Ontologiebasierte domänenspezifische Datenbereinigung in Data Warehouse Systemen

Stefan Brüggemann

Institut OFFIS, 26121 Oldenburg, Germany,
email: brueggemann@offis.de,
WWW home page: http://www.offis.de

Zusammenfassung. Datenbereinigung wird häufig im Transformationsschritt von ETL-Prozessen in Data Warehouses durchgeführt. Existierende Ansätze fokussieren auf syntaktische Korrektheit der vorliegenden Daten, indem beispielsweise Formatprüfungen und schemabasierte Bereinigungen durchgeführt werden. Weitere Ansätze konzentrieren sich auf semantische Prüfungen in Form von Duplikaterkennungen oder unerlaubten Werten. Data Warehouses werden oft in speziellen Anwendungsdomänen, wie beispielsweise dem Gesundheitswesen, eingesetzt. Dort existiert domänenspezifisches Wissen über die zu bereinigenden Daten, welches in bisherigen Ansätzen nur unzureichend berücksichtigt wird.
In diesem Beitrag wird ein ontologiebasierter Ansatz zur Datenbereinigung vorgestellt, in welchem dieses Wissen geeignet modelliert werden kann und während der Datenbereinigung Konzepthierarchien berücksichtigt werden können, welche semantische Zusammenhänge in Datenbeständen erkennen lassen können.

1 Einführung

Ein Data Warehouse System (DWS) ist eine physische Datenbank, die eine integrierte Sicht auf (beliebige) Daten darstellt [1]. In DWS werden Daten aus verschiedenartigen Datenquellen integriert, um auf diesen Daten Analysen durchzuführen, beispielsweise in der Krebsepidemiologie oder in betriebswirtschaftlichen Kontexten. Die Datenintegration findet meist im Rahmen von ETL-Prozessen (Extraktion, Transformation, Laden) [2] statt. Während der Extraktion wird festgelegt, welche Daten der Quellsysteme in das DWS überführt werden, in der Transformationsphase werden zum einen Konvertierungen zwischen Datenstrukturen, Kodierungen und Datentypen durchgeführt, zum anderen liegt ein Schwerpunkt auf der Datenbereinigung (Data Cleaning) im Rahmen der Optimierung der Datenqualität. Das Laden letztlich schreibt die Daten in die Datenbank des Ziel-DWS.

Auf Basis der Datenbestände von DWS-Systemen werden häufig weitreichende Entscheidungen getroffen [3]. Dies verlangt nach einer sehr hohen Datenqualität [4], da fehlerhafte oder inkonsistente Daten zu Fehlentscheidungen führen können („Garbage in, Garbage out“[5]).

Eine einheitliche Definition des Begriffs *Data Cleaning* gibt es nicht, in allen Verwendungen des Begriffs wird allerdings eine Verbesserung der Qualität betrachteter Daten angestrebt. Fehler in Daten können auf Tupelebene, beispielsweise bei der Verwendung falscher Werte, falscher Formate, fehlender Werte, aber auch auf Schemaebene, wie Mißachtungen referentieller Integrität oder Eindeutigkeitsverletzungen, existieren.

In diesem Beitrag wird zunächst ein Überblick über bisherige Vorgehensweisen im Data Cleaning, besonders über Ansätze, welche Ontologien zur Wissensmodellierung nutzen, gegeben. Im Anschluß daran wird ein Ansatz zur domänenspezifischen Datenbereinigung im ETL-Prozeß von DWS mittels Ontologien vorgestellt. Mit einer Zusammenfassung schliesst dieser Beitrag ab.

2 Bestehende Ansätze

Existierende Ansätze im Data Cleaning fokussieren oftmals auf syntaktische und semantische Fehler. Einen Überblick hierzu geben unter anderem [5–7]. Frameworks und Werkzeuge wie AJAX [8] und Potter's Wheel [9] bieten unter anderem Methoden zur schemabasierten Datenbereinigung, zur Typ- und Formatprüfung, zur Lösung des *Object Identity Problem* in der Duplikaterkennung, zur Detektion von fehlenden Werten oder Verletzungen referentieller Integritäten.

Wie auch in [7] dargestellt, existieren bisher kaum Ansätze zur Datenbereinigung unter Ausnutzung domänenspezifischen Wissens. In [10] werden Ontologien zum Data Cleaning von XML-Dateien genutzt. Hier wird von einer existierenden Ontologie ausgegangen, zu welcher dann eine Abbildung der DTD definiert wird. Diese Ontologie wird dann benutzt, um ein Dokument auf Validität und sogenannte *Ontologie-Validität* zu prüfen. Hier wird allerdings nicht auf die Erstellung der Ontologie und auf Bereinigung der Daten eingegangen. [11] verwendet Ontologien zur Datenintegration, wobei Schemakonflikte beispielsweise mit Hilfe von Konzepthierarchien aufgelöst werden. Domänenspezifisches Wissen wird in [12] mittels Ontologien modelliert. Auf diesen Ontologien können Nutzer dann Aktionen definieren, die auf den zugrunde liegenden Daten durchgeführt werden. Dort wird jedoch nicht auf Erstellung und Pflege der Ontologie eingegangen.

