. Diese Art der Adaption eignet sich vorwiegend für adaptive Lernsysteme. In diesen können Seiten mit Informationen zu Themen vom Benutzer verborgen werden, welche für dessen Lernerfolg entweder uninteressant sind oder erst zu einem späteren Zeitpunkt genügend Wissen zur Abarbeitung zur Verfügung steht [Bru07]. Des weiteren wird die Menge der Information verringert, welche der Benutzer aufnehmen muss. Er muss sich weniger Texte von Links durchlesen, bevor er eine Entscheidung bezüglich der Seite, auf die er wechseln möchte, trifft.

Es gibt zwei unterschiedliche Arten, Links von Anwendern zu verbergen [Bru07]. Diese können auch in Kombination verwendet werden.

- Das Verstecken von Links bedeutet, dass alle grafischen Besonderheiten entfernt werden. Der Link sieht anschließend aus wie normaler Text. Beim Lesen merkt ein Anwender also keinen Unterschied und nimmt die Verknüpfung als Textphrase wahr. Die Möglichkeit, per Klick zur verknüpften Seite zu gelangen, besteht jedoch weiterhin.

- Bei der Deaktivierung eines Links wird die eigentliche Weiterleitung beim Klick nicht ausgeführt. Die visuellen Anzeigen für das Vorhandensein eines Links sind jedoch unverändert. Dies kann den Benutzer frustrieren, falls er weiß, dass eine verknüpfte Seite vorhanden ist, diese jedoch nicht erreicht werden kann.

Für web-basierte Informationssysteme, welche nicht die Vermittlung von Wissen als Hauptaufgabe besitzen, gibt es keine triftigen Gründe, Verlinkungen von Seiten vor dem Benutzer zu verbergen. Zu diesen zählen auch kommerzielle Systeme, wie etwa jene, die Unterkünfte zur Übernachtung anbieten.

5.4.1.4 Erweiterung von Links

Bei dieser Variante werden Links durch visuelle oder textuelle Elemente angereichert [KKP01]. Damit soll die Aufmerksamkeit des Benutzers auf die entsprechenden Verknüpfungen gelenkt werden. Für den Benutzer interessante Inhalte

sollen demnach eine größere Chance bekommen, als nächstes aufgerufen zu werden [Bru07]. Eine Möglichkeit ist die Verwendung von Icons vor oder nach einem Link [Bru07]. Als Text selbst könnten etwa Rufzeichen vorangestellt sein, um die Wichtigkeit zu veranschaulichen. Beim Überfahren mit der Maus können kleine Popover angezeigt werden [Bru07]. Weiters kann die Form des eigentlichen Textes geändert werden (etwa fett geschrieben, eine andere Schriftart oder Farbe) [Bru07]. Diese Art der Adaption verändert die eigentliche Struktur einer Webseite nicht. Dies bedeutet, dass die Sortierung der Links stabil bleibt [Bru07]. Weiters lassen sich damit unterschiedliche Stufen von Relevanz realisieren [Bru07]. Aufgrund dieser Vorteile wird diese Art der Anpassung am häufigsten eingesetzt [Bru07].

In Tourismussystemen könnten eventuell auf Übersichtsseiten jene Links angereichert werden, hinter denen für den Anwender interessante Gruppen von Unterkünften dargestellt werden. Ist von einem Benutzer beispielsweise bekannt, dass er gerne ganze Ferienwohnungen bucht, kann diese Kategorie hervorgehoben werden. Andererseits lässt sich diese Information auch direkt bei der Filterung von Suchergebnissen anwenden. Die Nützlichkeit ist deshalb in Frage gestellt.

5.4.1.5 Dynamische Erzeugung von Links

Durch die Generierung von neuen Links können zuvor nicht verknüpfte Webseiten miteinander verlinkt werden [KKP01]. Dies geschieht automatisch, ohne manuelle Eingriffe. Im Unterschied zu anderen Techniken wird es also dem System überlassen, Navigationsmöglichkeiten zwischen relevanten Inhalten einzufügen [Bru07]. Als Gegensatz dazu können Links auch wieder entfernt werden, wenn diese nicht mehr von Bedeutung sind [KKP01]. Dadurch wird dem Benutzer nur angezeigt, was für ihn tatsächlich Relevanz hat. Es gibt drei unterschiedliche Methoden zur Erzeugung von Links [Bru07].

- Neue Verknüpfungen zwischen Seiten können zum Beispiel durch die Auswertung von Klickströmen als sinnvoll erachtet werden. Diese werden dann permanent in die Liste der existierenden Links aufgenommen. Falls eine Vielzahl von Nutzern über Umwege von einer Webseite zu einer anderen gelangen, kann eine direkte Verknüpfung als nützlich angesehen werden.

- Aufgrund von ähnlichen Inhalten von Seiten kann zwischen diesen verlinkt werden. Dadurch können einem Anwender zur aktuell betrachteten Information passende Inhalte angeboten werden.

- Der aktuelle Kontext des Anwenders kann evaluiert werden, um entsprechend von Zielen oder Interessen geeignete Links in die aktuell betrachtete Seite einzufügen. In Lernsystemen etwa lassen sich zu einer aktuellen Seite benötigte Hintergrundinformation verlinken, falls diese dem Benutzer noch nicht geläufig ist.

In Systemen mit Bezug auf den Tourismus ist am ehesten die erste der genannten Techniken sinnvoll. Durch die Auswertung von Navigationsmustern aller Anwender können eventuell zeitsparende Verknüpfungen angelegt werden. Falls etwa ein Großteil der Benutzer von der Startseite zuerst auf Seite mit allen Kategorien klickt und danach die Kategorie von Angeboten genauer betrachtet, würde es Sinn ergeben, direkt eine Verknüpfung von der Start- auf die Angebotsseite zu erstellen.

5.4.2 Mechanismen zur Adaption

Einfache Mechanismen zur Entscheidung über eine Anpassung der Linkstruktur basieren auf der Auswertung von simplen Daten. Eine Quelle ist die Anzahl der Seitenaufrufe in der Vergangenheit [Bru07]. Dies Information kann danach etwa visuell oder durch eine Änderung der Reihenfolge von Links in einer Liste repräsentiert werden. Häufig besuchte Seiten werden dabei hervorgehoben. Weiters können bestimmte Ereignisse als Anlass für die Adaption verwendet werden [Bru07]. In Lernsystemen kann so zum Beispiel die Verlinkung zu neuen Themen freigeschalten werden, sobald bestimmte Voraussetzungen erfüllt sind. Dies kann erweitert werden, indem der Lernfortschritt in hierarchisch angeordneten Themen propagiert wird und somit zuerst alle Elemente eines Themas erlernt sein müssen, bevor weitere übergeordnete Elemente zur Verfügung stehen.

Eine andere Methode, um zu bestimmen, ob ein Anwender die Möglichkeit zur Navigation von einer Seite zu einer anderen hat, sind Analysen über die Inhalte. Hierbei werden die repräsentativen Inhalte mit den Daten des Benutzerprofils oder den Inhalten von aktuell besuchten Dokumenten verglichen [Bru07]. Besteht ein gewisser Grad an Ähnlichkeit, wird ein Link zwischen den entsprechenden

Seiten eingefügt oder mit zusätzlicher Information angereichert. Diese Techniken sind auch jene, die in Empfehlungssystemen verwendet werden.

Weiters können soziale Eigenschaften verwendet werden. Die Idee dahinter ist, dass Anwender üblicherweise Vorschlägen von anderen Personen folgen [Bru07]. Eine Möglichkeit ist die Ermittlung von aktuell besuchten Seiten. Links zu diesen können einem Anwender visuell anders dargestellt werden [Bru07]. Weiters können historische Navigationsprozesse aufgezeichnet werden und damit die Popularität der einzelnen Links für aktuelle Benutzer dargestellt werden [Bru07]. Diese erhalten damit eine Unterstützung in der Auswahl der nächsten zu besuchenden Seite. Alternativ können auch kollaborative Methoden verwendet werden, um wie in Empfehlungssystemen Bewertungen von anderen Anwendern über eine Seite als Ursprung zur Adaption zu verwenden.

5.5 Adaptive Darstellung von Inhalten

Die angepasste Darstellung von Inhalten dient dazu, einem Anwender nur Informationen von einer Seite zu präsentieren, die für ihn relevant sind. Im Gegensatz zur Filterung von Daten wird bei dieser Technik der Adaption der Inhalt eines Dokumentes betrachtet und entschieden, welche einzelnen Teile für einen Benutzer von Relevanz sind [BCC07]. Dieser kann mittels unterschiedlichen Methoden auf die Informationen aufmerksam gemacht werden, die vom System als bedeutend eingestuft sind. Bei Empfehlungssystemen werden dagegen bereits vor der Darstellung die Daten nach Relevanz sortiert und gefiltert.

