Association for Information Systems
AIS Electronic Library (AISeL)

Wirtschaftsinformatik Proceedings 2009 | Wirtschaftsinformatik

2009

MODELLVERGLEICH MITTELS CLUSTERANALYSE AM BEISPIEL EINER AUTOMATISIERTEN ÄHNLICHKEITSANALYSE FÜR OLAPBERICHTSSPEZIFIKATIONEN

Ralf Knackstedt
European Research Center for Information Systems (ERCIS) der Westfälischen Wilhelms-Universität Münster

Marc Oliver Deinert
European Research Center for Information Systems (ERCIS) der Westfälischen Wilhelms-Universität Münster

Jörg Becker
European Research Center for Information Systems (ERCIS) der Westfälischen Wilhelms-Universität Münster

MODELLVERGLEICH MITTELS CLUSTERANALYSE AM BEISPIEL EINER AUTOMATISIERTEN ÄHNLICHKEITSANALYSE FÜR OLAP-BERICHTSSPEZIFIKATIONEN

Ralf Knackstedt, Marc Oliver Deinert, Jörg Becker[1]

Kurzfassung
Der Modellvergleich stellt eine grundlegende Operation auf konzeptionellen Informationssystemmodellen dar. Im Rahmen der Restrukturierung eines historisch unkontrolliert gewachsenen Berichtswesens ermöglicht er die Identifikation ähnlicher Berichtsspezifikationen. Dies ist Voraussetzung dafür, mehrere ähnliche Berichte durch einen gegebenenfalls parametrisierbaren Bericht zu ersetzen, der die gleichen Informationsbedarfe abdeckt. Aufgrund der Vielzahl von Berichten innerhalb eines Unternehmens muss der Modellvergleich möglichst weitgehend automatisiert erfolgen. Der Beitrag untersucht die Einsatzfähigkeit der Clusteranalyse zum automatisierten Vergleich von OLAP-Berichtsspezifikationen. Es wird gezeigt, dass durch methodische Anpassungen der verwendeten Clusteranalyse Ergebnisverbesserungen erzielt werden können. Außerdem wird die Einbindung des Ansatzes in eine Entwicklungsumgebung für Data-Warehouse-basierte OLAP-Systeme vorgestellt.

1. Modellvergleich im Kontext der Berichtswesenvereinfachung

Der Modellvergleich stellt eine grundlegende Operation auf konzeptionellen Modellen dar. In quantitativer Form ist es sein Ziel, für zwei miteinander verglichene Modelle den Wert einer Kennzahl zurückzugeben, welche die Ähnlichkeit der beiden Modelle misst (ein Wert von 1 bzw. 0 kann dabei z. B. bedeuten, dass die Modelle vollständig gleich bzw. unterschiedlich sind) [9, 13]. Häufig liegt in der Praxis eine Vielzahl an Modellen vor, so dass deren manueller Vergleich sehr aufwändig ist. Unter diesen Voraussetzungen ist es von Bedeutung, Verfahren zu entwickeln, die den Modellvergleich möglichst weitgehend automatisieren, um so seine Wirtschaftlichkeit zu erhöhen.

Dieser Umstand ist im OLAP-basierten Berichtswesen häufig gegeben. In vielen Unternehmen ist der Berichtsbestand historisch gewachsen. Aufgrund fehlender bzw. nicht genutzter Metainformationen weisen viele Unternehmensberichtswesen erhebliche Redundanzen und unnötige Komplexität auf. Ziel einer Berichtswesenvereinfachung ist es deshalb, eine gegebene

[1] European Research Center for Information Systems (ERCIS) der Westfälischen Wilhelms-Universität Münster, D-48149 Münster, Leonardo-Campus 3.

Menge an Berichten im Unternehmen in ihrer Anzahl und/oder Struktur so zu verändern, dass der Aufwand zur Pflege des Berichtswesens gesenkt wird und dabei die Abdeckung eines gegebenen Informationsbedarfs durch das Berichtswesen im Unternehmen sich zumindest nicht verschlechtert. Sehr ähnliche Berichte können dabei häufig zu gemeinsamen Berichten zusammengefügt werden, die gegebenenfalls beim Aufruf mit Selektionsmerkmalsausprägungen geeignet parametrisiert werden. Voraussetzung für die Durchführung der Sanierung eines unkontrolliert gewachsenen Berichtswesens, ist der Vergleich der Berichtsspezifikationen. Ein durchschnittliches betriebliches Berichtswesen weist oftmals über 200 Berichte auf. Eine manuelle Durchführung des Vergleichs dieser Berichte gestaltet sich in diesem Fall aufwändig. Gegebenenfalls können auch die kognitiven Fähigkeiten überfordert werden, die notwendig sind, die Vergleichsobjekte konsistent in Beziehung zueinander zu setzen. Mit einem automatisiert durchgeführten Modellvergleich ließen sich die Wirtschaftlichkeit und Durchführbarkeit der Vereinfachung bestehender Berichtswesen erhöhen bzw. sicherstellen.

Der Modellvergleich lässt sich dabei in ein Vorgehensmodell zur Berichtswesenvereinfachung einordnen (vgl. Abbildung 1). Um die Berichtsspezifikationen für die Analyse zur Verfügung zu haben, sollte das Einlesen der Berichtsspezifikationen über Konverter und standardisierte Austauschprotokolle in ein Modellierungswerkzeug unterstützt werden. Die so eingelesenen Modelle können parallel zum Gegenstand alternativer Verfahren zur Identifikation von Vereinfachungspotenzialen gemacht werden. Neben einfachen kennzahlenbasierten Auswertungen [15] und fortgeschrittenen statistischen Verfahren können zur Generierung von Veränderungsideen auch Berichtswesen-Referenzmodelle verwendet werden. Eine umfassende Softwarewerkzeugunterstützung sollte die Konsolidierung der multimethodisch ermittelten Vereinfachungspotenziale unterstützen. Die resultierenden Berichtsspezifikationen sollten nach Möglichkeit automatisiert in die Berichtssysteme zurückgeschrieben und implementiert werden, wie dies für OLAP-Berichtsspezifikationen bereits seit längerem diskutiert wird [11].

