Ad-hoc Datentransformationen für Analytische Informationssysteme

Christian Lüpkes
OFFIS - Institut für Informatik
Escherweg 2
26121 Oldenburg, Deutschland
christian.luepkes@offis.de

ABSTRACT

Beim Betrieb von Data Warehouse Systemen kann es zu einem Semantic Shift kommen. Dieser bezeichnet eine Veränderung der Bedeutung von Dimensionselementen und kann bei Nichtbeachtung zu Informationsverlust und fachlich inkorrekten Analyseergebnissen führen. In dieser Arbeit wird ein graph-basierter Ansatz vorgeschlagen, welcher die Änderungen zwischen Dimensionen als Überleitungen verwalten und für Analysen zur Verfügung stellen kann. Dadurch wird es möglich, Anfragen in Analytischen Informationssystemen unter Berücksichtigung eventueller Semantic Shifts zu beantworten. Dieser Ansatz verzichtet dabei auf eine kennzahlbasierte Approximation und nutzt die Überleitungen klassischer Adaptionsverfahren. Der eingeführte Ansatz wird kritisch hinsichtlich bestehender Ansätze diskutiert und exemplarisch in verschiedenen Domänen durchgeführt.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Administration — *Data warehouse and repository*; H.2.8 [**Database Management**]: Applications; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design, Data Analysis

Keywords

Data warehouse, Schema Versioning, OLAP, Temporal Data Warehouse

1. EINLEITUNG

Die am meisten verwendete Architektur für Analytische Informationssysteme ist die des Data Warehouses mit Metadaten, welche die gespeicherten Daten beschreiben und einer auf diesen Metadaten aufbauender Auswertungssoftware. Die Metadaten werden dabei in streng hierarchisch organisierten Taxonomien, sogenannten Dimensionen, gespeichert. Dimensionen beschreiben, wie die Daten analysiert werden können.

Das Data Warehouse ist dabei nach der Definition von Inmon eine themenorientierte, integrierte, stabile Sammlung zeitbezogener Daten, welche als Datenbasis zur Analyse dient [9]. Data Warehouses haben also immer einen Zeitbezug, bieten aber keine hochentwickelten Konzepte, um mit Änderungen in den Metadaten über die Zeit umzugehen. Klassisch wird davon ausgegangen, dass die Metadaten über die Zeit weitgehend stabil sind [7] [11]. Falls die Metadaten in Einzelfällen doch angepasst werden müssen, werden die gespeicherten Daten einfach den neuen Metadaten entsprechend umcodiert, die sogenannte *Instanzadaption* [2] [11]. Der Nachteil dieses Ansatzes ist, dass beim Umcodieren üblicherweise ein Informationsverlust entsteht. Zudem wird durch die Änderung der Metadaten und die Instanzadaption eine Wiederholung früherer Anfragen unmöglich. Außerdem besteht bei klassischen Systemen keine Möglichkeit die spezifischen Informationen der Metadatenänderungen zu speichern, da die Metadaten selbst nicht zeitbezogen gespeichert werden [12].

2. PROBLEMBESCHREIBUNG

Um das identifizierte Problem des *Semantic Shift* bei Datenanalysen zu verdeutlichen, soll an dieser Stelle zunächst ein Beispiel aus der Arbeit des Autors im deutschen Gesundheitswesen gegeben werden. Dort werden alle Diagnosen nach der *ICD*-Klassifikation, der ***I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems*, codiert. Die Klassifikation selbst beinhaltet sowohl beschreibende als auch ordnende Metadaten und wird als Dimension zur Datenanalyse verwendet. Die deutsche Modifikation der *WHO*-ICD, ICD-GM (***G**erman **M**odifikation*), wird dabei jedes Jahr durch eine Expertengruppe des DIMDI, ***D**eutsches **I**nstitut für **M**edizinische **D**okumentation und **I**nformation* aktualisiert [3] [4] [5].