5.5.1 Adaptionsmethoden für Inhalte

Um einem Anwender nur Informationen von einem Dokument zu präsentieren, die für ihn relevant sind, gibt es unterschiedliche Ansätze. In allen Fällen muss das entsprechende Benutzerprofil ausgewertet werden, damit evaluiert werden kann, welche Teile einer dargestellten Webseite angezeigt oder hervorgehoben werden sollen und welche nicht [BCC07]. Für die Anpassung gibt es einerseits die Möglichkeit, unterschiedliche Seiten oder Inhalte vorab zur Verfügung zu stellen und aus diesen die passenden Elemente auszuwählen [BCC07]. Weiters können bedeutende Teile direkt im Prozess der Darstellung eruiert werden. Die unterschiedlichen

Ansätze haben weiters Einfluss auf die Organisation der zur Verfügung stehenden Daten [BCC07].

5.5.1.1 Multiple vorgefertigte Seiten oder Fragmente

Die einfachste Variante, um Inhalte einer Seite dynamisch an Benutzerbedürfnisse anzupassen, ist die Verwendung unterschiedlicher vorgefertigter Seiten. Diese einzelnen Variationen werden für die Anzeige zur Verfügung gestellt [BCC07]. Basierend auf den Informationen im Benutzerprofil eines Anwenders wird zur Laufzeit die zutreffende Version ausgewählt [BCC07]. Diese wird anschließend präsentiert. Der klare Nachteil bei dieser Methode ist, dass die Anzahl der benötigten Seiten von der Menge der Daten im Informationssystem sowie der unterschiedlich gewünschten Ausprägung abhängt [BCC07]. Wenn viele einzelne Teile einer Seite angepasst werden müssen, wird dies sehr schnell zu aufwändig. Am sinnvollsten ist die Anwendung bei einer geringen Anzahl an Profilausprägungen (Stereotypen), sowie einer überschaubaren Menge von verfügbaren Inhaltsseiten.

Die Idee von unterschiedlichen, fertigen Seiten kann auch innerhalb eines Dokuemnts angewendet werden. In diesem Fall ist eine Webseite aus einzelnen Fragmenten zusammengesetzt. Diese sind so aufgeteilt, dass sie sich in der Art der dargestellten Information unterscheiden [BCC07]. Die Fragmente müssen ebenso wie bei der Variante mit unterschiedlichen Seiten vorab erstellt werden. Für die Auswahl von Fragmenten für eine Seite gibt es zwei Möglichkeiten. Zum einen besteht die Möglichkeit der Verwendung von optionalen Fragmenten [BCC07]. In dieser Variante wird eine Webseite in eine gewisse Anzahl von Fragmenten aufgeteilt, welche Bedienungen enthalten [BCC07]. Bei der Präsentation werden entsprechende Fragmente für die zur Verfügung stehenden Platzhalter ausgewählt [BCC07]. Dabei können an entsprechende Stellen nur Fragmente verwendet werden, welche den definierten Voraussetzungen entsprechen. Alternativ können veränderbare Fragmente verwendet werden [BCC07]. In diesem Fall gibt es für die einzelnen Teile einer Webseite unterschiedliche, passende Fragmente [BCC07]. Aus diesen wird zur Laufzeit jenes ausgewählt, welches am besten auf die aktuellen Benutzereigenschaften und den Kontext passt [BCC07]. Dies kann etwa für Menschen mit besonderen Bedürfnissen verwendet werden, um Inhalte in anderen Formaten darzustellen. Im Gegensatz zur Auswahl ganzer Seiten besteht bei der Verwendung von einzelnen Seitenteilen ein höherer Aufwand bei der Selektion der geeigneten Fragmente [BCC07]. Weiters kann es schwer sein, eine ganze Seite

in sich konsistent und kohärent darzustellen. Der Vorteil ist jedoch, dass mit einer überschaubaren Menge von vorhandenen Inhaltsteilen eine große Anzahl an Seiten erstellt werden kann.

In einem Informationssystem zur Vermittlung von Unterkünften kann die Information auf Seiten, welche Übernachtungsmöglichkeiten beschreiben, üblicherweise gut in voneinander abgegrenzte Fragmente eingeteilt werden, wenn diese die selbe Struktur aufweisen. Falls nur freie Texte zur Darstellung verwendet werden, ohne eine gewisse Strukturierung in eine festgelegte Form, ist die Anwendung dieser Methoden nicht mehr so einfach. Unter der Annahme, dass Freizeitaktivitäten immer als separater Block präsentiert werden, können diese für Anwender hervorgehoben werden, falls Interesse besteht.

5.5.1.2 Alternative Varianten basierend auf Informationskonzepten

Im Gegensatz zu fertigen Varianten von Seiten oder Fragmenten kann vorhandene Information auch dynamisch in Elemente eingeteilt werden. Diese Elemente müssen für die Darstellung passend zum aktuellen Kontext ausgewählt und anschließend strukturiert werden, was auch als Inhaltsplanung bezeichnet wird [BCC07]. Weiters werden Methoden aus der natürlichen Spracherzeugung benötigt, damit einzelne Konzepte zu Texten verknüpft werden können, die keine Vermutung auf automatische Verarbeitung zulassen. Dies erlaubt die Adaption ohne eine vorgegebene Menge an unterschiedlichen Texten zu benötigen [BCC07]. Damit wird einem Betreiber Aufwand erspart und es ist eine Skalierung auf große Informationsmengen gegeben.

Bei der Auswahl des Inhaltes werden entsprechend relevante Informationen aus der Domäne identifiziert. Hierzu wird für jedes Inhaltselement berechnet, zu welchem Grad es für die aktuelle Situation von Bedeutung ist [BCC07]. Dabei werden Daten aus dem Benutzerprofil eines Anwenders verwendet, um zu entscheiden, ob das entsprechende Element für ihn von Interesse ist. In einem Tourismussystem können etwa alle Eigenschaften einer Unterkunft als eigenes Element oder Fakt betrachtet werden. Aus diesen wird eine bestimmte Anzahl ausgewählt, welche dem Benutzer präsentiert werden sollen. Somit werden ihm nur die Informationen angezeigt, die er sehen möchte. Geht aus einem Profil etwa hervor, dass ein Benutzer an Freizeitaktivitäten interessiert ist, können alle Inhalte des Dokumentes, welche als solche klassifiziert sind, als darzustellende Informationen ausgewählt werden.

Nachdem die Auswahl jener Elemente erfolgt ist, welche dargestellt werden sollen, muss eine geeignete Struktur ermittelt werden [BCC07]. Dies dient dazu, dass bestimmte Elemente, die nur gemeinsam dargestellt werden, nicht an verschiedenen Stellen der Webseite präsentiert werden [BCC07]. Weiters wird die Reihenfolge der Elemente auf der Seite bestimmt. Hierbei können etwa die einzelnen Teile entsprechend ihrer Relevanz sortiert werden [BCC07]. Zusätzlich besteht die Möglichkeit, durch geeignete Techniken Blöcke hervorzuheben, um den Benutzer darauf aufmerksam zu machen [BCC07]. Einem Anwender, welchem etwa Freizeitaktivitäten wichtiger sind als zusätzliche Raumausstattung in einer Unterkunft, kann diese Eigenschaften weiter oben auf der Seite angezeigt bekommen. Zusätzlich kann der Text beispielsweise in anderer Farbe dargestellt werden.

5.5.2 Darstellungstechniken

Durch unterschiedliche Möglichkeiten zur Darstellung von Informationen können einzelne Seitenelemente hervorgehoben werden, so dass diese vom Benutzer schneller wahrgenommen werden [BCC07]. Bei Methoden, welche sich auf die Relevanz von Inhalten beziehen, wird der Anwender entlastet, in dem er nicht die komplette Information aufnehmen muss [KKP01]. Der für ihn relevante Inhalt wird anders dargestellt, als nicht relevante Informationen. Weiters gibt es die Möglichkeit zur Präsentation von Daten durch unterschiedliche Formate, um etwa besondere Bedürfnisse von Anwendern zu berücksichtigen [KKP01].

In der Regel werden vorgefertigte Elemente verwendet, um diese nach den Bedürfnissen des Benutzers anzuzeigen [BCC07]. Dabei gibt es zwei unterschiedliche Arten, nach denen die Techniken klassifiziert werden können. Jene, die sich nach dem Fokus orientieren, zeigen unter Umständen nicht die vollständigen Informationen eines Dokumentes an, während andere, die sich nach dem Kontext orientieren, darauf bedacht sind, alle Elemente zu präsentieren [BCC07]. Bei der ersten Variante können eventuell wichtige Inhalte verborgen bleiben wohingegen bei der zweiten Kategorie die Aufmerksamkeit des Benutzers auf die wesentlichen Elemente schwinden kann [BCC07].