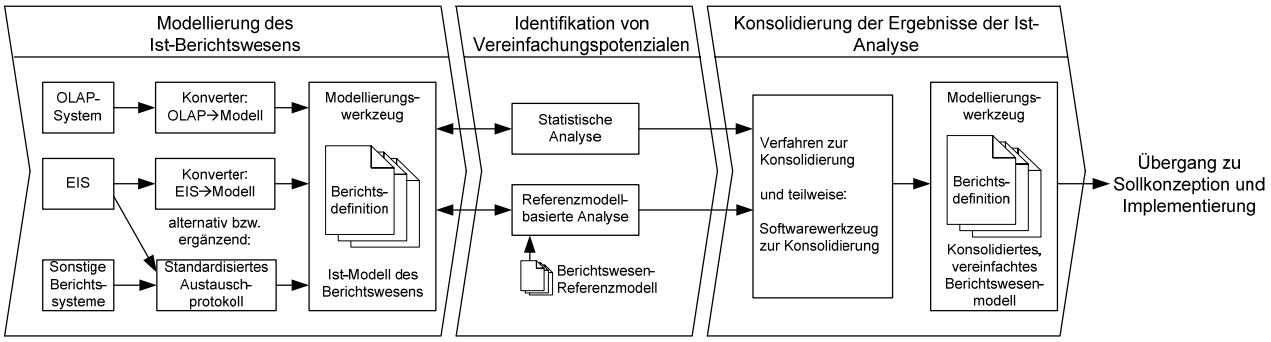

Abbildung 1: Aufgaben und IT-Artefakte der Berichtswesenvereinfachung

Der vorliegende Beitrag verfolgt als einen Lösungsbaustein zur Realisierung einer umfassenden softwarebasierten Unterstützung der Berichtswesenvereinfachung die Forschungsfrage, wie sich der Vergleich von Berichtsspezifikationen mittels statistischer Methoden automatisieren lässt. Als zu untersuchendes konkretes statistisches Verfahren wird die Clusteranalyse gewählt. Zunächst wird die Struktur der OLAP-Berichtsspezifikationen vorgestellt, die dem automatisierten Vergleich unterzogen werden sollen (Abschnitt 2). Um eine Vergleichsbasis für die Ergebnisse der Clusteranalyse zu erhalten, wurde eine mehrteilige Befragung zur Ähnlichkeit von OLAP-Berichtsspezifikationen vorgenommen (Abschnitt 3). Im Anschluss wird in einem zweistufigen Vorgehen die Clusteranalyse für den Modellvergleich adaptiert und ihre softwaretechnische Unterstützung gezeigt. Ein Ausblick schließt den Beitrag ab (Abschnitt 4).

2. OLAP-Berichtsspezifikationen

Für die fachliche Konzeption von OLAP-Systemen wurde eine Vielzahl von Modellierungsansätzen entwickelt. Gemeinsam ist den Ansätzen, dass sie der Eingrenzung von aus Fakten bestehenden Datenmengen dienen, wobei sich Fakten aus Bezugsobjekten und Kennzahlen zusammensetzen. Zur Gliederung der Fülle an Modellierungsansätzen lassen sich fünf Gruppen unterscheiden [6]: *Klassische Ansätze* lehnen sich eng an Modellierungssprachen an, die für die Fachkonzeption auf relationalen Datenbanken basierender, operativer Anwendungssysteme etabliert sind, wie z. B. das Entity-Relationship-Modell [4,8]. *Erweiterungen klassischer Datenmodellierungsansätze* haben die klassischen Ansätze modifiziert, wobei insbesondere der Modellierung von Navigationsräumen, den Kennzahlenbeziehungen und der Modellierung von Bezugsobjektausprägungen Beachtung geschenkt wird [2,20]. *Rein multidimensionale Modellierungsansätze* verzichten auf eine Anlehnung an bestehende Datenmodellierungsansätze und schaffen spezielle Repräsentationsformen [10,11]. Eine gesonderte Gruppe bilden diejenigen Modellierungsansätze, die dem Bereich der *Scientific and Statistical Data Bases* entstammen, der sich bereits seit den achtziger Jahren mit der Spezifikation multidimensionaler Datenräume befasst [7,18]. Darüber hinaus werden *objektorientierte und erweitert objektorientierte Modellierungsansätze* vorgeschlagen [12,19]. Für die Untersuchung des Modellvergleichs mittels Clusteranalyse verwenden wir einen speziellen rein multidimensionalen Modellierungsansatz, der wesentliche Sprachkonstrukte der OLAP-Spezifikationssprachen berücksichtigt (vgl. Abbildung 2). Dabei werden Navigationsräume modelliert, die von Dimension(sausschnitt)en und Kennzahlensystemen aufgespannt werden. Die Dimensionen gliedern Bezugsobjekte (wie z. B. einzelne konkrete Filialen, Artikel oder Tage) in hierarchischer Form. Über einzelne Basisbezugsobjekte wie z. B. die Artikel oder Filialen eines Unternehmens können unterschiedliche orthogonale Dimensionen modelliert werden (z. B. Filialen nach Geographie oder nach Renovierungsstand). Da Berichtsempfänger nicht jeweils das Recht oder die Notwendigkeit besitzen, die gesamten Dimensionen zu analysieren, wird die Modellierung von Dimensionsausschnitten vorgesehen. Die Auswahl an relevanten Dimension(sausschnitt)en wird um Kennzahlen ergänzt. Der Navigationsraum legt auf diese Weise den Informationsbedarf eines OLAP-Berichts fest, der vorsieht, dass alle ausgewiesenen Kennzahlen jeweils hinsichtlich aller angegebenen orthogonalen Dimensionen bzw. Dimensionsausschnitte auswertbar sein sollen.

