Evaluation adaptiver Systeme und Verhaltenskomplexität

Stephan Weibelzahl, Gerhard Weber
Institut für Psychologie
Pädagogische Hochschule Freiburg,
D-79117 Freiburg
weibelza@uni-freiburg.de, werbergeh@ph-freiburg.de

Zusammenfassung

Für die Evaluation adaptiver Systeme wird Verhaltenskomplexität als objektives und valides Kriterium vorgeschlagen. Adaptivität vereinfacht die Interaktion zwischen Benutzer und System. Dies schlägt sich in einer Reduktion der Verhaltenskomplexität nieder. Für die Berechnung wird das beobachtete Verhalten als Netz von Zuständen und Übergängen repräsentiert. An einem empirischen Datensatz werden vier verschiedene Komplexitätsmaße verglichen, von denen sich zwei als sehr brauchbar erweisen.

Keywords: empirische Evaluation, Evaluationskriterien, Verhaltenskomplexität, Cognitive Task Analysis, Validität

1 Evaluation adaptiver Systeme

Die Evaluation adaptiver Systeme wurde bisher sehr vernachlässigt. So wurde in den ca. 53 Beiträgen, die auf der *Seventh International Conference on User Modeling 1999* (Kay, 1999) vertreten waren, nur in 7 Fällen auf dieses Thema hingewiesen.

Soweit empirische Untersuchungen überhaupt angestellt werden, kommen meist objektive Kriterien zum Einsatz: benötigte Zeit, Geschwindigkeit, Anzahl

Tabelle 1: Einteilung verschiedener Kriterien zur Evaluation adaptiver Systeme

	Berücksichtigung kognitiver Prozesse	
	nein	*ja*
objektive Kriterien	Zeit Geschwindigkeit Anzahl der Dialogschritte	Verhaltenskomplexität
subjektive Kriterien		Usability-Fragebögen

der Dialogschritte (z.B. Litman & Pan, 1999), *accuracy* und *precision* (z.B. Billsus & Pazzani, 1999) oder Lernerfolg (z.B.Specht, 1998). Alle diese Maße bewerten zwar inwieweit ein bestimmter Aspekt der Zielsetzung des jeweiligen Systems erreicht wurde, aber eines der Hauptziele von Adaptivität — nämlich die Interaktion möglichst angenehm und zur Zufriedenheit des Benutzers zu gestalten — wird nur sehr indirekt erfasst. Kognitive Vorgänge beim Benutzer und dessen Erleben werden kaum oder gar nicht berücksichtigt.

Eine Möglichkeit diesem Problem zu begegnen ist die Verwendung von Usability-Fragebögen, also den Benutzer direkt nach seiner Zufriedenheit (z.B. Encarnação & Stoev, 1999) oder der erlebten Aufgabenschwierigkeit (z.B. Bares & Lester, 1997) zu fragen. Dabei treten allerdings zwei Probleme auf: Erstens ist Usability ein sehr umfassendes Konstrukt und enthält Teilaspekte wie etwa Verfügbarkeit oder Fehlerrobustheit (Oppermann, Murchner, Reiterer, & Koch, 1992), auf die Adaptivität normalerweise gar nicht abzielt. Bei einem derart groben Messinstrument verschwinden die gesuchten Effekte leicht in der Gesamtvarianz. Ein zweites Problem bei der Verwendung von Usability Fragebögen zur Evaluation adaptiver Systeme liegt im Verfahren selbst. Im Idealfall bemerken die Benutzer gar nicht, dass sie während der Interaktion individuell unterstützt wurden, und haben dann keinen Vergleichsmaßstab für eine Bewertung.

Die Messung der Verhaltenskomplexität, die im folgenden genauer beschrieben wird, zeigt eine Möglichkeit auf, gleichzeitig das Verhalten der Benutzer objektiv zu beschreiben, und dennoch kognitive Prozesse nicht zu vernachlässigen (siehe Tabelle 1).