5.5.2.1 Fokus-orientierte Möglichkeiten

Bei fokus-orientierten Methoden werden dem Anwendern nur Inhalte dargestellt, an denen er laut seinem Benutzerprofil interessiert ist. Des weiteren gibt es für ihn keine Möglichkeit, die restlichen Informationen eines Dokumentes einzusehen [BCC07]. Zu dieser Variante zählen die Seiten- beziehungsweise Fragmentauswahl, sowie alle Methoden zur Behandlung von Informationselementen, wie sie bereits beschrieben wurden [BCC07]. Ein Nachteil solcher Methoden ist, dass Elemente vom Benutzer verborgen bleiben können, welche dieser für relevant hält. Aus diesem Grund ist der Mechanismus zur Adaption sowie ein aktuelles Benutzerprofil von großer Bedeutung [BCC07]. Falls der Prozess nicht zuverlässig funktioniert, hat der Benutzer keine Möglichkeit, an die gewünschten Informationen zu gelangen [BCC07]. Des weiteren gibt es für den Anwender keine Möglichkeit zur Steuerung über die angezeigten Elemente [BCC07]. Lediglich Einstellungen betreffend des Profils können sich auf die Auswahl der anzuzeigenden Daten auswirken.

Dieses Art der Anpassung wird vorwiegend in adaptiven Lernsystemen verwendet. In diesen ist es einfacher, ein genaueres Abbild von Benutzern zu erstellen. [BCC07]

5.5.2.2 Kontext-orientierte Anpassungen

Bei den kontext-orientierten Möglichkeiten zur Adaption wird darauf Wert gelegt, dem Anwender alle verfügbaren Informationen eines Dokumentes zu präsentieren [BCC07]. Um die für ihn wichtigen Elemente hervorzuheben, werden Änderungen in der Darstellungsform angewandt. Dadurch soll erreicht werden, dass ein Benutzer die für ihn relevanten Informationen rasch auffinden kann, aber dennoch Zugang zu weiteren Inhalten hat, falls er diese benötigt. Hierzu gibt es folgende Techniken [BCC07] [KKP01].

- Zusammenklappbare Textelemente können für den Benutzer relevante Informationsblöcke aufgeklappt darstellen. Alle nicht relevanten Inhalte werden zugeklappt präsentiert. Eine prägnante Überschrift oder Ähnliches gibt einen Hinweis auf die dahinter verborgenen Daten. Bei Bedarf kann der Anwender also diese für ihn als sekundär wichtigen Informationen betrachten, indem er auf die Überschrift klickt.

- Durch unterschiedliche Farbstärke der selben Farbe kann ein Dimmeffekt angewandt werden. Hierbei können wichtige Elemente etwa in schwarz dargestellt werden. Je nach Relevanz werden unterschiedliche Grautöne verwendet, um die restliche Information zu präsentieren. Dadurch ist das gesamte Dokumente sichtbar und ein Benutzer soll auf die stark saturierten Datenblöcke aufmerksam gemacht werden.

- Auf die gleiche Weise können unterschiedliche Farben für die einzelnen Grade an Relevanz eingesetzt werden.

- Durch Sortierung können relevante Blöcke weiter oben auf einer Seite angezeigt werden. Unwichtigere Informationen werden nach unten geschoben.

- Skalierung des Textes kann angewendet werden, um Text in größerer Schrift darzustellen, falls dieser von Bedeutung ist. Durch die Anwendung von mehreren Schriftgrößen können unterschiedliche Grade an Relevanz abgedeckt werden.

Zusammenklappbare Elemente ist die einzige Technik, bei dem der Benutzer nicht die gesamte Information auf Anhieb präsentiert bekommt [BCC07]. Sortierung hingegen besitzt als alleinige Methode die Eigenschaft, dass die vorgegebene Struktur der Elemente nicht erhalten bleibt [BCC07]. Weiters bietet das Verstecken von Inhalten keine Möglichkeit, die Priorität von Informationen zu vermitteln [BCC07]. Dies ist auch für Dimmeffekte der Fall, wenn alle nicht relevanten Daten in gleichem Farbton präsentiert werden. Die unterschiedliche Einfärbung von Text, sowie Sortierung und Skalierung bieten die Möglichkeit der Übertragung von Priorität [BCC07].

All diese Methoden lassen sich auf gleiche Weise in einem Tourismussystem verwenden. Zum Beispiel kann Information über die Raumausstattung verborgen werden, falls aus dem Benutzerprofil darauf geschlossen werden kann, dass dieser Inhalt von geringer Relevanz für einen Anwender ist.

5.5.2.3 Generierung von angepassten Texten

Durch die Generierung von Textelementen kann der eigentlichen Text dynamische an Benutzerbedürfnisse angepasst werden. Hierbei ist ein einfacher Ansatz, spezielle Textstellen anpassbar zu machen [KKP01]. In adaptiven Lernsystemen

kann so je nach Wissensstand eines Anwenders ein Textelement unterschiedlicher Komplexität eingefügt werden [KKP01]. Eine weitere Variante besteht darin, einzelne Wissenselemente mithilfe von Prozessen der natürlichen Sprachgenerierung so zu einem Text zusammenzufügen, dass dieser wie manuell geschriebener Text wahrgenommen wird. Dadurch kann der Grad der Personalisierung sehr stark an einzelne Benutzer angepasst werden. Diese haben das Gefühl, dass speziell auf ihre Bedürfnisse eingegangen wird.

Üblicherweise sind folgende Schritte notwendig, um aus verfügbaren Informationen individuelle Texte zu generieren [RD97].

1. Der Inhalt, welcher im erzeugten Text vermittelt werden soll, muss ermittelt werden. Hierbei werden Nachrichten erstellt, die von der Applikation abhängig sind und Daten zusammenfassen, welche gemeinsam im Text vorkommen sollen. Das Ergebnis ist ein eine formale Darstellung der Informationen, so dass diese in den folgenden Schritten verwendet werden können. Die Daten können dabei in Einheiten, Konzepte und Beziehungen zwischen Einheiten unterteilt werden.

2. Anschließend muss geplant werden, in welcher Reihenfolge und Struktur die verfügbaren Nachrichten im Text vorkommen sollen. Dies erlaubt die Gruppierung von Nachrichten zu Absätzen und weiteren Einheiten. Dadurch soll der Text leicht verständlich präsentiert werden. Weiters soll sichergestellt werden, dass zusammengehörige Elemente auch zusammengehörig dargestellt werden.

3. Als Nächstes können Nachrichten zusammengefasst werden. Dies ist nicht zwingend erforderlich, kann jedoch die Lesbarkeit und den Textfluss positiv beeinflussen. Nachrichten können so etwa in einzelne Sätze verschmolzen werden.

4. Der nächste Schritt ist die Entscheidung über die zu verwendenden Wörter und Textphrasen zum Ausdruck von Einheiten und Konzepten, auch als Lexikalisierung bezeichnet. Dies kann etwa durch vorgefertigte Wörter oder Textteile erfolgen, welche einzelnen Elementen zugeordnet werden. Durch die Verwendung von Listen und Regeln für den Sprachgebrauch kann die Textqualität gesteigert werden. Dadurch kann das System aus unterschiedlichen Alternativen auswählen.

5. Danach muss ermittelt werden, auf welche bereits bekannten Domänenelemtente (Wörter und Phrasen) für die Generierung eines Satzes der Blockes zurückgegriffen werden kann. Dies hängt vom aktuellen Kontext ab. Die Generierung solcher referenzierenden Ausdrücke ist ähnlich zur Lexikalisierung. Im Unterschied dazu wird jedoch diskriminierend entschieden, welche Phrasen passend sind.

6. Der letzte Schritt besteht darin, den fertigen Text unter Berücksichtigung von grammatischen Regeln zu generieren. Dazu zählen etwa die Verwendung von Singular oder Plural sowie die Groß- und Kleinschreibung von Wörtern.