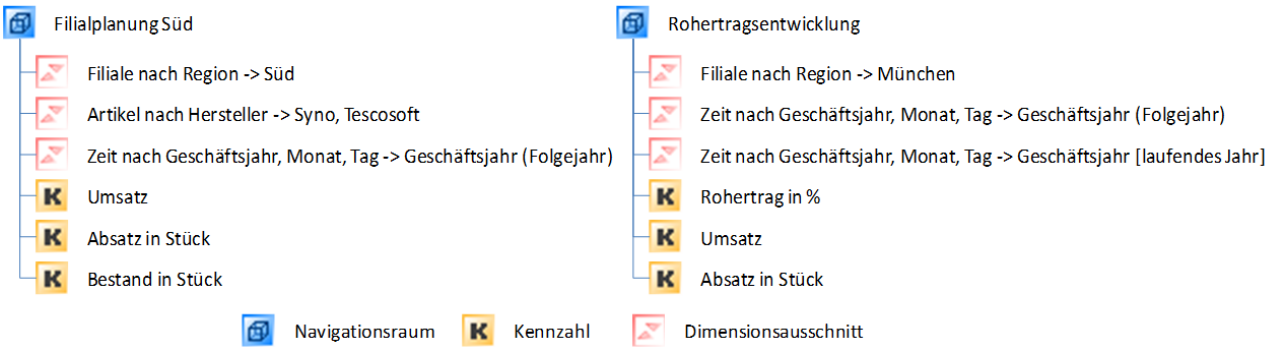

Abbildung 2: Beispiele der modellierten Navigationsräume

3. Befragungsbasierte Erhebung der Ähnlichkeit von Berichtsspezifikationen

Die vorgestellte Modellierungssprache wurde genutzt, um Teilnehmern einer Veranstaltung zur Modellierung und Umsetzung von OLAP-Systemen Paare von Navigationsraumspezifikationen vorzulegen, die diese hinsichtlich ihrer Ähnlichkeit zu bewerten hatten. Durch thematische Ausrichtung der Veranstaltung waren die Teilnehmer mit dem Konzept von Navigationsräumen

vertraut. Bei den Teilnehmern handelte es sich um Studenten der Wirtschaftsinformatik und ihre Betreuer, welche an der Veranstaltung „Führungsinformationssysteme“ teilgenommen haben. Für die Untersuchung wurden 10 Navigationsräume modelliert (vgl. Tabelle 1). Die Navigationsräume „Filialplanung Nord“ und „Filialplanung Süd“ sind dabei sehr ähnlich. Der Navigationsraum „Kundenanalyse Markenware Hamburg“ hingegen ist sehr speziell.

Tabelle 1: Übersicht der verwendeten Navigationsräume

Name	Kennzahlen	Dimensionsausschnitte
Umsätze gesamt -> aktuell	Bestand Stück, Bestand Verkauf, Bestandskalkulation, Umsatz kalkuliert, Umsatz erzielt, Nachlass in %	Filiale nach Region -> uneingeschränkt, Artikel nach Hersteller -> uneingeschränkt, Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Rohertragsentwicklung	Rohertrag in %, Rohertrag, Nettoumsatz erzielt, Umsatz erzielt	Filiale nach Region -> München, Zeit nach Geschäftsjahr, Monat, Tag -> [Folgejahr], Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Kundenanalyse Markenware Hamburg	Kundenanzahl, Bestellmenge Stück, Nachlass in %, Umsatz erzielt	Artikel nach Preisklasse -> Markenware, Filiale nach Region -> Hamburg, Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Filialplanung Süd	Umsatz erzielt, Umsatz kalkuliert, Bestand Verkauf, Bestand Verkauf (Plan)	Filiale nach Region -> Süd, Artikel nach Hersteller -> Syno, Tescosoft, Zeit nach Geschäftsjahr, Monat, Tag -> [Folgejahr]
Filialplanung Nord	Umsatz erzielt, Umsatz kalkuliert, Bestand Verkauf, Bestand Verkauf (Plan)	Filiale nach Region -> Nord, Artikel nach Hersteller -> Syno, Tescosoft, Zeit nach Geschäftsjahr, Monat, Tag -> [Folgejahr]
Filialauswertung Konsolen	Umsatz erzielt	Filiale nach Größe -> Groß, Artikel nach Hersteller -> Syno, Tescosoft, Artikel nach Sortiment -> Konsolen, Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Filialauswertung Hersteller	Umsatz erzielt	Filiale nach Region -> Hamburg, Artikel nach Hersteller -> Minixtor, Trekstar, Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Fahrplan 2005	Umsatz erzielt, Bestand Einkauf, Umsatz kalkuliert	Artikel nach Preisklasse -> uneingeschränkt, Zeit nach Geschäftsjahr, Monat, Tag -> [Folgejahr], Zeit nach Geschäftsjahr, Monat, Tag -> [laufendes Jahr]
Bestand Nord	Bestand Stück, Bestand Einkauf, Bestand Verkauf	Artikel nach Sortiment -> uneingeschränkt, Zeit nach Geschäftsjahr, Monat, Tag -> [gewählter Tag], Filiale nach Region -> Nord
Aktionsauswertung	Bestand Stück, Bestand Verkauf, Absatz Stück, Umsatz erzielt, Nchlass in %	Artikel nach Preisklasse -> Billigware, Filiale nach Region -> uneingeschränkt, Artikel nach Hersteller -> uneingeschränkt