Die Aktualisierungen bestehen darin, dass neu identifizierte Erkrankungen einen Code zugewiesen bekommen, Erkrankungen zusammengefasst werden oder einzelne Krankheitsbereiche neu unterteilt werden. So wurde zum Beispiel im Jahr 2006 der Code *J09* für die neu identifizierte Vogelgrippe eingeführt.

Um die Daten zwischen den Jahren transformieren zu können, stellt das DIMDI zusätzlich sogenannte Überleitungen in einem Datenbankformat zur Verfügung. In den Abbildungen 1 und 2 sind diese exemplarisch in Ausschnitten für die Jahre 2005 bis 2007 gezeigt. Der Buchstabe A zeigt dabei an, ob eine Überführung automatisch in der durch die Spal-

23^{rd} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 31.05.2011 - 03.06.2011, Obergurgl, Austria.

icd_code2005	icd_code2006	auto2005_2006	auto2006_2005
J10.0	J10.0	A	A
J10.0	J09		A
J10.1	J10.1	A	A
J10.8	J10.8	A	A

Abbildung 1: Ausschnitt der offiziellen ICD Überleitungen zwischen den Jahren 2005 und 2006

tenüberschrift festgelegten Richtung möglich ist.

icd_code2006	icd_code2007	auto2006_2007	auto2007_2006
J09	J09	A	A
J10.0	J10.0	A	A
J10.1	J10.1	A	A
J10.8	J10.8	A	A

Abbildung 2: Ausschnitt der offiziellen ICD Überleitungen zwischen den Jahren 2006 und 2007

Im Jahr 2005 auf 2006 ist es zum Beispiel möglich, den Code *J10.0* zwischen den Jahren umzucodieren. Allerdings gibt es noch einen zweiten Eintrag von *J10.0* auf *J09* bei dem nur eine Umcodierung von 2006 auf 2005 zugelassen ist.

In Analytischen Informationssystemen werden die Daten meist nach der aktuellsten ICD-Definition gespeichert. Daten aus dem Jahr 2005 würden also im Jahr 2006 und 2007 umcodiert. Entsprechend den Überleitungsregeln aus den Abbildungen 1 und 2 würde der Wert *J10.0* syntaktisch gleich bleiben und nicht umcodiert werden.

Eine typische Anfrage wäre nun der Art *„Zeig mir die jährliche Summe aller behandelten J10.0 Patienten der Jahre 2005, 2006 und 2007“* welche das in Abbildung 3 gezeigte Ergebnis liefert. Im Jahr 2006 gab es also im Vergleich zum Vorjahr eine Abnahme von *J10.0* Patienten die sich im Jahr 2007 noch deutlicher fortsetzt.

Summe der behandelten J10.0 Patienten nach Jahr

2005	2006	2007
18347	17913	17548

Abbildung 3: Summe aller behandelten J10.0 Patienten der Jahre 2005, 2006 und 2007

2.1 Darstellung als Graph

In Abbildung 4 ist exemplarisch ein Ausschnitt der ICD-Klassifikation für nachgewiesene sonstige Influenzaviren der Jahre 2005 bis 2007 abgebildet. Der Graph repräsentiert dabei die offizielle Taxonomie der ICD Codes und die gerichteten Kanten repräsentieren die offiziell als gültig definierten Transformationsregeln für Umcodierugnen der drei abgebildeten Jahre 2005, 2006 und 2007, wie sie in den Tabellen der Abbildungen 1 und 2 definiert sind.