In den meisten Systemen zur Erstellung von individualisierten Texten werden mehrerer dieser Schritte zusammengefasst, so dass diese in drei Phasen erledigt werden. In der Textplanung wird der zu verwendende Inhalt spezifiziert, sowie die Struktur und Sortierung eruiert. Diese Schritte werden zusammengefasst, weil sie häufig schlecht voneinander zu trennen sind. Anschließend werden Sätze geplant, in dem die Aggregation von Nachrichten, die Lexikalisierung und die Erstellung von referenzierenden Ausdrücken durchgeführt wird. Den Abschluss bildet die linguistische Realisierung. [RD97]

Mithilfe dieser Technik können etwa Beschreibungstexte von Unterkünften für Benutzer individuell generiert werden. Dies ist jedoch mit zusätzlichem Aufwand verbunden, wie durch die Erläuterung der einzelnen Schritte verdeutlicht wurde.

5.5.2.4 Anpassung von Medien

Durch die Darstellung von Inhalten in unterschiedlichen Formen können zum einen besondere Bedürfnisse von Benutzern berücksichtigt werden. Dabei können unterschiedliche Gründe für die Vorliebe eines bestimmten Formates vorhanden sein. Zum Beispiel können blinde Menschen aus Bildern keine Information gewinnen. Es muss ein beschreibender Text vorliegen, der von spezieller Software in Sprache umgewandelt werden kann. Weiters sind manche Informationen besser in einem spezifischen Format dargestellt [BCC07]. Der Kontext eines Anwenders kann ebenfalls entscheidend für die Auswahl eines Mediums sein [BCC07]. So kann ein Benutzer etwa durch Vibration des Mobilgerätes auf Ereignisse aufmerksam gemacht werden, wenn der Helligkeitssensor vermuten lässt, dass sich das Gerät in einer Tasche befindet. Eine weitere Entscheidungsquelle sind techni-

sche Ressourcen [BCC07]. In einem Mobilfunknetz können Videos eventuell nur sehr langsam an Endgeräte übertragen werden. Deshalb könnten diese bei geringen zur Verfügung stehenden Bandbreiten durch eine Serie von Bildern ersetzt werden.

Die adaptive Darstellung durch unterschiedliche Medien lässt sich in zwei Ansätze unterteilen [BCC07].

- Die Anpassung kann basierend auf vorher festgelegten Regeln erfolgen. Diese Art wird mehrheitlich angewandt.

- Alternativ kann die unterschiedliche Verwendung von Medien als Optimierungsproblem angesehen werden. Hierbei werden unterschiedliche Metriken definiert, aufgrund welcher die passenden Medien ausgewählt werden. Der Vorteil hierbei ist, dass keine umfangreichen Sätze an Regeln definiert werden müssen.

5.6 Verwendung der Methoden in Tourismussystemen

Aufgrund der Eigenschaften von kommerziellen web-basierten Informationssystemen, die im Bereich Tourismus angesiedelt sind (Vermittlung von Unterkünften), eigenen sich besonders die Techniken für personalisierte Suchen sowie Empfehlungssysteme für den Einsatz zur adaptiven Auswahl und Darstellung von Inhalten. Das Hauptziel von Anwendern in solchen Webapplikationen ist üblicherweise, aufgrund von gewissen Vorgaben, passende Unterkünfte zu finden. Je nach Eigenschaften und Charakter der Benutzer können dabei mehr oder weniger spezifische Anfragen an die Suche gestellt werden. Durch die Erweiterung dieser Anfragen um Daten aus dem Benutzerprofil oder Filterung basierend auf Vorlieben des Anwenders können diese Informationen für die Ermittlung von personalisierten Ergebnissen verwendet werden. Weiters können etwa Impulskäufer mit Vorschlägen aus dem Empfehlungssystem dazu gebracht werden, Objekte zu buchen, nach welchen sie eigentlich gar nicht suchen wollten. Diese Funktion kann auch helfen, falls die Charakteristiken von Anwendern darauf schließen lassen, dass diese in ihrer Entscheidungsfindung generell unschlüssig sind. Durch Empfehlungen kön-

nen sie aus mehreren Objekten auswählen und somit Impulse für spätere Suchen erhalten.

Weiters können angepasste Darstellungsformen der Inhalte zum Abschluss von Transaktionen beitragen. Anwender, welche die für sie wichtigen Informationen priorisiert vorfinden, können anhand dieser direkt Entscheidungen über die Buchung treffen, ohne dabei die gesamte Webseite nach diesen Daten durchforsten zu müssen. Weiters ist es von großem Vorteil, die Präsentation von Information an besondere Bedürfnisse anzupassen. Durch die Änderung des Mediums kann für viele Benutzer die Verständlichkeit erhöht oder überhaupt erst ermöglicht werden. Die einfachste Form wäre die Ersetzung von visuellen Inhalten durch Textpassagen. Blinde Anwender können diese Information durch Software verarbeiten lassen. Diese wandelt den Text in Sprache um, so dass der Informationsgehalt verständlich wird. Diese Methode kann auch verwendet werden, wenn von Personen bekannt ist, ob sie Daten lieber in textueller oder visueller Form aufnehmen.

Die Anpassung der Struktur ist in solchen Systemen wahrscheinlich eher von untergeordneter Bedeutung und der damit verbundene Aufwand nicht gerechtfertigt. Üblicherweise sind die wichtigsten Funktionen wie die Suche und Empfehlungen über Unterkünfte direkt auf der Startseite der Webapplikation verfügbar. Auch Links zu Sonderangeboten sind üblicherweise direkt auf dieser Seite beworben. Eventuell könnten die einzelnen Kategorien, in welche Übernachtungsmöglichkeiten eingeteilt werden, an die gespeicherten Eigenschaften im Benutzerprofil angepasst werden. Um den Anwender jedoch nicht zwischen Interaktionen zu verwirren, sollte nur eine einmalige Berechnung der Sortierreihenfolge stattfinden, welche anschließend über einen längeren Zeitraum verwendet wird. Weiters muss hier abgewogen werden, ob sich der Aufwand für die Berechnung lohnt.

Kapitel 6

Vergleich von Personalisierungsmaßnahmen in Produktivsystemen

In folgendem Abschnitt werden Methoden der Personalisierung in bekannten webbasierten Informationssystemen erläutert. Dabei werden die Unternehmen Amazon.com[1], Google[2], Netflix[3] sowie Yahoo![4] behandelt. Aufgrund der spärlich zur Verfügung stehenden Informationen ist kein direkter Vergleich zwischen den Systemen möglich. Deshalb werden diese getrennt voneinander betrachtet. Die verwendeten Techniken werden dabei auf Basis der verfügbaren Daten mehr oder weniger ausführlich behandelt. Weiters können sich die Anwendungen in Anzahl und Art der Personalisierungsmaßnahmen unterscheiden. Die genannten Unternehmen wurden gewählt, weil für den Tourismusbereich keine relevante Literatur verfügbar ist.

6.1 Amazon.com

Amazon.com, eine Webapplikation zum Verkauf von Büchern, startete 1995 [Fun16a]. Ende der 1990er Jahre wurde das Sortiment auf CDs erweitert [Fun16a]. Kurze Zeit später folgten diverse Elektronikprodukte [Fun16a]. Heute bietet Amazon.com ein breit gefächertes Angebot an Büchern, CDs, Elektronikprodukten,

[1] http://www.amazon.com
[2] http://www.google.com
[3] http://www.netflix.com
[4] http://www.yahoo.com

Spielzeug, Haushaltsmittel, Kleidung und vielem mehr [Fun16a]. Hauptsächlich verwendet Amazon.com Empfehlungssysteme, um seinen Kunden Vorschläge über Produkte zu geben, welche für sie von Interesse sein könnten [LSY03]. Weiters wird dem Kunden die Möglichkeit gegeben, Empfehlungen zu bewerten [LSY03]. Ähnlich zu Produkten, welche in traditionellen Geschäften nahe von Kassen platziert sind, liefert Amazon.com Empfehlungen basierend auf Objekten im Einkaufswagen eines Benutzers [LSY03]. Das bedeutet, die Produkte, welche Impulskäufer ansprechen sollen, werden speziell an deren aktuelle Bedürfnisse angepasst. Neben der Verwendung auf Webseiten werden Empfehlungen auch via E-Mail an Anwender versendet [LSY03].

