Malte B. Blanken

»Wer die Wahl hat, hat die Qual! – Empfehlungsdienste sollen helfen

Insbesondere Online-Shops und Anbieter von Medieninhalten im Internet stellen oftmals ein derart umfangreiches Angebot zur Verfügung, welches der normale Nutzer nicht mehr überblicken kann. Viele solcher Inhalteanbieter, wie Amazon.com und Last.fm, setzen daher vermehrt Empfehlungsdienste ein. Der vorliegende Artikel soll erläutern, wie diese grundsätzlich funktionieren.

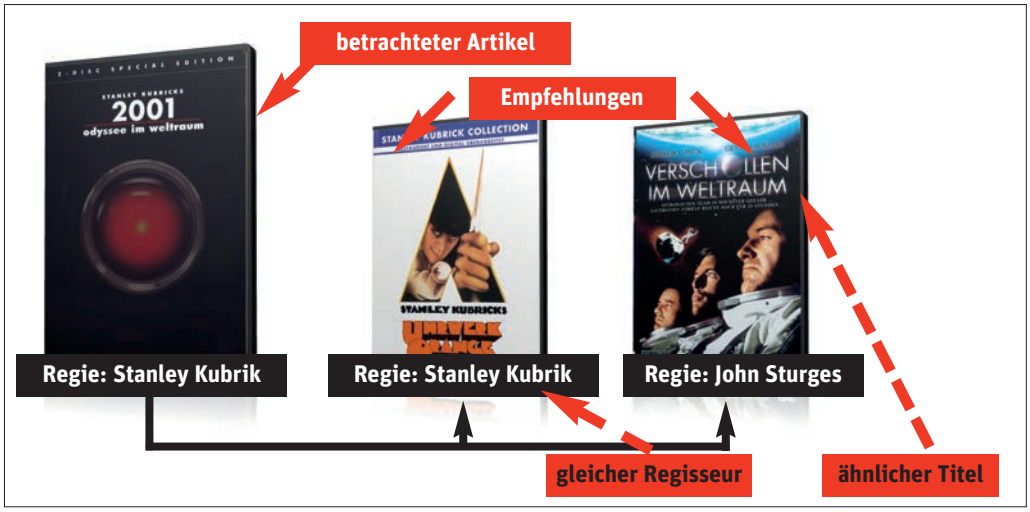


Abb. 1: Inhaltsbasierte Empfehlungen

Was Empfehlungsdienste leisten und wie sie funktionieren, kann gut anhand von Online-Shops wie Amazon.com gezeigt werden. In diesem Bereich finden Empfehlungsdienste zurzeit auch ihr Haupteinsatzgebiet. Sieht sich der Kunde eines solchen Online-Shops einige Artikel an, werden ihm sogleich weitere Artikel vorgeschlagen, welche ihn interessieren könnten.

Abhängig von der Datenbasis, welche der Empfehlungsgenerierung zugrunde liegt, wird ein Empfehlungsdienst als *inhaltsbasiert* (engl. content-based), *kollaborativ* (engl. collaborative) oder *hybrid* bezeichnet.

Bei inhaltsbasierten Verfahren werden die Inhalte auf Ähnlichkeiten hin untersucht. Hat sich beispielsweise ein Kunde eines Online-Shops für den Film *2001 – Odyssee im Weltraum* interessiert, bekommt er weitere Filme von Stanley Kubrik und Filme mit ähnlichem Titel, wie *Verschollen im Weltraum*, empfohlen (siehe Abb. 1).

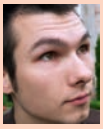
Bei kollaborativen Verfahren werden hingegen nicht die Inhalte, sondern die Anwender miteinander verglichen. Diese Verfahren können ebenfalls anhand von Online-Shops veranschaulicht werden: Haben dort mehrere Kunden einige gleiche Produkte erworben, so werden diese Kunden als „ähnlich" betrachtet. Aus deren jeweilig unterschiedlichen Einkäufen können dann Empfehlungen für die anderen Kunden abgeleitet werden (siehe Abb. 2).

Bei hybriden Verfahren werden inhaltsbasierte und kollaborative Methode miteinander kombiniert. Es werden also sowohl Informationen über die Inhalte als auch über die Nutzer des Systems verwendet.

In jedem Fall werden entweder eine Vorauswahl oder aber alle zur Verfügung stehenden Inhalte – und das können Millionen sein – in eine Reihenfolge gebracht. Die Inhalte bekommen einen Wert zugeordnet, der die jeweilige „Interessantheit" ausdrückt. Anschließend können die „interessantesten" Inhalte als Empfehlungen präsentiert werden.