2 Komplexität und kognitive Aufgabenanalyse

Die Grundidee kann man folgendermaßen skizzieren: Die meisten adaptiven Systeme zielen darauf ab, die Interaktion zwischen Benutzer und System zu vereinfachen. Adaptivität soll dazu führen, dass der Benutzer schneller und vor allem einfacher das sich gesteckte Ziel erreicht. Dabei verschiebt sich die Verteilung der Aufgaben (Jameson, 1999). Ein adaptives System übernimmt Teilaufgaben, wie z.B. Planungs-, Sortier- oder Auswahlfunktionen oder vereinfacht die Struktur der Aufgabe selbst, indem es Lösungswege oder Funktionen anbietet. In jedem Fall bedeutet dies eine Vereinfachung der kognitiven Struktur der Interaktion.

Dies soll an verschiedenen Beispielen verdeutlicht werden: Ein adaptives Hilfesystem (Encarnação & Stoev, 1999) weiß, welche Ziele der Benutzer im Augenblick verfolgt, und selektiert dementsprechend passende Hilfetexte. Das wiederum erspart dem Benutzer, selbst einen Überblick über die verschiedenen Themen und Begriffe zu gewinnen, und die passenden auszusuchen. Ein adaptives Produktpräsentationssystem (Jörding, 1999) vereinfacht die Informationssuche, indem es genau die Art und Form von Information anbietet, die der Benutzer sucht. Adaptive Lernprogramme (Specht, 1998) unterstützen die Navigation im Kurs durch Link Annotation und Curriculum Sequenzierung. Die Lernenden werden dabei von der Aufgabe befreit, sich zu merken, welche Themenbereiche sie bereits absolviert haben, und in welcher Reihenfolge die weiteren Lehreinheiten zu bearbeiten sind. Sie können sich daher voll auf das eigentliche Ziel — eben das Lernen — konzentrieren.

3 Messung der Komplexität des Verhaltens

Adaptivität reduziert also die Komplexität der Interaktion. Doch wie läßt sich dies messen?

3.1 Theoretische Ansätze der Cognitve Task Analysis

Die Cognitive Task Analysis bietet verschiedene Methoden an, wie Aufgaben und Ziele eines Benutzers repräsentiert werden können. Im folgenden wird nur noch der Begriff Aufgabe (task) verwendet, auch wenn es sich tatsächlich, wie an den oben erwähnten Beispielen deutlich wird, häufig auch um selbst gesetzte Ziele des Benutzer handelt.

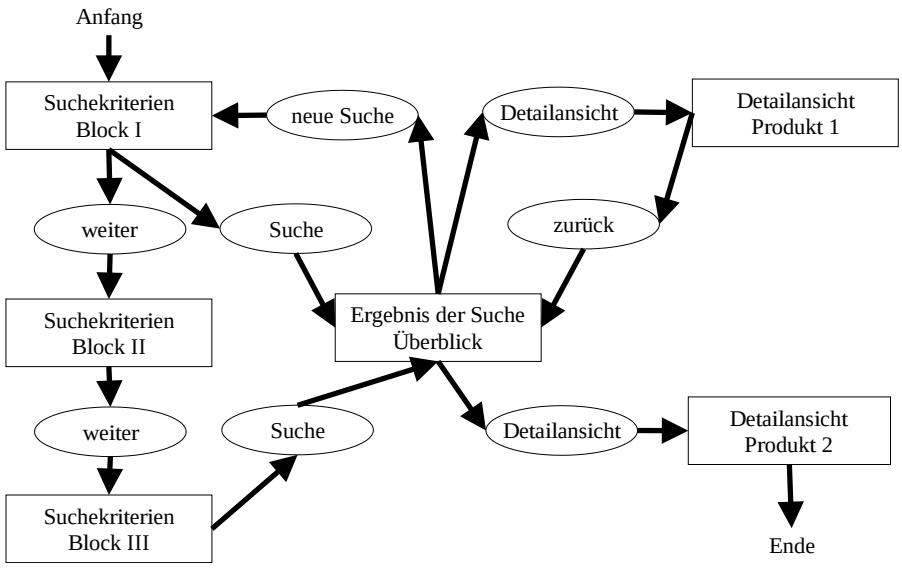

Abbildung 1: Beispiel für ein Petri-Netz, das aus der Interaktion eines Benutzers mit einem Produktberatungssystem gewonnen wurde. Zustände sind als Rechtecke, Übergänge als Ellipsen gekennzeichnet.