Jedem der 16 Teilnehmer wurden 9 Paare von Navigationsräumen vorgelegt. Die Zusammenstellung der Fragebögen war zufällig. Bei jedem Paar sollte auf einer Skala von 0 bis 10 die Ähnlichkeit angegeben werden, wobei 0 für sehr unähnlich und 10 für sehr ähnlich stand. Bei 10 modellierten Navigationsräumen ergeben sich insgesamt 45 Paare, die es zu vergleichen galt. Jeder Paarvergleich wurde von zwei bis vier Teilnehmern bewertet. Abbildung 3 zeigt eine Matrix, in der die auf einen Wert zwischen null und eins normierten Mittelwerte der paarweisen Ähnlichkeiten aller Navigationsräume angegeben sind, wobei eins maximal ähnlich und null maximal unähnlich bedeutet.

	1	2	3	4	5	6	7	8	9	10
1										
2	0,53									
3	0,71	0,30								
4	0,50	0,45	0,28							
5	0,74	0,38	0,35	0,91						
6	0,63	0,00	0,19	0,50	0,56					
7	0,53	0,52	0,68	0,58	0,71	0,54				
8	0,35	0,65	0,41	0,55	0,41	0,43	0,21			
9	0,25	0,37	0,15	0,45	0,53	0,45	0,30	0,36		
10	0,58	0,38	0,53	0,53	0,48	0,35	0,23	0,41	0,23	

1 - Umsätze Gesamt -> Aktuell
2 - Rohertragsentwicklung
3 - Kundenanalyse Markenware Hamburg
4 - Filialplanung Süd
5 - Filialplanung Nord
6 - Filialauswertung Konsolen
7 - Filialauswertung Hersteller
8 - Fahrplan 2005
9 - Bestand Nord
10 - Aktionsauswertung

Abbildung 3: Ähnlichkeiten anhand der Befragungsergebnisse

Der zweite Teil des Fragebogens bestand aus einer Sammlung von möglichen Begründungen für die Ähnlichkeit zweier Navigationsräume. Jeder Begründung sollte ein Wert zwischen 1 und 10 zugeordnet werden, der die Wichtigkeit dieser Begründung für die Ähnlichkeit angibt. Ziel war es, eine Rangliste an Merkmalen zu erhalten, anhand derer man OLAP-Berichtsspezifikationen vergleichen kann. An die Delphi-Methode angelehnt, wurde eine zweite Befragungsrunde durchgeführt. Dabei wurden jedem Teilnehmer zu jeder Frage seine letzte Antwort sowie der Mittelwert und die Standardabweichung der letzten Runde angegeben. Da von den ursprünglich 16 Teilnehmern nur 12 Teilnehmer an der zweiten Runde der Delphi-Befragung teilgenommen haben, wurden für die Relevanzeinschätzung der Merkmale nur die Angaben dieser 12 Teilnehmer verwendet. Abbildung 4 zeigt die bewerteten Vergleichskriterien nach den Medianergebnissen der zweiten Runde sortiert. Als wichtigste Kriterien können besonders die ersten vier Begründungen angesehen werden, da diese besonders hohe Bewertungen von den Teilnehmern zugesprochen bekamen.

	Median		**Modalwert**		**Standardabweichung**	
	1. Runde	2. Runde	1. Runde	2. Runde	1. Runde	2. Runde
Gleiche Dimensionsausschnitte (DA)	9	9	9	9	1,07	0,79
Gleiche Kennzahlen	8	8	8	8	0,94	0,72
Gleiche Dimension auf die sich DA beziehen	7,5	8	8	8	1,38	1,22
Gleiche Dimensionsgruppe auf die sich DA beziehen	6,5	7	7	7	1,64	1,60
Gleiche Anzahl an Kennzahlen	3,5	4	3	3	2,19	1,31
Sehr ähnliche Bezeichner der Navigationsräume	4	3,5	5	4	2,31	0,78
Gleiche Anzahl an DA	3,5	3	3	4	2,23	1,03
Gleiche Gesamt Anzahl	2	2	1	1	2,62	1,56
Gleiche Reihenfolge der DA	2	2	2	2	0,94	0,85
Gleiche Reihenfolge der Kennzahlen	2,5	2	1	1	1,38	0,90
Gleiche Symbole der Modellelemente	1,5	1	1	1	1,48	0,65

Abbildung 4: Relevanzeinschätzungen der Befragungsteilnehmer

4. Modellvergleich mittels Clusteranalyse

4.1. Einstufiger Ansatz

Ziel der weiteren Untersuchung war es, zu untersuchen, wie mittels Clusteranalyse die Einschätzungen der Teilnehmer aus der Befragung möglichst exakt approximiert werden können. Dafür standen verschiedene algorithmische Ansätze mit verschiedenen Parametrisierungen zur Verfügung. Dabei wurde mit der hierarchischen Clusteranalyse zunächst ein nach [1] in der Praxis weit verbreitetes Verfahren umgesetzt. Dazu müssen für jedes Modell, welches betrachtet werden soll, die Vergleichsmerkmale festgelegt werden. Mit der hierarchischen Clusteranalyse konnten die ersten beiden Kriterien aus der Befragung umgesetzt werden („Gleiche Dimensionsausschnitte“ und „Gleiche Kennzahlen“). Da die hierarchische Clusteranalyse im standardmäßigen Ansatz nicht mit strukturierten Merkmalen arbeitet, wurden die Begründungen „Gleiche Dimension auf die sich Dimensionsausschnitt bezieht“ und „Gleiche Dimensionsgruppe auf die sich Dimensionsausschnitt bezieht“ in dieser Entwicklungsstufe nicht umgesetzt. Als Parameter der hierarchischen