In einem Data Warehouse werden die zu analysierenden Daten in der Regel auf der feinsten verfügbaren Klassifikationsstufe vorgehalten. Veranschaulicht handelt es sich also um die Ausprägungen der Blätter. Falls eine Analyse der Erkrankungen *J10.1* oder *J10.8* durchgeführt werden soll,

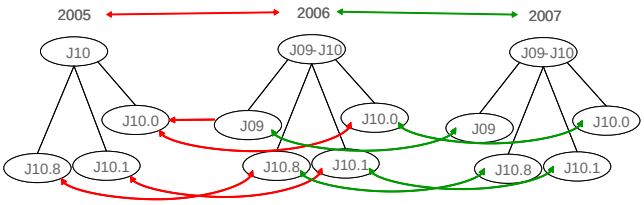

Abbildung 4: Darstellung dreier Teilgraphen der ICD-GM Metadaten für Influenzaviren und deren Überleitung über die Jahre 2005 bis 2007

ist dies unproblematisch, da es keinerlei Änderungen in der Datenbeschreibung gab; zu erkennen an der Existenz der bijektiven Kanten zu den gleichen Knoten jedes Jahres. Das Problem des *Semantic Shift* tritt auf, wenn *J10.0* für die Jahre 2005 bis 2007 untersucht werden soll. Es ist nicht eindeutig, welche Bedeutung von *J10.0* verwendet werden soll. Falls die Bedeutung des Codes *J10.0* vom Jahr 2005 beabsichtigt ist, sollte für die Folgejahre auch der Code *J09* berücksichtigt werden. Falls die Semantik von 2006 oder 2007 gemeint ist, muss dem menschlichen Analysten bewusst sein, dass es eine inhaltliche Änderung vom Jahr 2005 zu 2006 für *J10.0* gab auch wenn die Daten syntaktisch identisch sind und Transitionen in beide Richtungen existieren.

Das analysespezifische Hintergrundwissen des Fachexperten ist, dass *J10.0* ein Sammelknoten für nicht genauer bestimmte Influenzaviren ist. Wie bereits oben erwähnt, wurde in 2006 die Vogelgrippe identifiziert und als neuer Code *J09* eingefügt. Dadurch wurde die Bedeutung von *J10.0* als alle unbestimmten Influenzaviren zwar nicht verändert, aber verglichen mit 2005 *fehlen* nun die Vogelgrippefälle. Für statistische Analysen auf solch einer feingranularen Ebene würden daher fehlerhafte Ergebnisse geliefert. Bei einer Analyse auf den Elternknoten *J10* des Jahres 2005 oder *J09-J10* der Jahre 2006 und 2007 wären die Resultate korrekt, da alle Transformationskanten auf Kindknoten verweisen.

Für die Ergebnisse in Abbildung 3 bedeutet dies, dass die Abnahme der *J10.0* Erkrankungen auch darin begründet liegt, dass Krankheitsfälle in *J09* codiert wurden, die vorher in *J10.0* enthalten waren.

2.2 Weitere Domänen

Der *Semantic Shift* kann nicht nur in der medizinischen Dokumentation beobachtet werden, sondern auch in anderen Bereichen. So kann man z.B. für die Entwicklung der Länder Europas von 1988 bis 2000 bedingt durch den Zusammenbruch des Warschauer Pakts ähnliches feststellen. Allerdings muss dort beachtet werden, dass es sich bei den abgeleiteten und angepassten Metadaten nicht um gesetzlich vorgegebene Dimensionsstrukturen handelt, sondern um von Fachexperten erstellte Dimensionen. Dies ist der Normalfall bei Data Warehouses. Die Dimension soll alle Länder im Herzen Europas widerspiegeln. Bis zum Jahr 1991 gab es die beiden eigenständigen deutschen Staaten *BRD* und *DDR*. In der Dimension wären diese dann als Blätter verfügbar. Mit der Wiedervereinigung wird das Blatt *DDR* gelöscht und die dazugehörigen Daten der *BRD* zugeordnet. Der Begriff *BRD* ist also syntaktisch gleich geblieben, beschreibt nun aber einen deutlich größeren Bereich.

Würde man die Daten der BRD betrachten, so könnte man z.B. in 1991 eine deutliche Steigerung der Einwohner-

zahl feststellen. Dies wäre aber nicht durch hohe Geburtsraten begründet, sondern durch die größere betrachtete Fläche infolge der Wiedervereinigung mit der DDR.