In der GOMS Analyse werden die Ziele, Operatoren, Methoden und Entscheidungsregeln des Benutzers zu einer Hierarchie zusammengestellt. Je mehr Unterziele und Operatoren benötigt werden, desto komplexer die Aufgabe (Card, Moran, & Newell, 1983). Andere Ansätze beschreiben das Verhalten des Benutzers mit Hilfe von Grammatiken (z.B. Backus Naur Form bei Reisner, 1984). Sowohl GOMS als auch die Grammatiken haben jedoch eher normativen Charakter (Rasmussen, 1997). Sie analysieren, wie der Benutzer sich idealerweise verhalten sollte. Fehler oder mangelndes Wissen des Benutzers sind nur mit Schwierigkeiten zu repräsentieren.

Rauterberg, Schluep, and Fjeld (1997) umgehen dieses Problem, indem sie ein Verfahren zur Beschreibung des Verhaltens eines einzelnen Benutzers vorschlagen. Die Interaktion läßt sich begreifen als ein Netz von Systemzuständen, zwischen denen der Benutzer über festgelegte Übergänge (z.B. Mausklicks, Tastendrücke, Menüauswahl, etc.) hin und her “springt”. Abbildung 1 zeigt einen Ausschnitt aus einem solchen Netz für ein Produktberatungssystem.

Ein solches Netz, bei dem es sich aus formaler Sicht um ein Petri-Netz handelt, kann aus Protokolldaten extrahiert werden. Dazu werden die vorhandenen

Systemzustände festgelegt. Dann kann ermittelt werden, welchen “Weg” der Benutzer genommen hat.

Ein einfaches System von geringer Komplexität zeichnet sich nun durch eine besonders einfache Struktur dieses Netzes aus. Ein solches Vorgehen hat den Vorteil, dass nicht ein einzelner *one best way* festgelegt werden muss, wie dies etwa bei GOMS der Fall ist, sondern der jeweils individuelle Weg nachgezeichnet werden kann.

Aus kognitiver Sicht kann ein solches Petri-Netz wie folgt interpretiert werden: Der Benutzer benötigt zur Erledigung der Aufgabe eine Repräsentation der verschiedenen Zustände des Systems, der vorhandenen Operatoren, der Systemstruktur und der Entscheidungsstruktur (Rauterberg, Schluep, & Fjeld, 1997). Diese muss, soweit sie noch nicht vorhanden ist, erst noch erworben werden. Je mehr Zustände und Übergänge für die Erledigung der Aufgabe repräsentiert werden müssen, desto (subjektiv) komplexer ist die Aufgabe.

3.2 Berechnung der Komplexität

Rauterberg (1992) vergleicht vier verschiedene Maße der Komplexität, die aus Theorien der Petri-Netze bzw. der Graphen-Theorie abgeleitet wurden: Ein sehr einfaches Maß ist C_{state}. Als Komplexität gilt hier die Anzahl der verschiedenen aufgetretenen Zustände (*S*tates) des Systems.

$$C_{state} = S \tag{1}$$

Es wird sehr schnell klar, dass für die Berechnung der Komplexität auch die Anzahl der Relationen zwischen den Zuständen berücksichtigt werden muss. Ansonsten würde ein System, das sehr viele Funktionen auf einer einzigen Seite bündelt, als besonders einfach gelten. Daher berechnet C_{fan} das Verhältnis zwischen Zuständen und Übergängen (*T*ransitions). Es stellt damit ein Maß für die relative strukturelle Komplexität dar.

$$C_{fan} = \frac{T}{S} \tag{2}$$

Ein drittes Maß berechnet die Anzahl der im Netz enthaltenen Zyklen. Es zeigt also an, wie häufig ein Benutzer wieder zu einem früheren Zustand zurückgekehrt ist.

$$C_{cycle} = T - S + P \tag{3}$$

Die Korrekturkonstante P ist in den hier betrachteten Netzen immer 1.