Der Algorithmus für Empfehlungen, welcher von Amazon.com verwendet wird, basiert auf den Ansätzen der kollaborativen Filterung. Dabei ist dieser laut eigenen Angaben für Vorschläge in Echtzeit geeignet, skalierbar auf massive Mengen an Daten und erzeugt qualitativ hochwertige Empfehlungen [LSY03]. Bei der von Amazon.com verwendeten Methode werden zu Artikeln, welche von Anwendern gekauft oder bewertet wurden, ähnliche Produkte gesucht [LSY03]. Diese werden dann in einer Empfehlungsliste dem Benutzer präsentiert [LSY03]. Anstelle dabei für jeden Artikel die Ähnlichkeit mit jedem anderen zu berechnen, werden die notwendigen Vergleiche eingeschränkt. Da viele Paare von Produkten keine gemeinsamen Kunden besitzen, welche diese gekauft oder bewertet haben, kann die entsprechende Berechnung übersprungen werden [LSY03]. Dies spart neben Rechenleistung auch Speichernutzung [LSY03]. Der grundlegende verwendete Algorithmus ist in Listing **??** dargestellt [LSY03]. Zuerst wird für jeden Artikel ermittelt, welche Kunden diesen gekauft haben. Danach werden alle restlichen Produkte eruiert, welche mit dem Kunden eine Verbindung besitzen [LSY03]. Dabei wird aufgezeichnet, dass ein Anwender beide Artikel gekauft hat [LSY03]. Anschließend werden die Ähnlichkeiten zwischen allen zuvor gemerkten Paaren berechnet [LSY03].

Listing 6.1: Pseudocode zur Berechnung der Ähnlichkeiten von Produkten

```
1 Fuer jedes Element I1 im Produktkatalog
2   Fuer jeden Kunden C, welcher I1 gekauft hat
3     Fuer jedes Element I2, welches von Kunde C gekauft wurde
4       Speichere die Information, dass Kunde C die Elemente I1
            und I2 gekauft hat
5   Fuer jedes Element I2
6     Ermittle die Aehnlichkeit zwischen I1 und I2
```

Die Ähnlichkeiten werden im Hintergrund berechnet, damit sie später für Empfehlungen zur Verfügung stehen. Bei einem Vergleich zwischen allen Produkten wäre der Berechnungsaufwand enorm. Dadurch, dass viele Kunden jedoch nur eine geringe Anzahl an Artikeln kaufen, reduziert sich die Komplexität. Die Effizienz kann weiter gesteigert werden, indem die Menge an Benutzern eingeschränkt wird. Hierbei werden exemplarische Anwender ausgewählt, welche Verkaufsschlager bezogen haben. [LSY03]

Basierend auf den verfügbaren Ähnlichkeiten werden in Echtzeit für einen Benutzer Artikel ausgewählt, welche zu dessen bereits gekauften oder bewerteten Objekten passen. Aus der Liste der ähnlichen Produkte wird eine Liste mit einer gewissen Anzahl, basierend auf dem Vergleich sowie der Popularität der einzelnen Artikel, präsentiert. [LSY03]

Die wichtigste Eigenschaft des Algorithmus ist, dass die teuren Berechnungen von Ähnlichkeiten zwischen Objekten nicht in Echtzeit durchgeführt werden, sondern offline im Hintergrund. Nur die exzellent skalierbare Auswahl jener Produkte, die dem Anwender angezeigt werden, kann bei der Interaktion von Benutzern mit dem System erfolgen. [LSY03]

6.2 Google

Google ging aus der Firma BackRub hervor, welche von Sergey Brin und Larry Page 1997 gegründet und 1998 in Google umbenannt wurde [Fun16b]. Bereits in den ersten wurde Jahren eine Suchmaschine für das Internet angeboten [Fun16b]. Mittlerweile ist Google die wohl bekannteste und am meisten aufgerufene Suche für Inhalte aus dem World Wide Web. Im Laufe der Zeit wurde das Unternehmen um immer mehr Dienstleistungen und Produkte erweitert. Dazu zählen unter anderem der Webbrowser Chrome, das Betriebssystem Android für Mobiltelefone, die Videoplattform YouTube, Google Docs zur Verwaltung von Dokumenten, der Onlinespeicher Google Drive, der E-mail Dienst Gmail sowie zahlreiche andere [Goo16]. Anwender erhalten von Google ein personalisiertes Suchergebnis, welches von unterschiedlichen Faktoren beeinflusst ist [Por14].

Im Jahr 2005 begann Google damit, für angemeldete Benutzer personalisierte Resultate auf Suchanfragen zurückzugeben. Dabei wurden vor allem die historisch durchgeführten Suchen als Referenz herangezogen [Por14]. Seit 2009 ist diese

Funktion auch für Anwender verfügbar, welche kein Benutzerkonto bei Google angelegt haben [Por14]. Mittlerweile haben unterschiedliche Faktoren Einfluss auf das Ergebnis einer Suche. Dazu zählen die folgenden [Por14].

- Der Aufenthaltsort eines Anwenders.
- Vergangene Suchen des Benutzers.
- Der gespeicherte Browserverlauf bei der Verwendung von Google Chrome.
- Soziale Daten aus dem Netzwerk Google+.

Der Aufenthaltsort wird dazu verwendet, um interessante Lokalitäten in unmittelbarer Nähe zu präsentieren. Hierzu werden Ergebnisse bei ortsbezogenen Anfragen (wie etwa die Suche nach Restaurants) so gereiht, dass lokale Betriebe und Dienstleister weiter oben in der Liste vorkommen [Por14]. Weiters wird diese Funktion auch im Betriebssystem Android auf Mobiltelefonen verwendet, um Empfehlungen basierend auf dem Standort zu Sehenswürdigkeiten sowie anderen Dienstleistungen abzugeben [Por14]. Neben der Darstellung in natürlicher Weise innerhalb der Ergebnisliste gibt es auch einen eigenen Bereich, in welchem nur lokale Unternehmen angezeigt werden [Por14]. Die darin vorkommenden Betriebe zahlen meistens für die Werbung.

Weiters werden historische Suchen für die Auswertung von aktuellen Anfragen verwendet. Hierbei werden verwandte Begriffe, welche mit der Abfrage in Verbindung stehen könnten, in die Suche aufgenommen [Por14]. So etwa werden bei der Suche nach *JavaScript* Textbücher zum Erlernen der Sprache angezeigt, falls in früheren Suchen Google bereits einmal mit der Eingabe *Bücher Programmieren* beauftragt wurde [Por14]. In ähnlicher Weise werden Webseiten ausgewertet, die besucht wurden. Dies geschieht bei der Verwendung des Browsers Google Chrome [Por14]. Dadurch können beim Anwender beliebte Seiten in der Ergebnisliste weiter vorne auftauchen als andere Inhalte [Por14].

Seit der Verfügbarkeit des sozialen Netzwerkes Google+ werden auch Daten von dem dort vorhandenen Profil in Suchen einbezogen. Dies betrifft vor allem Bewertungen zu Produkten, welche Freunde aus dem Netzwerk abgegeben haben. Weiters werden private Inhalte in den Ergebnissen angezeigt. Diese können entweder von den eigenen verwendeten Diensten wie Google+ oder Google Docs stammen, oder auch etwa geteilte Bilder oder Videos von Freunden im sozialen

Netzwerk sein. Dabei werden diese Resultate wie alle anderen in die Ergebnisliste aufgenommen. Lediglich kurze Texte informieren über die Eigenschaften, die zur Aufnahme in das Ergebnis geführt haben und dass es sich um privaten Inhalt handelt. Dies bedeutet jedoch nicht, dass diese Informationen auf einmal öffentlich zugänglich sind. [Sul12]

6.3 Netflix

Netflix wurde 1997 als DVD Verkaufs- sowie Verleihservice gestartet [Fun16c]. Dabei wurden alle Transaktionen online abgewickelt [Fun16c]. Die DVDs wurden via Post zum Kunden und wieder zurück zu Netflix transportiert [Fun16c]. Weiters wurden Abonnements angeboten, bei denen zu fixen Preisen pro Monat eine gewisse Anzahl an DVDs ausgeliehen werden durfte [Fun16c]. Aufgrund der beträchtlichen Fortschritte im Bereich des Internets wurde das Angebot 2007 um einen Onlinedienst für Streams erweitert [Fun16c]. Dadurch ist es Kunden möglich, zu jeder beliebigen Zeit und ohne Verzögerung angebotene Filme sowie Fernsehserien zu konsumieren. Mittlerweile ist Netflix mit diesem Angebot in über 190 Ländern vertreten [Net16]. Netflix setzt Empfehlungsmethoden ein, um seinen Anwendern Inhalte zu präsentieren, welche für diese relevant sind.

Auf der Startseite des Portals zum Streamen von Videos werden einem Benutzer unterschiedliche Kategorien mit Titeln präsentiert [Ama12]. Diese Kategorien werden passend zum Benutzerprofil ausgewählt, ebenso erfolgt die Auswahl der darin enthaltenen Titel und deren Reihung innerhalb der Gruppe [Ama12]. Hierbei wird jedoch nicht nur eine einzelne Person berücksichtigt. In einem Konto können mehrere Benutzer angelegt werden, welche die angebotenen Services nutzen können. Die Empfehlungen auf der Startseite sind so ausgelegt, dass sie auch Vorschläge für andere Anwender in der Familie beinhalten [Ama12]. Vorschläge werden immer mit einer Begründung versehen. Aus dieser ist ersichtlich, warum der entsprechende Film oder die Serie vorgeschlagen wurde [Ama12]. Dadurch soll Transparenz verdeutlicht und das Vertrauen des Benutzers gesteigert werden [Ama12]. Dieser soll dazu animiert werden, von ihm angesehene Titel zu bewerten [Ama12].