Die umgekehrte Richtung kann man bei der Tschechoslowakei beobachten. Bis 1990 war es die *ČSSR*, dann wurde das gleiche Land umbenannt in *ČSFR* und im Jahr 1992 aufgeteilt in die zwei Staaten Slowakei *SR* und Tschechien *ČR*. Für den letzt genannten Fall würde in der Dimension ein Blatt gelöscht und dafür zwei neue Blätter eingefügt. Die dazugehörige Transformationsregel wäre, dass es keine Möglichkeit gibt, *ČSFR* auf *SR* und *ČR* abzubilden, wohl aber in der Gegenrichtung.

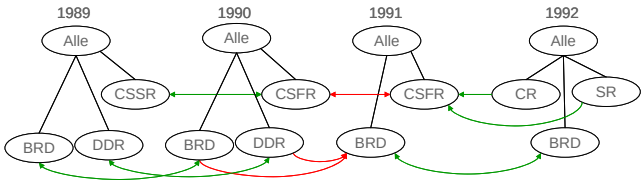

Abbildung 5: Darstellung von vier Ausschnitten einer Länderdimension für die Jahre 1989 bis 1992 und deren Überleitungen

Die graph-basierte Visualisierung der beschriebenen Dimensionsentwicklung ist in Abbildung 5 zusehen. Dazu muss gesagt werden, dass der Aufbau und die Entwicklung der Dimension von Fachexperten für Analysezwecke durchgeführt wurde. Die Dimensionen und Transformationen hätten auch auf andere Arten modelliert werden können.

3. EXISTIERENDE LÖSUNGANSÄTZE

Der erste Lösungsansatz für das Problem der sich ändernden Dimensionen wurde 1993 von Kimball postuliert [11]. Die Lösung besteht in der Umcodierung der Daten nach der jeweilig neuesten Dimensionsbeschreibung. Dies kann dabei in drei verschiedenen Arten geschehen. Der *Type 1* Ansatz überschreibt die alten Werte mit den neuen, umcodierten Werten. Die *Type 2* und *Type 3* Ansatz behält die alten Werte zusätzlich bei. Auf diese Weise können alte Werte in die neue Dimension transformiert, bzw. eingebunden werden. Der Nachteil aller dieser Ansätze ist aber, dass sie nicht in der Lage sind mit dem *Semantic Shift* syntaktisch gleicher Ausprägungen umzugehen. Es gibt also keine Unterstützung für Datenanalysen, die über verschiedene Versionen der Dimensionen hinausgehen, wenn sich die Bedeutung der Daten geändert hat.

Das Problem der Anfragen über mehrere Dimensionsversionen wurde 2006 in [8] als graphentheoretisches Problem diskutiert. Dabei wurden die Metadaten als sogenannte *Schemagraphen* repräsentiert. Für die Graphen wurden erlaubte Modifikationen definiert, welche die potentiellen Änderungen der Dimensionen wiedergeben. Wird eine Dimension durch eine Modifikation geändert, wird dies als neue Version in einem Graphen gespeichert. Basierend auf einer Graphenalgebra ist es dadurch möglich, Anfragen über verschiedene Dimensionsversionen hinweg zu stellen. Diesem Ansatz fehlt zum einen der Umgang mit dem *Semantic Shift* der Daten. Zum anderen erscheint er nicht praxisgerecht, da für historische Daten neu hinzugekommene Angaben nachträglich eingepflegt werden müssen, um Vergleiche über verschiedene Versionen zu ermöglichen.