Das vierte Komplexitätsmaß ($C_{density}$) gibt die Dichte des Netzes in Relation zur maximal möglichen Dichte an.

$$C_{density} = \frac{T}{S \times (S-1)} \tag{4}$$

Beispielhaft soll hier die Berechnung der jeweiligen Komplexitäten für das Netz in Abbildung 1 nachvollzogen werden: Die Anzahl der unterschiedlichen Zustände beträgt $S = 6$. Es traten $T = 8$ verschiedene Übergänge auf.

$$C_{state} = 6 \tag{5}$$

$$C_{fan} = \frac{8}{6} = 1,33 \tag{6}$$

$$C_{cycle} = 8 - 6 + 1 = 3 \tag{7}$$

$$C_{density} = \frac{8}{6 \times (6-1)} = 0,27 \tag{8}$$

Bisherige Erfahrungen bei der Anwendung dieser Maße in der Softwareevaluation (Rauterberg, 1992) ergaben, dass alle vier Varianten zwischen dem Verhalten von Anfänger und Experten differenzieren können. Dabei erwiesen sich C_{state} und $C_{density}$ als sensibel gegenüber der Variation der Aufgabe. Diese beiden Maße sind daher nur für experimentelle Designs geeignet, bei denen die Aufgabe konstant gehalten wird. Sie sollten jedoch nicht verwendet werden, um verschiedene Aufgaben miteinander in Beziehung zu setzen.

Bei der Evaluation von adaptiven Systemen werden unterschiedliche Varianten (z.B. adaptiv versus nicht adaptiv oder unterschiedliche Adaptionsentscheidungen) bezüglich der selben Aufgabe verglichen. Die selbe Aufgabe soll durch geeignete Anpassung weniger komplex werden. Alle vier oben aufgeführten Maße kommen daher für eine Evaluation in Betracht.

Beim Vergleich verschiedener Systeme muss zusätzlich noch ein Korrekturfaktor eingefügt werden, der angibt, wie stark sich die Komplexität der Gesamtsysteme unterscheidet. (Ansonsten würde das ideale System, das immer zu minimaler Verhaltenskomplexität führt, keine Verzweigungen aufweisen, sondern alle Funktionen seriell hintereinander anordnen.) Die Systemkomplexität kann mit den selben Methoden berechnet werden wie die Verhaltenskomplexität, es wird jedoch nicht das beobachtete Verhalten, sondern alle vom Benutzer erreichbaren Zustände und Übergänge der Berechnung zugrunde gelegt. Falls eine solche Berechnung nicht möglich ist, kann auch die Menge aller beobachteter Zustände und Übergänge verwendet werden.

Tabelle 2: Vergleich der vier verschiedenen Komplexitätsmaße nach Gruppen getrennt. Gruppe 1 wurde durch eine adaptive Komponente unterstützt, während Gruppe 2 das Standardsystem benutzte. Angegeben sind Stichprobengröße (N), Mittelwert ($\overline{x}$), Standardabweichung (σ) und statistische Signifikanz der Differenz zwischen den Gruppen (α).

	Gruppe	N	$\overline{x}$	σ	α
C_{state}	1	25	18,36	3,28	,029*
	2	17	20,70	3,31	
C_{fan}	1	25	1,47	0,17	,985
	2	17	1,47	0,21	
C_{cycle}	1	25	9,88	4,05	,003*
	2	17	13,88	4,04	
$C_{density}$	1	25	0,087	0,015	,029*
	2	17	0,100	0,023	

4 Empirische Befunde

Sind die genannten Komplexitätsmaße geeignet um adaptive Systeme zu evaluieren? Um diese Frage zu beantworten, wurde ein adaptives System herangezogen, für das bereits eine Evaluation nach den üblichen Kriterien (Zeit, Zufriedenheit) vorliegt. Das CASTLE System empfiehlt Ferienhäuser auf Grund von Benutzerpräferenzen (Weibelzahl & Weber, 1999).

In einem Laborexperiment hatten Versuchspersonen die Aufgabe, das für sie am besten geeignete Haus aus dem Katalog herauszusuchen. In einer Bedingung (Gruppe 1) wurden die Benutzer dabei durch eine adaptive Komponente unterstützt. Eine zweite Gruppe benutzte das selbe System ohne individuelle Unterstützung. Dabei zeigte sich, dass Gruppe 1 schneller ein passendes Haus fand, und anschließend in einem Fragebogen zufriedener mit der Benutzung war als Gruppe 2. Jedoch verfehlten diese Unterschiede statistische Signifikanz.