Da der Prozess der Empfehlungsgenerierung letzten Endes eine Auswahl von Inhalten darstellt, wird in der Literatur auch von inhaltsbasierten

DER AUTOR



Malte B. Blanken ist wissenschaftlicher Mitarbeiter und Lehrbeauftragter an der Hochschule Osnabrück. Zurzeit forscht er dort zum Thema Empfehlungsdienste.

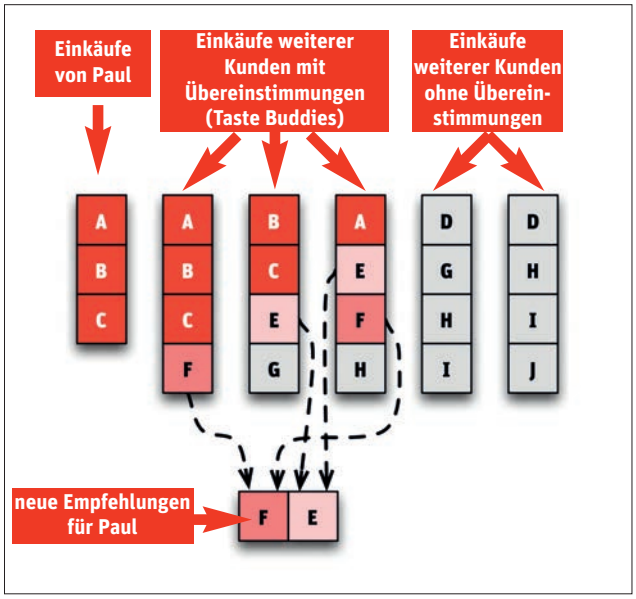


Abb. 2: Kollaborative Empfehlungen

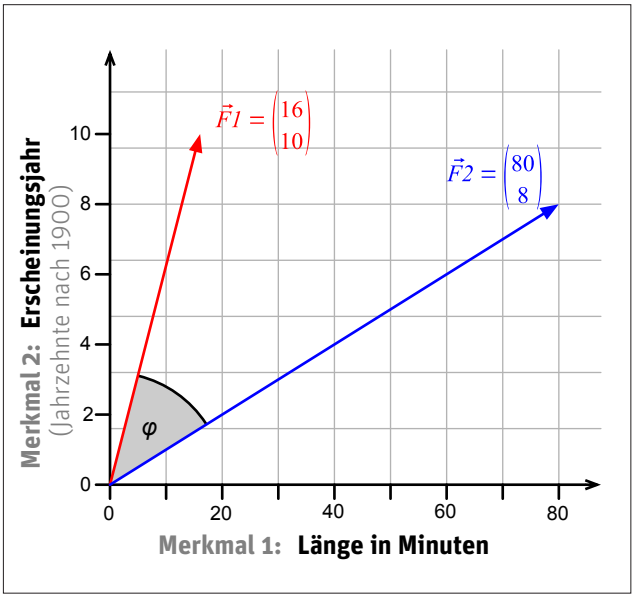


Abb. 3: Merkmalsvektoren zweier Inhalte (a und b) mit zwei Merkmalen und dem eingeschlossenen Winkel φ

und kollaborativen Filtern gesprochen. Diese zwei grundlegenden Filterverfahren sollen in den beiden folgenden Abschnitten etwas genauer betrachtet werden.

Inhaltsbasiertes Filtern

Das inhaltsbasierte Filtern setzt möglichst umfangreiche Kenntnisse über die Merkmale der Inhalte voraus. Bei einem Film könnten dies unter anderem Titel, Regisseur, Genre und Erscheinungsjahr sein. Die Menge dieser Merkmale muss nun mathematisch handhabbar gemacht werden. Typischerweise werden hierzu Vektoren genutzt. Ein Mathematiker würde sagen, dass ein Inhalt mit n Merkmalen auf einen n-dimensionalen Merkmalsvektor in einem n-dimensionalen Merkmalsraum abgebildet wird. Das bedeutet an dieser Stelle nichts anderes, als dass beispielsweise ein Inhalt mit den zwei Merkmalen Erscheinungsjahr und Länge in Minuten (also n = 2) als Punkt in einem zweidimensionalen Koordinatensystem dargestellt werden kann. Dabei repräsentiert die eine Koordinatenachse das Erscheinungsjahr, die andere die Länge in Minuten. Der zu einem gegebenen Punkt gehörige Vektor kann einfach als ein Pfeil vom Koordinatenursprung (also dem Punkt 0;0) zu diesem Punkt verstanden werden. Soll nun also ein Inhalt mit einem anderen verglichen werden, so werden die beiden dazugehörigen Merkmalsvektoren betrachtet. Um schließlich die Ähnlichkeit zu bestimmen, können verschiedene Verfahren der Mathematik herangezogen werden. Ein mögliches Maß für die Ähnlichkeit zweier Inhalte ist der Winkel zwischen ihren Merkmalsvektoren (siehe Abb. 3). Dabei gilt: Je kleiner der Winkel zwischen den Merkmalsvektoren ist, desto ähnlicher sind die Inhalte.