Die am weitest gehende Lösung für das präsentierte Problem wurde 2002 in [12] veröffentlicht. Ein formales Temporal-Modell für die Beschreibung von Änderungen in den Dimensionen wurde dazu eingeführt [7]. Es wurden entsprechende Transformationsfunktionen definiert, welche die erlaubten Datenänderungen beschreiben. Der Ansatz ermöglicht dabei Anfragen über verschiedene Versionen der Dimensionen hinweg, indem die Daten zur Anfragezeit adaptiert werden. Der Nachteil des Ansatzes liegt in der Realisierung der Instanzadaption durch die Verwendung von Matrizenmultiplikation. Jeder Wert einer Dimensionsversion muss von Fachexperten mit einem Koeffizienten versehen werden, der aussagt wie ähnlich der Wert dem Nachfolger in der verbundenen Dimensionsversion ist. Dies erlaubt eine Abschätzung, um den *Semantic Shift* zu lösen. Jedoch hat dies zwei Nachteile. Zum einen muss der Koeffizient für jede Verwendung der Dimension in einer Kennzahl individuell angegeben werden, da sich die Koeffizienten für z.B. Erkrankungs- und Sterberisiko unterschiedlich verhalten und deshalb die Koeffizienten nicht für alle Analysen gleich sind. Zum anderen wird das in den Transformationsdaten inhärente Wissen nicht dazu genutzt, genaue anstatt approximierten Ergebnissen zu liefern.

4. DER GRAPH-BASIERTE ANSATZ

Wie in der Problembeschreibung ausgeführt und in den Abbildungen 4 und 5 veranschaulicht, lassen sich Dimensionen als streng hierarchische Bäume mit einem Wurzelknoten darstellen. Die Blätter repräsentieren dabei in der Regel die im Data Warehouse speicherbaren Werte. Falls Analysen auf den Elternknoten durchgeführt werden sollen, werden diese standardmäßig aus den Kindelementen berechnet [2].

Bei den Dimensionen handelt es sich um von Fachexperten modellierte Metadaten, die nur zu bestimmten Zeitpunkten geändert werden. Deshalb ist es möglich, die Änderungen einer Dimension zusammen mit einer Versionsnummer zu speichern. Dieser Ansatz zur Beschreibung der zeitlichen Entwicklung wurde auch von [12] und [8] verfolgt. Anders aber als bei [11] soll keine Instanzadaption mit Informationsverlust vorgenommen werden, sondern die Transformationsregeln als gerichtete Kanten zwischen den Blättern zweier Dimensionsversionen gespeichert werden. Es wird verlangt, dass jede neue Version einer Dimension Transformationsregeln zu mindestens einem Vorgänger definiert. Dies ist keine Einschränkung, da es beim Fehlen von Transformationsregeln nicht um einen Nachfolger der Dimension sondern um eine vollständig neue, andere Dimension handelt.

Bei einer Anfrage an das Analytische Informationssystem soll ein Interpreter zwischen den Anwender und das Auswertungssystem geschaltet werden. Dieser Interpreter wertet die Transformationsregeln aus und stellt fest, ob in dem angefragten Zeitraum für die auszuwertenden Daten eine Änderung stattgefunden hat. Wenn dies nicht der Fall ist, wird die Anfrage ohne Nutzerinteraktion und ohne Änderungen durchgeführt. Falls jedoch zwei oder mehrere Dimensionsversionen von der Anfrage betroffen sind, wird der Interpreter mittels der ein- und ausgehenden Kanten der Knoten prüfen, ob auf zusätzliche Knoten über die Kanten zugegriffen werden kann. Wenn die Knoten für den gewünschten Zeitraum stabil sind, wird dem Nutzer die Veränderung der Dimension für seinen angefragten Ausschnitt als sogenannter Evolutionspfad angezeigt.

Da es nicht beabsichtigt ist, die Definition der Transfor-

mationsregeln auf genau einen Vorgänger und Nachfolger zu beschränken, kann es durchaus mehrere unterschiedliche Evolutionspfade geben, die zu unterschiedlichen Mengen von Knoten führen. Deswegen sollen die gefunden Evolutionspfade dem anfragenden Nutzer angezeigt werden, der dann den für seine Anfrage geeignetsten auswählen kann. Dabei ist festzustellen, dass die Bedeutung der Evolutionspfade immer der modellierten Realwelt einer Dimensionsversion entspricht. Dies führt dazu, dass die Daten dann ad-hoc zum Anfragezeitpunkt unter die ausgewählte Dimensionsversion transformiert werden. Die Datentransformation ist allerdings keine Instanzadaption, sondern eine Transformation eines Wertes auf eine Menge von Werten.