In einer Reanalyse der Protokolldaten wurde die Verhaltenskomplexität jedes Benutzers analysiert. Tabelle 2 zeigt die Ergebnisse. Dabei hat der Absolutwert der Komplexität keinerlei Bedeutung, entscheidend sind lediglich die Unterschiede zwischen den Gruppen. Es zeigt sich, dass C_{cycle} und $C_{density}$ tatsächlich zwi-

Tabelle 3: Vergleich der Korrelationen der vier verschiedenen Komplexitätsmaße mit der Zufriedenheit, der Gesamtdauer der Interaktion, sowie der selbst eingeschätzten Erfahrung im Umgang mit Computern und dem Internet ($N = 42$). Signifikante Ergebnisse sind mit * gekennzeichnet ($p <, 05$)

	Zufriedenheit	Gesamt-dauer	Computer-erfahrung	Internet-erfahrung
C_{state}	,15	,50*	,17	,23
C_{fan}	−,04	,53*	−,25	−,22
C_{cycle}	−,04	,57*	−,17	−,15
$C_{density}$	−,33*	−,07	−,29	−,47*

schen den Gruppen differenzieren können.

Zur weiteren Validierung wurde überprüft, inwieweit die Komplexitätsmaße mit dem objektiven Maß *benötigte Gesamtzeit* und dem subjektiven Maß *Zufriedenheit mit der Interaktion* zusammenhängen. Dazu wurden die in diesem Zusammenhang relevanten Items des *Questionnaire of User Interaction Satisfaction*, QUIS, übernommen (Chin, Diehl, & Norman, 1988). Höhere Werte stehen für höhere Zufriedenheit. Weiterhin wurde erwartet, dass Personen mit größerer Erfahrung im Umgang mit Computern und Internet eine geringer Verhaltenskomplexität zeigen. Die Versuchspersonen schätzten sich auf einer fünfstufigen Skala selbst ein. Wie in Tabelle 3 angegeben, zeigt sich deutlich, dass nur $C_{density}$ die subjektive Zufriedenheit der Benutzer widerspiegelt. Die Gesamtdauer hängt dagegen nur mit den übrigen drei Komplexitätsmaßen zusammen. Alle Maße außer C_{state} sprechen wie erwartet mäßig bis stark auf Unterschiede in der Computer- und Internet-Erfahrung an.

5 Diskussion

Während es mit traditionellen Maßen, wie *Gesamtdauer der Interaktion* und *Benutzer-Zufriedenheit* nicht gelungen war, die Unterschiede zwischen den beiden Gruppen statistisch abzusichern, konnten zwei der untersuchten Komplexitätsmaße deutlich zwischen den Gruppen differenzieren. C_{fan} zeigte nicht den erwarteten Effekt und scheint daher für die Evaluation adaptiver Systeme weniger gut ge-

eignet. Dagegen erwiesen sich C_{cycle} und $C_{density}$ als sehr brauchbar. Der Zusammenhang mit externen Kriterien wie Computer-Erfahrung und Internet-Erfahrung zeigt, dass die so berechneten Komplexitäten auf tatsächliche Unterschiede im Verhalten zurückgehen.

Vor allem $C_{density}$ dürfte als Kriterium für weitere Evaluationen interessant sein, da es in engem Zusammenhang mit der subjektiven Zufriedenheit steht, aber die Schwierigkeiten einer expliziten Befragung der Benutzer umgeht.

Das Maß Verhaltenskomplexität hat gegenüber verschiedenen anderen Kriterien den große Vorteil, dass nahezu beliebige Systeme miteinander verglichen werden können. Es ist daher möglich, nicht nur einfach die adaptiven Komponenten eines Systems auszuschalten, sondern stattdessen ein zweites, speziell für die Aufgabe konzipiertes System zu verwenden. Die Schwierigkeiten bei der Wahl einer geeigneten Alternative (Höök, 2000) werden so entscheidend reduziert.

Um die gemessene Verhaltenskomplexität auch kognitiv zu interpretieren, müssen drei Voraussetzungen erfüllt sein: Erstens müssen die verwendeten Systemzustände und Übergänge auch tatsächlich kognitiv repräsentiert sein. Diese Annahme scheint plausibel, sollte aber im Einzelfall noch einmal experimentell überprüft werden. In jedem Fall deutet dies darauf hin, dass nicht alle theoretisch unterscheidbaren Systemzustände in die Analyse der Komplexität eingehen sollten, sondern nur solche, die auch vom Benutzer als unterschiedlich wahrgenommen werden.