Clusteranalyse bei binären Merkmalen bieten sich verschiedenste Methoden an, welche in der entsprechenden Literatur ausgiebig diskutiert werden [1, 5, 16]. Als Koeffizienten zur Ähnlichkeitsberechnung lassen sich zwei Typen identifizieren, auf der einen Seite der M-Koeffizient für symmetrische binäre Merkmale und auf der anderen Seite der S-Koeffizient für nicht-symmetrische binäre Merkmale [5]. Bei OLAP-Modellen handelt es sich dabei um binäre Merkmale der Form „vorhanden/nicht vorhanden", welche nicht-symmetrische Merkmale darstellen [5]. Verwendet wird hierbei der Tanimoto-Koeffizient, der sich in vielen Disziplinen bewährt hat [17]. Der Tanimoto-Koeffizient berechnet sich aus der Anzahl der übereinstimmenden Merkmale, dividiert durch die Anzahl aller Merkmale, die in beiden Objekten vorhanden sind.

	1	2	3	4	5	6	7	8	9	10
1										
2	0,25									
3	0,46	0,12								
4	0,23	0,12	0,20							
5	0,47	0,05	0,27	0,20						
6	0,45	-0,10	0,10	0,28	0,34					
7	0,58	0,27	0,30	0,47	0,60	0,41				
8	0,05	0,28	0,21	0,18	0,04	0,18	-0,07			
9	0,16	0,37	0,15	0,34	0,28	0,33	0,30	0,24		
10	-0,08	0,19	0,37	0,23	0,18	0,26	0,13	0,21	0,13	

Einfache Fehlersumme	10,6845	**Durchschnitt**	0,2374
Summe der absoluten Fehler	11,1974	**Durchschnitt**	0,2488
Summe der quadrierten Fehler	3,6224	**Durchschnitt**	0,0804

Abbildung 5: Differenz der Ähnlichkeiten aus Befragung und Clusteranalyse

Ein Vergleich der durchgeführten Clusteranalyse mit den Ähnlichkeitsmaßen der Befragung zeigt einige charakteristische Ergebnisse (vgl. Abbildung 5). Auffällig ist zunächst, dass nahezu alle Werte bei der Befragung größer waren als bei der Clusteranalyse. Trotzdem kann man ähnliche Tendenzen erkennen. Die beiden Navigationsräume mit der höchsten Ähnlichkeit sind in beiden Fällen die Navigationsräume „Filialplanung Süd" und „Filialplanung Nord". Auch für die geringen Ähnlichkeitswerte geht die Aussage der Vergleichsgruppe in eine ähnliche Richtung. Insbesondere in den Extremen zeigt sich also eine allgemeine Übereinstimmung, d. h. dass Paare mit einer hohen Ähnlichkeit in der Befragung auch in der automatischen Clusteranalyse eine hohe Ähnlichkeit aufweisen. Es sind jedoch auch teils extreme Abweichungen zu erkennen. So wurde die Ähnlichkeit der Objekte 1 und 7 von den Studenten um 0,58 größer eingestuft, als von der automatischen Clusteranalyse. Diese großen Abweichungen finden sich nicht nur bei einigen weiteren Paaren, wie z. B. den Objekten 5 und 7, sondern auch in den Fehlerwerten. Diese sind ebenfalls in der Abbildung dargestellt und zeigen die Summe aller Abweichungen an. Die hohe absolute Fehlersumme von 11,1974 deutet ebenfalls auf große Diskrepanzen hin. Die hohe Abweichung könnte daraus resultieren, dass nur zwei der vier Begründungen für die Ähnlichkeit der Modelle in dieser Phase der Clusteranalyse betrachtet werden.

4.2. Mehrstufiger Ansatz

Um zu überprüfen, ob sich die Ergebnisse der Clusteranalyse hinsichtlich der Abweichung von den Befragungsergebnissen noch verbessern lassen, wurden in einem weiteren Iterationsschritt auch die beiden weiteren als sehr relevant identifizierten Begründungen der Ähnlichkeit von Berichtsspezifikationen berücksichtigt. Diese betreffen die Struktur der Dimensionsausschnitte. Sind zwei Dimensionsausschnitte Ausschnitte aus derselben Dimension, so sollte dieser Faktor in der berechneten Ähnlichkeit berücksichtigt werden. Die klassischen Ähnlichkeitsmaße der

Clusteranalyse können diese nur implizit in den Dimensionsausschnitten vorhandene Struktur aber nicht abbilden. Also ist eine Modifikation der klassischen Maße nötig. Während bei der klassischen Berechnung des Tanimoto-Koeffizienten die Anzahl der Übereinstimmungen verwendet wird, so muss in der modifizierten Form die Definition der Übereinstimmung angepasst werden. Vorschläge für solche Anpassungen basierend auf Merkmalsgewichtungen finden sich bei [5], wobei die hier verwendete Anpassung im Folgenden beschrieben wird. Wird für einen Dimensionsausschnitt eines Objekts, wie z. B. „Filiale nach Region -> Süd“ keine Übereinstimmung im Objekt gefunden, mit dem verglichen wird, muss die übergeordnete Dimension überprüft werden. Es wird dann für die Dimension „Filiale nach Region“ geprüft, ob das zu vergleichende Objekt einen Dimensionsausschnitt enthält, der ebenfalls aus dieser Dimension abgeleitet wurde. Wird eine solche Übereinstimmung gefunden, so erhöht sich die Anzahl der Merkmale um eins und die Anzahl der Übereinstimmungen erhöht sich um einen Faktor α zwischen 0 und 1. Dieser Faktor, mit dem die Übereinstimmung bei übergeordneten Dimensionen gewichtet wird, geht dann als neuer Parameter in die Definition des Ähnlichkeitsmaßes ein.