4.1 Beispiel der Lösungsidee

Falls der Nutzer eine Anfrage der Art *„Gib mir die Summe aller behandelten J10.0 Patienten der Jahre 2005 bis 2007“* stellt, wird der Interpreter die Werte *J10.0* und die ICD-GM Dimensionsversionen 2005, 2006 und 2007 identifizieren. Den Transformationsregeln in Abbildung 4 folgend, wird der Interpreter zwei verschiedene Arten von J10.0 feststellen: Die Version 2005 hat zwei eingehende Kanten aus der Version 2006, einmal vom *J10.0* als auch vom *J09* Knoten. Der Interpreter kann also feststellen, dass der Knoten *J10.0* Version 2005 geteilt wurde. Nun prüft der Interpreter die identifizierten Knoten des Jahres 2006 und findet zusätzlich nur bidirektionale Kanten zu den Knoten des Jahre 2007, was bedeutet, dass keine Änderung stattgefunden hat.

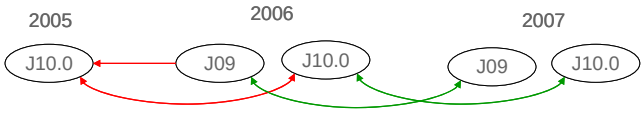

Abbildung 6: Lösungsvorschlag mit Erweiterung der Anfragemenge, Konzept *J10.0* Version 2005

Abbildung 7: Lösungsvorschlag mit Beschränkung des Anfragebereichs, Konzept *J10.0* Version 2006 und 2007

Dem Nutzer wird als Zwischenergebnis seiner Anfrage mitgeteilt, dass zwei verschiedene Interpretationen von *J10.0* für den Zeitraum 2005 bis 2007 identifiziert wurden. Es werden dann diese zwei verschiedenen Evolutionspfade zur Auswahl angeboten: In Abbildung 6 wird die Erweiterung der Anfrage für die Jahre 2006 und 2007 um den Wert *J09* vorgeschlagen, damit die Anfrage der Bedeutung von *J10.0* im Jahr 2005 entspricht. Die zweite angebotene Lösung ist in Abbildung 7 zu sehen. Hier wird die Bedeutung von *J10.0* der Jahre 2006 und 2007 vorgeschlagen.

Werden Anfragen auf höheren Ebenen der Dimension, wie z.B. die *Gib mir die Summer aller behandelten Fälle von „Grippen durch nachgewiesene Influenzaviren“* welche dem ICD Gruppencode *J10* des Jahres 2005 bzw. *J09-J10* der Jahre 2006 und 2007 entspricht, ist dies auch ohne weiteres möglich. Für alle Kindelemente von *J10* werden die Evolutionspfade ausgewertet. Da alle Kindelemente in den Dimensionsversionen unter einem Elternelement sind, wird davon ausgegangen, dass die Elternelemente gleich sind. Die Anfrage wird dann direkt ausgeführt. Sollte es in irgendeinem gültigen Evolutionspfad eines beliebigen Kindes mehrere Elternknoten geben, würden dem Anwender wieder die verschiedenen Optionen wie bei den Einzelelementen angezeigt. Der Anwender wird also immer über Bedeutungsänderungen und Systembrüche automatisch graphisch informiert und kann die für seine Zwecke geeignete Anfrage auswählen.