Ein Winkel zwischen zwei Vektoren kann jedoch nur berechnet werden, wenn die beiden Vektoren Elemente des gleichen Merkmalsraumes sind. Oder weniger mathematisch ausgedrückt: Die betrachteten Inhalte müssen die gleichen Merkmale besitzen, müssen vom gleichen „Typ“ sein. Ein Buch kann nicht mit einem Fotoapparat verglichen werden, ein Auto nicht mit einem Gartenstuhl. Und mehr noch: Es bereitet sogar schon Probleme, eine digitale Kamera mit einer analogen zu vergleichen. Die „Ähnlichkeitsberechnung“ funktioniert also nur für Inhalte mit vergleichbaren Merkmalsvektoren. Da diese Berechnung aber ein zentraler Bestandteil des inhaltsbasierten Filterns ist, gilt die gleiche Einschränkung auch für das Filterverfahren selbst.

Oft sind die Inhalte, aus denen Empfehlungen generiert werden sollen, jedoch vom gleichen Typ. Beispiele hierfür sind Empfehlungsdienste in Online-Buchläden und -Videotheken oder auch solche für Fernsehinhalte. Der letztgenannte Anwendungsfall wird derzeit an der Hochschule Osnabrück im Forschungsprojekt Next Generation-PVR (*http://einfach.st/hso*), im welchem der Autor dieses Artikels mitwirkt, untersucht.

Beim inhaltsbasierten Filtern werden üblicherweise die Ähnlichkeiten eines jeden Inhaltes zu jedem anderen ermittelt. Dies bedeutet, dass sehr viele Berechnungen durchgeführt werden müssen. Beispielsweise sind bei 1.000 Inhalten knapp 500.000 Ähnlichkeitswerte zu berechnen, bei 10.000 Inhalten bereits knapp 50 Millionen. Diese Berechnungen müssen allerdings nur durchgeführt werden, wenn neue Inhalte hinzukommen – und das geschieht in vielen Anwendungsfällen, wie einem Online-Shop, relativ selten. Die Ergebnisse werden in einer Datenbank gespeichert. Hat nun der Kunde eines Online-Shops ein Produkt gekauft oder positiv bewertet, können ihm sofort ein paar ähnliche Produkte empfohlen werden. Dazu müssen dann keine weiteren Berechnungen durchgeführt werden – es genügt eine einfache Datenbankabfrage. Diese könnte beispielsweise die sechs ähnlichsten Pro-

dukte (die mit dem größten Ähnlichkeitswert) liefern.

Hat der Kunde bereits mehrere Produkte gleichen Typs gekauft oder bewertet, kann dieses Mehr an Informationen genutzt werden, um die Empfehlungen durch Überlagerung der Empfehlungen zu den einzelnen Produkten zu verbessern.

Kollaboratives Filtern

Im Gegensatz zum inhaltsbasierten werden beim kollaborativen Filtern keinerlei Informationen über die Inhalte benötigt. Es genügt, jeden Inhalt durch eine eindeutige Identifikationsnummer oder Zeichenkette (oder einen Buchstaben, wie in Abb. 2) zu repräsentieren, denn die Auswahl, welche Inhalte einem Nutzer empfohlen werden, ist ausschließlich abhängig von Gemeinsamkeiten mit anderen Nutzern bezüglich ihrer Bewertung von Inhalten. Diese Bewertung erfolgt oftmals implizit, zum Beispiel durch Kaufen eines Produktes oder Anschauen eines Filmes. Das kollaborative Filterverfahren gründet also, wie der Name bereits andeutet, auf der (indirekten) Zusammenarbeit der Nutzer eines Systems.

Um nun den Prozess des kollaborativen Filterns zu erläutern, sei nochmals auf Abb. 2 verwiesen. In der dort dargestellten Situation hat der Kunde mit Namen Paul bereits die Produkte A, B und C gekauft (wobei hier der Kaufvorgang einer positiven Bewertung gleichkommt). Dem System sind fünf weitere Kunden und deren Einkäufe bekannt. Bei drei gibt es eine mehr oder weniger starke Korrelation mit den Einkäufen von Paul. Solche Kunden werden als Taste Buddies bezeichnet, also als „Kumpel" mit ähnlichem Geschmack. Die Produkte, welche zwar von den Taste Buddies gekauft wurden, nicht aber von Paul, sind potenzielle Empfehlungen für Paul. Von diesen werden die am häufigsten vorkommenden Paul präsentiert. Im hier gezeigten Beispiel sind dies F und E.