Für die Empfehlung von Titeln werden vorhergesagte Bewertungen mit der Beliebtheit einzelner Objekte kombiniert [Ama12]. Dies stellt sicher, dass einerseits nicht nur Nischentitel vorgeschlagen werden (basierend auf Bewertungen alleine)

und andererseits nicht zu viele populäre Filme oder Serien vorgeschlagen werden (basierend auf Beliebtheit alleine) [Ama12]. Diese grundlegenden Informationen werden mit weiteren Daten angereichert. Dazu zählen unter anderem die folgenden [Ama12].

- Bewertungen von Anwendern.
- Listen von Titeln, welche Anwender beabsichtigen anzusehen, sowie Suchbegriffe.
- Metadaten von Filmen sowie Fernsehserien.
- Historische Empfehlungen sowie deren Darstellung und die entsprechende Benutzerreaktion.
- Soziale Daten aus verbundenen Netzwerken.
- Externe Informationen wie Kritiken oder Umsätze einzelner Titel.
- Demographische Daten, Aufenthaltsort sowie Sprache.

Um aus diesen Informationen Empfehlungen zu generieren, werden unterschiedlichste Ansätze aus dem Bereich maschinelles Lernen eingesetzt. Dabei werden sowohl Techniken aus dem nicht überwachten Lernen wie auch überwachten Lernen verwendet. Beispiele sind Clustering, lineare Regression, Assoziationsregeln und die Faktorisierung von Matrizen. Der Grund für die große Menge an unterschiedlichen Methoden ist, dass gewisse Techniken besser für spezielle Aufgaben geeignet sind als andere. Durch den Einsatz mehrerer Algorithmen soll die Qualität gesteigert werden. [Ama12]

Einzelne Teile des Empfehlungsprozesses können zu unterschiedlichen Zeiten ausgeführt werden. Dabei werden Algorithmen für das maschinelle Lernen offline trainiert. Weiters werden andere, zeitintensive Berechnungen im Hintergrund durchgeführt. Empfehlungen selbst werden in Echtzeit zur Laufzeit auf Grundlage des aktuellen Kontextes ermittelt. Hierbei können Daten aus vorhergehenden, offline ausgeführten Berechnungen miteinbezogen werden. Weiters gibt es eine Komponente, die fast in Echtzeit agiert. Diese kümmert sich darum, dass Interaktionen von Benutzern aufgezeichnet werden. Die daraus gewonnenen Daten können zusammengefasst und für vorhandene Algorithmen verwendet werden. [AB16]

6.4 Yahoo!

Yahoo! wurde 1994 als Jerry's Guide to the World Wide Web gegründet [Fun16d]. Die Umbennenung in Yahoo! folgte im selben Jahr [Fun16d]. Das Unternehmen bietet Verknüpfungen zu anderen Webseiten an, die dem Anwender bei der Navigation durch das World Wide Web helfen sollen [Fun16d]. Weitere Dienste sind unter anderem kostenlose E-mail Konten (Yahoo! Mail), Chats zwischen Anwendern (Yahoo! Chat) sowie personalisierte Nachrichten (Yahoo! News) [Fun16d]. Laut dem Unternehmen selbst werden hochgradig personalisierte Erlebnisse für Anwender erstellt, indem diese mit Inhalten verbunden werden, welche für sie am bedeutsamsten sind [Yah16]. Durch Werbungen von andernen Unternehmen, welche den Benutzern vorgeschlagen werden, wird Geld eingenommen [Yah16]. Yahoo! bietet in einigen ihrer Applikationen personalisierten Inhalt basierend auf dem Standort der Anwender. Weiters wird eine persönliche Seite zur Verfügung gestellt, auf welcher der Benutzer für ihn interessante Informationen angezeigt bekommt.

Mit My Yahoo! bietet das Unternehmen eine angepasste Variante von Yahoo! für jeden Benutzer, der aus einer großen Zahl angebotener Module auswählen kann, welche angezeigt werden sollen [MPR00]. Der Inhalt dieser wird automatisch bei Änderungen aktualisiert. Weiters können innerhalb der einzelnen Komponenten meist benutzerindividuelle Anpassungen vorgenommen werden [MPR00]. Bei Fernsehprogrammen kann etwa der Ort einbezogen werden, um Sender entsprechend zu filtern und die Zeiten anzupassen [MPR00]. Das Gleiche gilt für Module mit Inhalten bezüglich Sport. Die dargestellten Mannschaften werden je nach Ort angepasst [MPR00].

Yahoo! Search ist die Suchfunktion im Yahoo! Ökosystem. Bei der Suche werden entsprechend zusätzliche Informationen angezeigt, falls diese bezüglich der eingegebenen Schlagwörter relevant sind [MPR00]. Wird zum Beispiel nach einem aktuellen Kinofilm gesucht, werden ein Pressebild, Schauspieler und ein Link auf die Spielzeiten in Kinos angezeigt [MPR00]. Falls ein Ort vom Benutzer bekannt ist, wird diese ebenfalls bei lokal vorhandene Kinos berücksichtigt [MPR00]. Auf ähnliche Weise werden Suchanfragen behandelt, die auf Produkte oder Dienstleistungen schließen lassen. Es wird in den Ergebnissen automatisch ein Link zu Yahoo! Yellow Pages eingefügt, auf denen passend zum Aufenthaltsort und den Stichwörtern in der Suchabfrage Unternehmen angezeigt werden [MPR00].

Basierend auf dem Ort, in dem sich der Anwender aufhält, kann Yahoo! noch weitere Komponenten anpassen. Darunter fallen etwa das angezeigte Wetter, Neuigkeiten aus der Umgebung, in der nächsten Zeit stattfindende Ereignisse, Verkehrsmeldungen für die unmittelbare Umgebung sowie andere Objekte. Dabei ist die Eingabe eines Ortes nicht zwingend vorgeschrieben. Dem Benutzer wird lediglich die Option gegeben, die damit verbundenen Möglichkeiten zu nutzen. [MPR00]

Kapitel 7

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Personalisierung von webbasierten Informationssystemen. Dabei wird die Vermittlung von Unterkünften als Beispiel herangezogen. Durch die Adaption von Inhalten sowie derer Darstellung kann eine Webapplikation auf persönliche Bedürfnisse und Interessen zugeschnitten werden. Die Vorteile für Anwender sind hierbei gezielteres, schnelleres Auffinden von Informationen sowie die Entdeckung von neuen, bis dato unbekannten Inhalten. Auf Seite des Betreibers bindet man seine Kunden durch deren Zufriedenheit und generiert so potentiell mehr Gewinn. Doch es bestehen auch Nachteile, wie etwa die unerwünschte Sammlung und Auswertung von Benutzerdaten. Weiters müssen in großen Systemen Berechnungen im Hintergrund ausgeführt werden, weil diese in Echtzeit nicht mehr schaffbar sind.

Im ersten Schritt müssen die Eigenschaften der Anwender analysiert und daraus ein Benutzerprofil erstellt werden. Dieses soll möglichst realitätsnah sein und zu jedem Zeitpunkt die entsprechende Person im System widerspiegeln. Deshalb muss eine ständige Aktualisierung erfolgen. Unterschiedliche Identifikationsmethoden dienen zum Erkennen von Anwendern im System. Die gängigste ist hierbei die Verwendung von Benutzername und Passwort. Weiters gibt es verschiedene Möglichkeiten, ein Benutzerprofil darzustellen. Die einfachste Variante stellt dafür ist ein Vektor von Stichwörtern. Durch die Erweiterung auf Konzepte oder die Verwendung von semantischen Netzen wird der Nachteil von Polysemie minimiert und es stehen mehr Möglichkeiten für die Personalisierung zur Verfügung, wie etwa die Darstellung von verwandten Konzepten. Aufgrund der Einfachheit und der geringen Häufigkeit von sprachlichen Hürden eigenen sich Profile basierend auf Stichwörter jedoch ebenfalls gut für Tourismussysteme.