4.2 Vorteile des Ansatzes

Es wird erwartet und angestrebt, dass der vorgestellte Ansatz die folgenden Vorteile bietet:

- Durch den graphenbasierten Ansatz, der auf eine kennzahlabhängige Approximation verzichtet, ist es möglich die Überführungsregeln für alle Analyseanfragen zu verwenden, welche die Dimension beinhaltet. Dies ist eine deutliche Erweiterung gegenüber [12].
- Da die Überfühungsregeln auch in klassischen Adaptionsverfahren wie [11] benötigt werden, ist kein zusätzlicher Arbeitsaufwand der Fachexperten notwendig, um die Kanten bereit zustellen.
- Durch den graphenbasierten Ansatz ist es bei mehreren Überleitungsregeln pro Dimensionsversion möglich, gröbere Versionen einer Dimension zu überspringen. Gröber meint dabei, dass Fehlen einzelner Knoten, die in späteren Versionen wieder eingefügt wurden. Bei einer Umcodierung des Datenbestandes wäre dies ein irreversiebler Informationsverlust.
- Der Import und die Haltung der Daten wird vereinfacht, da die Daten in ihrer originären Version gespeichert werden können. Die Daten müssen nicht in eine einzige Version umcodiert werden.
- Da der Nutzer zwischen verschiedenen inhaltlichen Interpretationen eines Wertes wählen kann, ist das Anfragesystem mächtiger als klassische Systeme. Zudem erlaubt dies die Wiederholung historischer Analysen, da die Datenbasis nicht umcodiert und die Dimensionsdaten genauso erhalten bleiben.

4.3 Zu untersuchende Fragestellungen

Um sicherzustellen, dass ein Data Warehouse zusammen mit einer OLAP Analyse Anwendung die vorgestellten Funktionen und insbesondere Vorteile erfüllen kann, muss untersucht werden, welche Konsistenzbedingungen die Überleitungsregeln als auch die Metadaten einhalten müssen. Zudem sind die Anforderungen an die Datenrepräsentation und Speicherung der Transformationsregeln und zusätzlichen Versionsinformationen in den Metadaten als auch der Datenhaltung zu untersuchen. Ein weiterer Bereich ist, wie sich die Methoden auf verschieden Datenarten (Integer, Boolean, Nominal) als auch verschiedene Analyse Operationen (Sum, Max, Min, Average) anwenden lassen. Da in Analysen auch oft mehrere verschiedene Dimensionen genutzt werden, muss als letzter wichtiger Punkt noch die Anwendbarkeit auf mehrere Dimensionen durchdacht werden.

4.4 Evaluation

Um den Ansatz mit seinen Konzepten und festgelegten Anforderungen zu evaluieren, wird ein Prototyp auf Basis

von *MUSTANG - Multidimensional Statistical Data Analysis Engine* [1] [13] umgesetzt werden. Dies ist ein kommerzielles Daten Analyse Tool, welches insbesondere für Analysen im Gesundheitswesen, z.B. Krebsregistern, eingesetzt wird.

Da das vorgestellte Thema durch zwei Projekte mit Klinikdaten motiviert wurde, bei denen sich der *Semantic Shift* als problematisch erwiesen hatte, soll das Konzept in diesen evaluiert werden. Dabei handelt es sich zum einen um Daten deutscher Kliniken der Jahre 2006 bis 2010. Hier sollen Fragen der Versorgungsforschung auf einer feingranularen Ebene ausgewertet werden, was bisher nicht möglich war. Zum anderen geht es in einem Forschungsprojekt der EU darum, für spezielle Herzschrittmacherpatienten statistisch valide Muster zu identifizieren, die in historischen Patientendaten früherer Fälle enthalten sind. Die Patientendaten stammen dabei aus den Jahren 2006 bis 2011 eines österreichischen Universitätsklinikums, in dem später die Anwendung erfolgt. Hier liegt der Fokus darauf, alte Codierungen akkurat unter die aktuellste Version zu subsumieren, damit die Muster auf aktuelle Fälle angewendet werden können.