Eine weitere Voraussetzung ist, dass alle Benutzer das selbe Ziel verfolgen, bzw. die gestellte Aufgabe in gleicher Weise interpretieren. Auch diese Annahme sollte explizit, z.B. durch die Methode des Lauten Denkens, überprüft werden.

Die dritte Voraussetzung ist, dass alle betrachteten Übergange mit den selben Kosten verbunden sind. So würde sich die Komplexität des Netzes z.B. auch ändern, wenn drei einfach zu lernende Übergange von einem adaptiven System durch einen einzigen hoch komplexen Schritt ersetzt würden. Aus kognitiver Sicht ist diese Art der Anpassung aber sicher nicht von Vorteil. In dem hier betrachteten System ist diese Voraussetzung wahrscheinlich erfüllt, da Übergänge immer nur Sprüngen zwischen mehreren fast statischen Seiten entsprechen. In anderen Fällen muss möglicherweise eine Gewichtung der einzelnen Übergänge mit in die Berechnung der Komplexität eingehen.

Literatur

Bares, W. H., & Lester, J. C. (1997). Cinematographic user models for automated realtime camera control in dynamic 3D environments. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (p. 215-226). Vienna, New York: Springer Wien New York. (Available from http://um.org)

Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (p. 98-108). Vienna, New York: Springer Wien New York.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, New Jersey: Erlbaum.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of ACM CHI'88 Conference on Human Factors in Computing* (pp. 213–218).

Encarnação, L. M., & Stoev, S. L. (1999). Application-independent intelligent user support system exploiting action-sequence based user modeling. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (p. 245-254). Vienna, New York: Springer Wien New York.

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting With Computers*, *12*(4), 409–426.

Jameson, A. (1999). *User-adaptive systems: An integrative overview.* (Tutorial presented at the Seventh International Conference on User Modeling, Banff, Canada, June 20th 1999)

Jörding, T. (1999). Temporary user modeling for adaptive product presentation in the web. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (p. 333-334). Vienna, New York: Springer Wien New York.

Kay, J. (Ed.). (1999). *Proceedings of the seventh international conference on user modeling (UM99).* Wien: Springer.

Litman, D., & Pan, S. (1999). Empirically evaluating an adaptable spoken dialogue system. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (p. 54-64). Vienna, New York: Springer Wien New York.

Oppermann, R., Murchner, B., Reiterer, H., & Koch, M. (1992). *Software-ergonomische Evaluation: der Leitfaden EVADIS II* (2 ed.). Berlin: de Gruyter.

Rasmussen, J. (1997). Merging paradigms: Decision making, management, and cognitive control. In R. Flin, E. Salas, M. Strub, & L. Marting (Eds.), *Decision making under stress: Emerging paradigms and applications* (pp. 67–85). Aldershot, England: Ashgate.

Rauterberg, M. (1992). A method of a quantitative measurement of cognitive complexity. In G. van der Veer, M. Tauber, S. Bagnara, & M. Antalovits (Eds.), *Human-computer interaction: Tasks and organisation* (pp. 295–307). Rom: CUD.

Rauterberg, M., Schluep, S., & Fjeld, M. (1997). How to model behavioural and cognitive complexity in human-computer interaction with Petri nets. In H. Inooka (Ed.), *Proceedings of the 6th IEEE International Workshop on Robot and Human Communication* (pp. 320–325). Piscataway: IEEE.

Reisner, P. (1984). Formal grammar as a tool for analyzing ease of use. In J. Thomas & M. Schneider (Eds.), *Human Factors in Computing Systems* (pp. 53–78). Norwood, New Jersey: Ablex.

Specht, M. (1998). Empirical evaluation of adaptive annotation in hypermedia. In T. Ottmann & I. Tomek (Eds.), *Proceedings of the 10th World Conference on Educational Telecommunications, ED-MEDIA & ED-Telecom '98, Freiburg, Germany* (pp. 1327–1332). Charlottesville, VA.

Weibelzahl, S., & Weber, G. (1999). Benutzermodellierung von Kundenwünschen durch Fallbasiertes Schließen. In T. Jörding (Ed.), *Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen, ABIS-99.* Magdeburg. (availabe at http://www-mmt.inf.tu-dresden.de/joerding/abis99/proceedings.html)