Um Ähnlichkeiten zwischen einem Benutzerprofil und vorhandenen Dokumenten in einem Informationssystem berechnen zu können, müssen beide in vergleichbarer Form vorliegen. Da es sich bei den meisten Inhalten von Webseiten um unstrukturierten Text handelt, muss dieser zuerst aufbereitet werden, so dass eine Repräsentation durch Stichwörter möglich ist. Diese werden anschließend gewichtet, um die Relevanz der einzelnen Wörter im Dokument zu ermitteln. Durch unterschiedliche Methoden der Informationsbeschaffung können die repräsentativen Formen von Benutzerprofil und Dokumenten miteinander verglichen werden. Dadurch wird die Relevanz einzelner Inhalte für den Benutzer ermittelt, welche etwa für die personalisierte Suche oder Empfehlungssysteme verwendet wird. Hierbei ist das gebräuchlichste Modell das sogenannte Vektorraum-Modell. In diesem werden die Stichwortvektoren im Raumn dargestellt und durch die Kosinus-Ähnlichkeit (der Kosinus des Winkel zwischen den Vektoren) die Ähnlichkeit berechnet. Durch Clustering von Dokumenten kann dieser Vorgang optimiert werden, da bei Steigender Anzahl an Webseiten im System sonst die Vergleichsberechnung sehr lange dauern würde.

Die zwei wichtigsten Methoden zur Personalisierung von Inhalten sind die Adaption der Suche sowie Empfehlungssysteme. Bei personalisierten Suchprozessen werden die Ergebnisse, welche auf eine Anfrage zurückgegeben werden, gefiltert. Damit sollen nur für den jeweiligen Anwender relevante Inhalte präsentiert werden. Je nach Zeitpunkt der Verwendung des Benutzerprofils gibt es hierbei drei Ansätze. Entweder die Filterung erfolgt direkt im Prozess, oder es geschieht eine Erneute Sortierung nach der Informationsbeschaffung. Weiters ist es möglich, direkt die Anfrage mit Daten aus dem Benutzerprofil zu erweitern. Bei Empfehlungssystemen werden einem Anwender Inhalte vorgeschlagen, an denen er eventuell interessiert ist, ohne dass dieser danach sucht. Bei inhalts-basierten Systemen werden direkt die Dokumente mit den Eigenschaften des Benutzerprofils verglichen. Bei kollaborativen Systemen werden bereits vorhandene Bewertungen von ähnlichen Benutzern oder Objekten verwendet, um für ein aktuelles Dokument die Relevanz zu einem Anwender zu bestimmen. Die Navigation im Informationssystem kann durch verschiedene Methoden unterstützt werden, welche Links entweder visuell oder funktionell verändern. Weiters können dynamisch Verknüpfungen in Dokumente eingefügt oder gelöscht werden. Hierbei sollte darauf geachtet werden, dass in wichtigen Navigationselementen wie einem Hauptmenü nur Eigenschaften des Benutzerprofils als Grundlage der Adaption verwendet werden, welche über einen längeren Zeitraum stabil sind. Dadurch wird sichergestellt, dass sich zum Beispiel nicht bei jeder neuen Sitzung die Reihenfolge

der Elemente ändert. Unterschiedliche Möglichkeiten zur Darstellung helfen dem Benutzer, für ihn relevanten Inhalt aus einem Dokument schneller zu erfassen. Hierbei wird zwischen fokus-orientierten und kontext-orientierten Methoden unterschieden. Bei ersteren kann Information verborgen bleiben, dafür muss der Anwender nicht den gesamten Dokumenteninhalt aufnehmen. Im Gegensatz dazu wird bei zweiteren Techniken der gesamte Inhalt eines Dokumentes dargestellt und die wichtigen Elemente von unwichtigen hervorgehoben. Des weiteren können Texte individuell für Benutzer generiert und unterschiedliche Medien basierend auf Bedürfnissen von Anwendern oder technischen Einschränkungen verwendet werden. Die wohl wichtigsten Konzepte zur Personalisierung in Systemen zum Angebot von Unterkünften sind die personalisierte Suche sowie Empfehlungssysteme.

Aus der Arbeit geht hervor, dass die Adaption von Inhalten und deren Darstellung keineswegs trivial ist. Der gesamte Prozess muss an die Gegebenheiten der jeweiligen Applikation angepasst werden. Nur wenn die einzelnen Komponenten aufeinander abgestimmt sind, kann eine Bereicherung für Kunden sowie Betreiber erzielt werden. Vor allem die unterschiedlichen Inhalte sowie Benutzerbedürfnisse in der Vielzahl an verfügbaren Informationssystemen bedingt die notwendige Spezialisierung.

Methoden zur Personalisierung sind allgegenwärtig in Webapplikationen eingesetzt. Vor allem die Adaption der Suche sowie Empfehlungen von Inhalten sind weit verbreitet. Aufgrund der Verschlossenheit der meisten Betreiber, ist nur oberflächlich bekannt, welche Maßnahmen getroffen werden. Details wie verwendete Algorithmen oder Ähnliches sind unmöglich zu eruieren. Lediglich Netflix gibt einen etwas tieferen Einblick in Bezug auf die Verwendung von Empfehlungssystemen in deren Streamingportal. Weiters gibt es von Amazon eine kurze Übersicht über den Einsatz von Vorschlägen zu Produkten.

Literatur

[AB16] Xavier Amatriain und Justin Basilico. *System Architectures for Personalization and Recommendation*. 2016. URL: `http://techblog.netflix.com/2013/03/system-architectures-for.html` [besucht am: 9. Feb. 2016].

[Ama12] Xavier Amatriain. „Mining large streams of user data for personalized recommendations“. In: *SIGKDD Explorations* 14.2 (2012), S. 37–48.

[BCC07] Andrea Bunt, Giuseppe Carenini und Cristina Conati. „Adaptive Content Presentation for the Web“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 409–432.

[BM07] Peter Brusilovsky und Eva Millán. „User Models for Adaptive Hypermedia and Adaptive Educational Systems“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 3–53.

[Bru07] Peter Brusilovsky. „Adaptive Navigation Support“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 263–290.

[Bur02] Robin D. Burke. „Hybrid Recommender Systems: Survey and Experiments“. In: *User Model. User-Adapt. Interact.* 12.4 (2002), S. 331–370.

[Bur07] Robin D. Burke. „Hybrid Web Recommender Systems“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 377–408.

[Chi93] David N. Chin. „Acquiring user models“. In: *Artif. Intell. Rev.* 7.3-4 (1993), S. 185–197.

[CMS09] W. Bruce Croft, Donald Metzler und Trevor Strohman. *Search Engines - Information Retrieval in Practice*. Pearson Education, 2009.

[Fox+92] Edward A. Fox u. a. „Extended Boolean Models“. In: *Information Retrieval: Data Structures & Algorithms*. 1992, S. 393–418.

[Fun16a] FundingUniverse. *Amazon.com, Inc. History*. 2016. URL: http://www.fundinguniverse.com/company-histories/amazon-com-inc-history/ [besucht am: 9. Feb. 2016].

[Fun16b] FundingUniverse. *Google, Inc. History*. 2016. URL: http://www.fundinguniverse.com/company-histories/google-inc-history/ [besucht am: 9. Feb. 2016].

[Fun16c] FundingUniverse. *Netflix, Inc. History*. 2016. URL: http://www.fundinguniverse.com/company-histories/netflix-inc-history/ [besucht am: 9. Feb. 2016].

[Fun16d] FundingUniverse. *Yahoo! Inc. History*. 2016. URL: http://www.fundinguniverse.com/company-histories/yahoo-inc-history/ [besucht am: 9. Feb. 2016].

[Gau+07] Susan Gauch u. a. „User Profiles for Personalized Information Access“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 54–89.

[Goo16] Google. *Über Google - Produkte*. 2016. URL: https://www.google.at/intl/de/about/products/ [besucht am: 9. Feb. 2016].

[Jan+10] Dietmar Jannach u. a. *Recommender Systems - An Introduction*. Cambridge University Press, 2010.

[Kay01] Judy Kay. „User modeling for adaptation“. In: *User Interfaces for All, Human Factors Series* (2001), S. 271–294.

[KB11] Yehuda Koren und Robert M. Bell. „Advances in Collaborative Filtering“. In: *Recommender Systems Handbook*. 2011, S. 145–186.

[KKP01] Alfred Kobsa, Jürgen Koenemann und Wolfgang Pohl. „Personalised hypermedia presentation techniques for improving online customer relationships“. In: *Knowledge Eng. Review* 16.2 (2001), S. 111–155.

[KL03] Kevin Keenoy und Mark Levene. „Personalisation of Web Search“. In: *Intelligent Techniques for Web Personalization, IJCAI 2003 Workshop, ITWP 2003, Acapulco, Mexico, August 11, 2003, Revised Selected Papers*. 2003, S. 201–228.