	1	2	3	4	5	6	7	8	9	10
1										
2	0,04									
3	0,29	-0,14								
4	-0,15	-0,19	-0,09							
5	0,08	-0,26	-0,03	0,00						
6	0,20	-0,38	-0,09	-0,03	0,03					
7	0,30	-0,08	0,21	0,01	0,14	0,15				
8	-0,11	0,19	-0,02	-0,02	-0,15	-0,02	-0,30			
9	-0,16	-0,05	-0,17	-0,05	0,01	-0,11	-0,14	-0,20		
10	-0,16	-0,06	0,07	-0,13	-0,18	0,00	-0,22	-0,07	-0,15	

Einfache Fehlersumme	-2,1976	**Durchschnitt**	-0,0488
Summe der absoluten Fehler	5,6417	**Durchschnitt**	0,1254
Summe der quadrierten Fehler	1,0744	**Durchschnitt**	0,0239

Abbildung 6: Differenz der Ähnlichkeiten aus Befragung und modifizierter Clusteranalyse

In diesem zweiten Schritt wurde also erneut eine Clusteranalyse durchgeführt, allerdings unter Verwendung des modifizierten Ähnlichkeitsmaßes (vgl. Abbildung 6). Als Faktor für die Gewichtung der Übereinstimmung zwischen den Dimensionen zweier Dimensionsausschnitte wurde exemplarisch der Wert 0,7 eingestellt. Betrachtet man hier die Ähnlichkeiten stellt man fest, dass im Gegensatz zum einstufigen Ansatz die Ähnlichkeiten im Durchschnitt größer sind als die Befragungswerte. Zusätzlich sind die Abweichungen wesentlich geringer geworden. War die größte Abweichung im einstufigen Ansatz noch 0,6, so liegt diese jetzt nur noch bei 0,38. Auch die absolute Fehlersumme ist hier mit 5,6417 gesunken.

4.3. Vergleich der Ansätze

Für den Vergleich der beiden Clusteranalysevarianten lassen sich mit der einfachen Fehlersumme, der Summe der absoluten Fehler und der Summe der quadrierten Fehler verschiedene Fehlermaße anwenden. Die einfache Fehlersumme hat die Eigenschaft, dass sich die Abweichungen nach oben und unten zu null summieren können. Die beiden verbleibenden Fehlermaße unterscheiden sich dahingehend, wie stark sie einzelne große Abweichungen gewichten. Für den einstufigen Ansatz zur Clusteranalyse zeigt Abbildung 5 jeweils die absoluten Fehler und eine Summe der quadrierten Fehler von 3,6224. In Abbildung 6 ist für den mehrstufigen Ansatz zu erkennen, dass nicht nur die Einzelfehler deutlich geringer ausfallen, als bei der einstufigen Betrachtung, sondern dass auch die Summe der quadrierten Fehler nur noch 1,0744 beträgt. Betrachtet man die beiden Fehlersummen

zeigt sich durch den Einsatz des neuen Algorithmus eine deutliche Verbesserung von 3,6224 auf 1,0744. Auch für die anderen Arten der Abweichungsmessung ergeben sich diese Tendenzen. Die Summe der absoluten Fehler sinkt von 11,1974 bei der einstufigen Clusteranalyse auf 5,6417 bei der mehrstufigen Clusteranalyse. Zusätzlich sinkt auch der durchschnittliche Fehler von 0,2488 auf 0,1254, was auf eine deutlich bessere Approximation der Befragungswerte hindeutet. Diese Abweichung kann durch die Wahl der korrekten prozentualen Ähnlichkeitsbewertung verschiedener Dimensionsausschnitte noch optimiert werden, wofür Mehrfachberechnungen für alternative Werte des Parameters α vorzunehmen sind. Des Weiteren legt das Ergebnis nahe, dass zusätzliche Erweiterungen des Clusteralgorithmus, die z. B. auch die unterschiedliche Zusammensetzung der Bezugsobjektmengen einzelner Dimensionsausschnitte berücksichtigen, weitere Verbesserungen hervorbringen können.

4.4. Softwarewerkzeug

Für die prototypische Realisierung einer Softwareunterstützung des Modellvergleichs wurde ein Metamodellierungswerkzeug um die Funktionalität der Clusteranalyse erweitert. Mit dem Metamodellierungswerkzeug lassen sich beliebige hierarchische Modellierungssprachen, wie z. B. die hier verwendete, anlegen. Nach der Sprachdefinition können der Definition entsprechende Modelle konstruiert werden. Die Clusteranalyse operiert direkt auf den von dem Modellierungswerkzeug verwalteten Modellen. Neben den üblichen Einstellung der Clusteranalyse, wie z. B. Clusterverfahren und zu verwendendes Distanz- bzw. Ähnlichkeitsmaß, ist bei der Parametrisierung der in das Modellierungswerkzeug integrierten Clusteranalyse die Angabe der zu verwendenden Merkmale in Form von Modellelementtypen (z. B. Dimensionsausschnitte und Kennzahlen) vorgesehen (vgl. Abbildung 7). Die hier präsentierten Ergebnisse der Clusteranalyse wurden auf diesem Wege ermittelt. Das Ergebnis der Clusteranalyse wird als Dendrogramm dargestellt. Nach Angabe einer gewünschten Anzahl an Clustern, werden die Modelle in einer dem Dendrogramm entsprechenden Gliederung sortiert und können aus dieser Anordnung heraus weiterbearbeitet werden.