5. ZUSAMMENFASSUNG

Dieses Paper stellt einen Ansatz vor, der akkurate Datenanalysen in einem Analytischen Informationssystem über sich ändernde Datengrundlagen ermöglicht. Die Datenänderungen können dabei sowohl syntaktischer als auch semantischer Natur sein. Änderungen der Daten werden dabei als verbindende Kanten zwischen verschiedenen Versionen einer Dimension modelliert und diese Dimensionen dabei als Graphenstruktur aufgefasst. Durch die Interpretation der Verbindungen zum Zeitpunkt einer Analyseanfrage, werden die möglichen Evolutionspfade identifiziert. Die Evolutionspfade repräsentieren dabei domänspezifisches Hintergrundwissen, wie z.B. die Bedeutungsänderung von Werten, den *Semantic Shift*. Der Nutzer kann dieses Hintergrundwissen visuell erfassen und sich für einen geeigneten Evolutionspfad entscheiden. Die Analyseanfrage wird dann zur Anfragezeit so umgewandelt, dass die Daten ad-hoc unter die gewählte Bedeutung des Evolutionspfads transformiert werden. Da die Evolutionspfade so berechnet werden, dass Sie inhaltlich identische und vergleichbare Mengen repräsentieren, sind die Anfrageergebnisse akkurat. Dies wird dadurch ermöglicht, dass die Daten in ihrem Originalformat gespeichert und die Transformationsregeln nur gespeichert aber nicht direkt auf die Daten angewendet werden. Mit dem vorgestellten Modell und den dazugehörenden Methoden sind keine verlustbehafteten Datentransformationen oder Abschätzungen notwendig.

6. REFERENCES

[1] Appelrath, H.-J., Rohde, M. , Thoben, W. , OFFIS e.V., MUSTANG - Multidimensional Statistical Data Analysis Engine: `http://www.offis.de/en/offis_in_portrait/structure/projects/detail/status/mustang.html`, (2011)

[2] Bauer, A., Günzel, H.: Data Warehouse Systeme. dpunkt.verlag, 3. überarbeitete und aktualisierte Auflage, (2009)

[3] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM 2005 Systematisches Verzeichnis. Systematisches Verzeichnis zur Internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme - German Modification. Deutsche Krankenhaus Verlags-Gesellschaft, (2004)

[4] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2006. Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2005).

[5] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2007. Band I: Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2006)

[6] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2011: Band I: Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2010)

[7] Eder, J, Koncilia, C., Morzy, T.,: The ℂOMET Metamodel for Temporal Data Warehouses. In Proc. of the 14th Int. Conference on Advanced Information Systems Engineering (CAISE02), pp. 83–99.Springer Verlag (LNCS) (2002)

[8] Golfarelli, M., Lechtenbörger, J., Rizzi, S., Vossen, G.: Schema versioning in data warehouses: enabling cross-version querying via schema augmentation. In DataKnowl. Eng., 59(2):435–459, 2006. Elsevier Science Publishers B. V., Amsterdam, (2006)

[9] Inmon, W. H.: Building the data warehouse (2nd ed.). John Wiley & Sons, Inc.,New York, NY, USA, (1996)

[10] Inmon, W. H., Strauss, D., Neushloss, G.: DW 2.0: The Architecture for the Next Generation of Data Warehousing. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (2008)

[11] Kimball, R.: Slowly Changing Dimensions. In DBMS online, `http://www.dbmsmag.com/9604d05.html` (1996)

[12] Koncilia; C. A.: The ℂOMET Temporal Data Warehouse. PhD thesis, Universität Klagenfurt (2002)

[13] Teiken, Y., Rohde, M., Mertens, M.: Mustang - Realisierung eines analytischen informationssystems im kontext der gesundheitsberichtserstattung. In K.-P. Fähnrich and B. Franczyk, editors, GI Jahrestagung (1), volume 175 of LNI, pages 253–258. GI, (2010)

APPENDIX

A. ACKNOWLEDGMENTS

The research leading to these results has received in part funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreement no. ICT-248240, iCARDEA project.