[Koc01] Nora Parcus de Koch. „Software engineering for adaptive hypermedia systems: reference model, modeling techniques and development process“. Diss. Ludwig Maximilians University Munich, 2001. ISBN: 3-87821-318-2.

[LGS11] Pasquale Lops, Marco de Gemmis und Giovanni Semeraro. „Content-based Recommender Systems: State of the Art and Trends“. In: *Recommender Systems Handbook*. 2011, S. 73–105.

[LSY03] Greg Linden, Brent Smith und Jeremy York. „Industry Report: Amazon.com Recommendations: Item-to-Item Collaborative Filtering“. In: *IEEE Distributed Systems Online* 4.1 (2003).

[Mic+07] Alessandro Micarelli u. a. „Personalized Search on the World Wide Web“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 195–230.

[MPR00] Udi Manber, Ash Patel und John Robison. „The business of personalization: experience with personalization of Yahoo!“ In: *Commun. ACM* 43.8 (2000), S. 35–39.

[MR01] San Murugesan und Annamalai Ramanathan. „Web Personalisation - An Overview“. In: *Active Media Technology, 6th International Computer Science Conference, AMT 2001, Hong Kong, China, December 18-20, 2001, Proceedings*. 2001, S. 65–76.

[MRS+08] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze u. a. *Introduction to information retrieval*. Bd. 1. Cambridge university press Cambridge, 2008.

[MSM07] Alessandro Micarelli, Filippo Sciarrone und Mauro Marinilli. „Web Document Modeling“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 155–192.

[Net16] Netflix. *Netflix : Overview*. 2016. URL: `http://ir.netflix.com/` [besucht am: 9. Feb. 2016].

[PB07] Michael J. Pazzani und Daniel Billsus. „Content-Based Recommendation Systems“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 325–341.

[Por14] Portent. *Guide to Personalized Search Results*. 2014. URL: `https://www.portent.com/blog/seo/personalized-search-results.htm` [besucht am: 10. Feb. 2016].

[RD97] Ehud Reiter und Robert Dale. „Building applied natural language generation systems“. In: *Natural Language Engineering* 3.1 (1997), S. 57–87.

[Rek+96] Y. Rekhter u. a. *RFC 1918: Address Allocation for Private Internets*. 1996. URL: `https://tools.ietf.org/html/rfc1918` [besucht am: 14. Feb. 2016].

[RRS11] Francesco Ricci, Lior Rokach und Bracha Shapira. „Introduction to Recommender Systems Handbook“. In: *Recommender Systems Handbook*. 2011, S. 1–35.

[Sch+07] J. Ben Schafer u. a. „Collaborative Filtering Recommender Systems“. In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, S. 291–324.

[SD10] Sergey A. Sosnovsky und Darina Dicheva. „Ontological technologies for user modelling“. In: *IJMSO* 5.1 (2010), S. 32–71.

[SFS05] Catharina M. Serino, Christopher P. Furner und Cindi Smatt. „Making it Personal: How Personalization Affects Trust Over Time“. In: *38th Hawaii International Conference on System Sciences (HICSS-38 2005), CD-ROM / Abstracts Proceedings, 3-6 January 2005, Big Island, HI, USA*. 2005.

[Sul12] Danny Sullivan. *Google's Results Get More Personal With 'Search Plus Your World'*. 2012. URL: `http://searchengineland.com/googles-results-get-more-personal-with-search-plus-your-world-107285` [besucht am: 10. Feb. 2016].

[Yah16] Yahoo. *Corporate Information - Yahoo.* 2016. URL: `https://info.yahoo.com//` [besucht am: 9. Feb. 2016].

[ZG07] Jie Zhang und Ali A. Ghorbani. „GUMSAWS: A Generic User Modeling Server for Adaptive Web Systems“. In: *Fifth Annual Conference on Communication Networks and Services Research (CNSR 2007), 14-17 May 2006, Fredericton, New Brunswick, Canada*. 2007, S. 117–124.

Abbildungsverzeichnis

Tabellenverzeichnis

Manuel Steiner

Lebenslauf

Holzing 45
3254 Bergland
Österreich
+43 (0) 664 444 89 78
manuel.steiner@mts-international.org

Ausbildung

Nov 2013 – jetzt **Diplomingenieur in Computer Science**, *Johannes Kepler Universität*, Linz.
Hauptrichtung: Intelligent Information Systems

Sep 2014 – Mai 2015 **Master of Science in Computer Science**, *University of Bradford*, Bradford (Großbritannien).
Erasmus+ Studenten Austauschprogramm

Sep 2010 – Okt 2013 **Bachelor of Science in Informatik**, *Johannes Kepler Universität*, Linz.

Sep 2004 – Jun 2009 **Matura**, *IT-HTL*, Ybbs a.d. Donau.
Hauptzweig: Netzwerktechnik

Sep 2000 – Jun 2004 **Bundesrealgymnasium**, Wieselburg.

Sep 1996 – Jun 2000 **Volksschule**, Petzenkirchen.

Masterarbeit

Titel *Personalisierung von web-basierten Informationssystemen und deren Applikation im Tourismus*

Betreuer a.Univ.-Prof. DI Dr. Wolfram Wöß

Beschreibung Unterschiedliche Personalisierungsmethoden zum Einsatz in Web Applikationen werden untersucht. Weiters werden Möglichkeiten zur Profilverwaltung von Benutzern eruiert

Bachelorarbeit

Titel *Zeit-basierte Synchronisierung für Datenbanken*

Betreuer a.Univ.-Prof. DI Dr. Wolfram Wöß

Beschreibung Unterschiedliche Methoden für Synchronisierung von Datenbanken werden evaluiert. Weiters wird ein eigener Mechanismus implementiert

Arbeitserfahrung

Mär 2014 – jetzt **Techniker, Systemadministrator**, *MTS Management Technik Systeme GmbH & CO KG*, Wieselburg.
Konfigurierung und Installation von Prozessleitsystemen, IT Administration und Support

Jul 2014 – Aug 2014
Jul 2011 – Aug 2011
Aug 2010 – Jan 2011 **Behindertenbetreuung**, *Lebenshilfe Niederösterreich*, Mühling.
Betreuung von Personen mit mentalen Beeinträchtigungen

Nov 2009 – Jul 2010 **Zivildienst**, *Lebenshilfe Niederösterreich*, Mühling.
Betreuung von Personen mit mentalen Beeinträchtigungen

Jul 2009 – Okt 2009	**Softwareentwickler, Systemadministrator**, *MTS Management Technik Systeme GmbH & CO KG*, Wieselburg. Entwicklung einer Projektmanagementsoftware, Entwicklung einer Verwaltungssoftware für Psychotherapeuten und Sozialberater, IT Administration
Jul 2008	**Softwareentwicklung Praktikum**, *Voglhuber GmbH*, Petzenkirchen. Praktikum in der Softwareentwicklung (Kontrollsoftware für verteilte Anzeigen)
Jul 2007	**Hardwareentwicklung Praktikum**, *Microtronics Engineering GmbH*, Mank.
Jul 2006 – Aug 2006	Praktikum in der Hardwareentwicklung (verteile Sensornetzwerke)
Jul 2005	**Nachbetreuung Praktikum**, *MTS Management Technik Systeme GmbH & CO KG*, Wieselburg. Praktikum in der Nachbetreuung von Prozessleitsystemen

Sprachen

Deutsch	**Muttersprache**	
Englisch	**Flüssig**	*IELTS 8.5, British Council, 2016*

EDV Erfahrung

Grundlagen	C++, Assembler (Intel x86), JavaScript, System C
Fortgeschritten	UML, HTML, CSS, Python, C für Mikrocontroller
Experte	C, Java, SQL
Verschiedenes	Linux, Windows, Office

Qualitäten

Persönliche Fähigkeiten	Starke analytische Fähigkeiten, starke Fähigkeiten unter minimaler Aufsicht zu arbeiten, gute Fähigkeiten in der Softwareentwicklung, gute Kenntnisse im Bereich intelligenter Informationssysteme und Sicherheit in Computersystemen, aufgeschlossen zur Erfahrungssammlung in anderen Bereichen
Erfahrung mit	Softwareentwicklung in Java mit SQL Datenbankanbindung, Softwareentwicklung für Mikrocontroller, Intelligenten Informationssystemen (Datenintegration, Informationsextraktion und -gewinnung), Sicherheit in Computersystemen (ISO 27000, Ethisches Hacking)
Internationale Erfahrung	für 8 Monate in Bradford (Großbritannien) gelebt

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe. Die vorliegende Masterarbeit ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, am 10. März 2016

Manuel Steiner