Um die Qualität der Empfehlungen deutlich zu erhöhen, können unter anderem die folgenden zwei Optimierungen vorgenommen werden:

Erstens sollten die Bewertungen (Einkäufe) besonders ähnlicher Kunden einen größeren Einfluss auf die Empfehlungen haben als die von solchen Kunden, mit denen es nur eine geringe Übereinstimmung gibt.

Zweitens sollten abgestufte Bewertungen genutzt werden. Damit ist gemeint, dass Kunden Produkte beispielsweise mit null bis fünf Sternen explizit bewerten können. Dies ermöglicht ein wesentlich differenzierteres Maß für die Interessensähnlichkeit von Kunden, als die bloße Anzahl übereinstimmender Einkäufe. Daher ist zu erwarten, dass mit dieser Optimierung „bessere" Taste Buddies gefunden werden. Nun können die bisherigen Einkäufe eines Kunden, wie zuvor, mit denen seiner Taste Buddies verglichen und daraus potenzielle Empfehlungskandidaten ermittelt werden. Aus diesen wiederum werden dann diejenigen als Empfehlungen präsentiert, welche von den Taste Buddies besonders gute Bewertungen erhalten haben.

Eine Eigenschaft des kollaborativen Filters ist, dass nur dann gute Empfehlungen generiert werden können, wenn zunächst einmal einige andere Nutzer des Systems mit möglichst ähnlichen Vorlieben gefunden wurden. Es gibt jedoch Situationen, in denen es solche guten Taste Buddies nicht gibt, nämlich dann, wenn ein Nutzer einen seltenen Geschmack, also ungewöhnliche Vorlieben hat. Bei dem in Abb. 2 dargestellten Beispiel könnte dies ein Kunde sein, der bislang nur das Produkt K gekauft hat. Es gäbe dann für diesen Kunden keine Taste Buddies und somit könnten auch keine Empfehlungen generiert werden. Ebenso können auch Produkte, die bislang noch von keinem Kunden gekauft wurden, nicht empfohlen werden. Neue Produkte können daher nicht über kollaborative Empfehlungen bekannt werden. Insgesamt funktionieren kollaborative Empfehlungen erst dann gut, wenn es bereits eine größere Anzahl von Nutzern gibt, die schon eine größere Anzahl von Inhalten bewertet haben. Dies wird auch als „Kaltstartproblem" bezeichnet.

	Pro	Contra
inhaltsbasiert	» Anonymität: Es werden keine Daten über die Nutzer benötigt. » Geschwindigkeit: Empfehlungen können aufgrund vorbereiteter Ähnlichkeitswerte sehr schnell generiert werden.	» Inhalte müssen vom gleichen Typ sein. » Es werden „hochwertige" und möglichst viele Merkmalsdaten benötigt.
kollaborativ	» Beziehungen zwischen Inhalten verschiedenen Typs (z. B. Bücher und Fotokameras) stellen kein Problem dar.	» Relativ große Benutzergemeinde und umfangreiche Informationen zu deren Interessen werden benötigt .

Tabelle 1: Vor- und Nachteile der zwei Filterverfahren

Kombination der zwei Verfahren

Die in den vorigen beiden Abschnitten vorgestellten Filterverfahren haben je ihre Stärken und Schwächen. In Tabelle 1 sind die erwähnten noch einmal zusammengefasst. Es liegt nahe, beide Verfahren zu kombinieren, um die jeweiligen Stärken zu nutzen und die Nachteile zu umgehen. Derartige hybride Verfahren werden ungefähr seit Mitte der 1990er-Jahre entwickelt. Heutzutage arbeiten die Empfehlungsdienste der meisten großen Inhalteanbieter, wie Amazon.com, YouTube und Last.fm, mit einer Art kombiniertem Filterverfahren. Um den konkreten Anforderungen der jeweiligen Anwendungsfälle Rechnung zu tragen, unterscheiden sich deren Verfahren jedoch mitunter deutlich. Der Artikel „Amazon.com recommendations: item-to-item collaborative filtering" von Linden, Smith und York (DOI: 10.1109/MIC.2003.1167344) erläutert beispielsweise, wie bei Amazon.com Empfehlungen generiert werden und wie dabei sowohl gute Performance (hohe Geschwindigkeit) als auch hohe Qualität erreicht werden können.¶