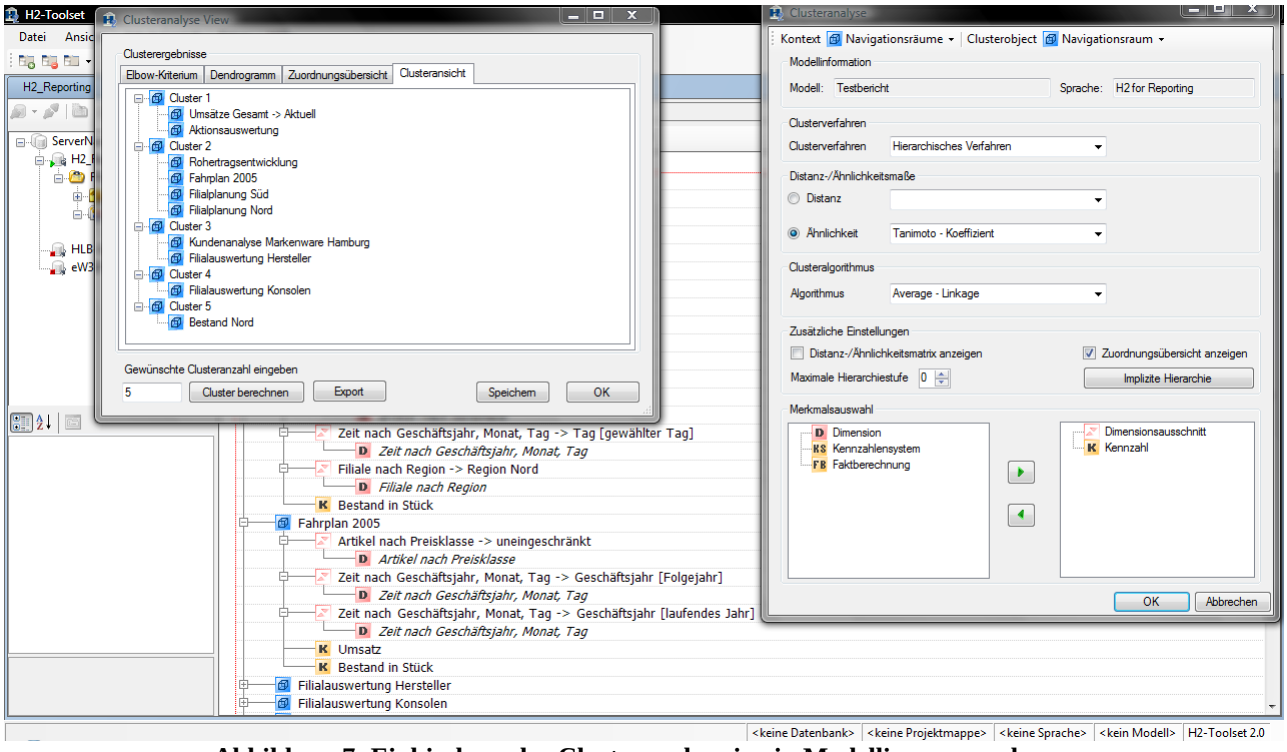

Abbildung 7: Einbindung der Clusteranalyse in ein Modellierungswerkzeug

5. Ausblick

Die vorgestellte Untersuchung zeigt, wie sich die Clusteranalyse zum Vergleich von OLAP-Berichtsspezifikationen einsetzen lässt. Auf der Basis der Ähnlichkeitswerte einzelner Berichtsspezifikationspaare ermittelt die Clusteranalyse homogene Berichtsgruppen. Diese Gruppenbildungen können genutzt werden, um ähnliche Berichte zu erkennen und diese dann zu konsolidieren, um eine allgemeine Datenreduktion und Vereinfachung des Berichtswesens zu erreichen (vgl. Abschnitt 1). Durch die Integration in ein Metamodellierungswerkzeug lässt sich das vorgestellte Verfahren des Modellvergleichs auf weitere hierarchische Modellierungssprachen anwenden. Die in diesem Beitrag verwendete Modellierungssprache wurde bewusst einfach gehalten, um die durchgeführte Befragung zu erleichtern. Neben komplexeren OLAP-Berichtsspezifikationssprachen [6] lässt sich der Ansatz z. B. auf eine hierarchische Modellierungssprache zur Konstruktion hybrider Leistungsbündel anwenden [3]. Ein Element dieser Sprache stellen dabei kundenindividuelle Leistungsbündel dar, die mit Hilfe eines Produktkonfigurators erstellt werden. Die Clusteranalyse bietet nun die Möglichkeit alle diese kundenindividuellen Modelle auf Ähnlichkeiten hin zu untersuchen, um neue Ideen für vorkonfigurierte Leistungsbündel zu generieren. Als weiteres Beispiel zur Erweiterung dieser Arbeit eignet sich auch die Modellierung von Webseiten [14]. Dabei werden die Hauptmerkmale von Webseiten identifiziert und in strukturierter Form in einem hierarchischen Modell dargestellt. Diese Webseiten können dann mit Hilfe der Clusteranalyse analysiert und in homogene Gruppen eingeteilt werden. Auf diese Weise unterstützt die Softwarelösung auch die Entwicklung von Referenzmodellen, was insbesondere für die Konstruktion von Berichtswesen-Referenzmodellen ausgenutzt werden kann. Kritisch anzumerken ist, dass als Vergleichsgrundlage für die Bewertung verschiedener Clusteranalyseansätze bisher ausschließlich Befragungen herangezogen wurden, an denen Probanden des akademischen Bereichs teilgenommen haben. Die Übertragbarkeit dieser Ergebnisse auf die Unternehmenspraxis bleibt bis zur Wiederholung der Untersuchungen mit Teilnehmern aus der Unternehmenspraxis offen.

6. Danksagung

Dieser Beitrag wurde durch die Förderung des BMBF Projektes „FlexNet“ (Flexible Informationssystemarchitekturen für hybride Wertschöpfungsnetzwerke; Förderkennzeichen 01FD0629) im Rahmen des Förderprogramms „Innovationen mit Dienstleistungen“ ermöglicht. Wir danken an dieser Stelle dem Projektträger Deutsches Zentrum für Luft- und Raumfahrt (DLR) für die Unterstützung.

7. Literaturhinweise

[1] BACKHAUS, K.; ERICHSON, B.; PLINKE, W.; WEIBER, R.: Multivariate Analysemethoden. 11. Auflage ed., Berlin 2006.

[2] BAEKGAARD, L.: Event-Entity-Relationship Modelling in Data Warehouse Environments. In: Proceedings of the ACM Second International Workshop on data Warehousing and OLAP (DOLAP ’99), S. 9 – 14, Kansas City 1999.

[3] BECKER, J.; BEVERUNGEN, D.; KNACKSTEDT, R.; MÜLLER, O.: Konzeption einer Modellierungssprache zur tool-unterstützten Modellierung, Konfiguration und Bewertung hybrider Leistungsbündel. In: Proceedings of the GI-Tagung Modellierung, Workshop Dienstleistungsmodellierung, S. 45-62, Berlin 2008.

[4] BECKER, J.; HOLTEN R.: Fachkonzeptuelle Spezifikation von Führungsinformationssystemen. In: Wirtschaftsinformatik, Band 40 Nr. 6, S. 483 – 493, 1998.

[5] BOCK, H.-H.: Automatische Klassifikation, Vandenhoeck & Ruprecht, Göttingen 1974.

[6] BÖHNLEIN, M.: Konstruktion semantischer Data-Warehouse-Schemata, 1. Auflage, Deutscher Universitäts-Verlag, Wiesbaden, 2001.

[7] CHAN, P.; SHOSHANI, A.: SUBJECT – A Directory Driven System for Organizing and Accessing Large Statistical Databases. In: Proceedings of the 7th International Conference on Very Large Data Bases (VLDB ’81), S. 553 – 563, Cannes 1981.

[8] CHEN, P. P.-S.: The Entity-Relationship Model – Toward a Unified View if Data. In: ACM Transaction on Database Systems, Band 1, S. 9 – 36, 1976.

[9] EHRING, M., KOSCHMIDER, A., OBERWEIS, A., Measuring similarity between semantic business process models, in: J. F. Roddick, A. Hinze (Hrsg.): Proceedings of the 4th Asia-Pacific Conference on Conceptual Modeling (APCCM 2007), Ballarat, Australia, 2007, S. 71-80.

[10] GABRIEL, R.; GLUCHOWSKI, P.: Semantische Modellierungstechniken für multidimensionale Datenstrukturen. In: HMD, Band 34, S. 18 – 37, 1997.

[11] HOLTEN, R.: Specification of management views in information warehouse projects. Information Systems, 28 (2003) 7, S. 709-751.

[12] HOLTHUIS, J.: Der Aufbau von data Warehouse-Systemen. Konzeption – Datenmodellierung – Vorgehen. Wiesbaden 1998.

[13] JUNG, J.-Y., BAE, J., Workflow clustering method based on process similarity, in: M. L. Gavrilova, O. Gervasi, V. Kumar, C. J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Hrsg.): Proceedings of the 6th International Conference on Computational Science and Its Applications (ICCSA 2006), Glasgow, Großbritanien, 2006, S. 379-389.

[14] RETSCHITZEGGER, W.; SCHWINGER, W.: Towards Modelling of DataWeb Applications – A Requirements‘ Prespective. In: Proceedings of the Americas Conference on Information Systems (AMCIS 2000), Band 1, S. 149 – 155, Long Beach 2000.

[15] SEIDEL, S.; JANIESCH, C.; WINKELMANN, A.: Softwareeinführung als Anlass zur Berichtswesenverbesserung. In: Becker, J.; Vering, O.; Winkelmann, A. (Hrsg.): Softwareauswahl und -einführung in Industrie und Handel. Vorgehen bei und Erfahrungen mit ERP- und Warenwirtschaftssystemen. Berlin et al. 2007, S. 219-242.

[16] SPÄTH, H.: Cluster Analyse Algorithmen zur Objektklassifizierung und Datenreduktion. In: Verfahren zur Datenverarbeitung, Editor: Späth, H.. Oldenbourg, München 1975.

[17] SPÄTH, H.: Partitionierende Cluster-Analyse bei Binärdaten am Beispiel von bundesdeutschen Hochschulen und Diplomstudiengängen. In: Zeitschrift für Operations Research, Band 21, S. B85 – B96, Physica-Verlag, Würzburg 1977.

[18] SU, S. Y. W.: SAM* - A Semantic Association Model for Corporate and Scientific Statistical Databases. In: Journal of Information Sciences, Band 29, S. 582 – 603, 1978.

[19] TOTOK, A.: Modellierung von OLAP- und Data-Warehouse-Systemen. Wiesbaden 2000.

[20] TRYFONA, N.; BUSBORG, F.; CHRISTIANSON, J. G. B.: starER – A Conceptual Model for Data Warehouse Design. In: Proceedings of the ACM Second International Workshop on data Warehousing and OLAP (DOLAP ’99), S. 9 – 14, Kansas City 1999.