3 Domänenspezifisches Data Cleaning

Bisherige Ansätze im Data Cleaning fokussieren nicht auf die Verwendung domänenspezifischen Wissens. Domänenexperten können Aussagen über Datenbestände treffen, die aus syntaktischen oder Datenschemaanalysen nicht herleitbar sind, da sie über domänenimmanentes Wissen über diese Daten verfügen. In der Krebsepidemiologie beispielsweise wird die TNM-Klassifikation [13] zur Beschreibung und anatomischen Einordnung bösartiger Tumoren verwendet. Die Kategorien T, N und M bezeichnen jeweils unterschiedliche Attribute von Tumoren und verfügen jeweils über einen eigenen Wertebereich. Kombinationen dieser Werte unterliegen strengen Restriktionen, die zusätzlich von weiteren Faktoren, wie beispielsweise der Lokalisation eines Tumors, abhängen können. Diese Werte

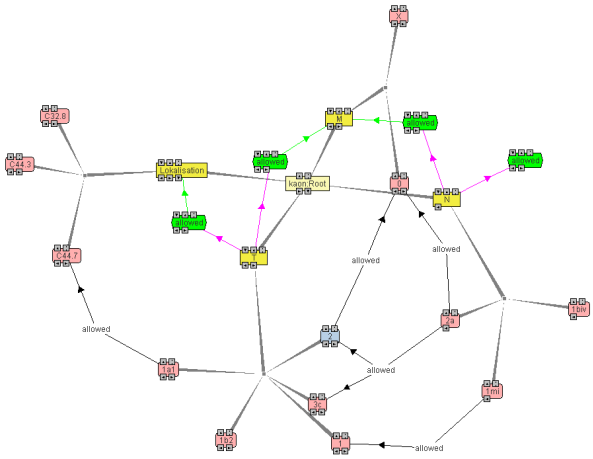

Abb. 1. TNM-Klassifikation mit Beispielinstanzen, erstellt mit KAON Workbench (http://kaon.semanticweb.org, zuletzt besucht am 24.04.2006)

können zwar jeweils eigenständig mit den in Kapitel 2 beschriebenen Methoden des syntaktischen Data Cleaning bereinigt werden, jedoch lassen sich Kombinationen nicht mehr auf triviale Weise prüfen und bereinigen.

Zur Modellierung dieses Wissens werden Ontologien eingesetzt. Ontologien dienen der formalen Spezifikation einer Konzeptualisierung [14]. Die Konzeptualisierung einer Anwendungsdomäne besteht aus der Klassifizierung von Konzepten und der Verbindungen untereinander im Sinne eines semantischen Netzes [15]. Abbildung 1 beschreibt einen Ausschnitt aus einer Ontologie mit den Konzepten T, N, M und Lokalisation, jeweils mit Beispielinstanzen (z.B. „2a“, “1mi“ und „1biv“ für N). Weiter sind mit „allowed“ Verbindungen zwischen diesen Instanzen definiert, welche gültige Verwendungen von Tupeln definieren. So sind beispielsweise die Paare („2a“, “0“) und („2a“,“3c“) gültig, aber nicht das Tripel („2a“, “0“, “3c“) als Instanzen von T, N und M.

Wird dieser Ansatz zur Datenbereinigung im ETL-Prozeß von DWS verwendet, so kann aus der dort vorhandenen Basisdatenbank eine Ontologie erstellt werden, wenn diese Datenbasis als valide angesehen wird. Die dort existierenden Daten werden Instanzen der erstellten Ontologie, und die dort vorhandenen Kombinationen werden valide Tupel in der Ontologie. Data Mining Technologien können eingesetzt werden, um solche validen Tupel aus der Datenbank zu extrahieren. Weiter können Benutzer, Domänenexperten wie beispielsweise medizinische Dokumentare, hier gültige Verbindungen definieren.

Dies ist ein Vorteil gegenüber der Benutzung virtueller Ontologien. Virtuelle Ontologien werden nicht explizit modelliert; stattdessen wird eine Abbildung auf die konkreten Daten innerhalb eines DWS erstellt. Der prinzipielle Vorteil, daß so die Ontologie stets aktuell ist, hat in DWS keine Bedeutung, da dort historische Daten gespeichert sind, die zu definierten Zeitpunkten modifiziert

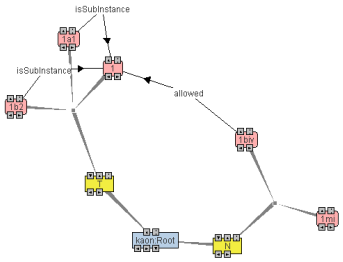

Abb. 2. Hierarchie zwischen Werten der TNM-Klassifikation

werden. In diesen virtuellen Ontologien könnten Experten das immanente Wissen nicht vermehren, da die Ontologie nicht manuell erweiterbar ist.

Im Transformationsschritt des ETL-Prozesses werden die zu integrierenden Daten zuerst syntaktisch bereinigt, anschliessend wird Data Cleaning mit Hilfe der Ontologie durchgeführt. Werden Datensätze hier als fehlerhaft erkannt, so sind mehrere Szenarien denkbar:

- Der identifizierte Datensatz wird gelöscht, dies kann zu einem Datenverlust führen, falls der Datensatz nicht vollständig fehlerhaft war.
- Der Datensatz wird zur späteren manuellen Bearbeitung gekennzeichnet, hier können Domänenexperten entscheiden, ob der Datensatz semantisch korrekt ist. In diesem Falle kann die Ontologie um das gewonnene Wissen erweitert werden.
- Der fehlerhafte Datensatz kann automatisch bereinigt werden, falls zwischen den Instanzen Hierarchien definiert sind. In Abbildung 2 sind die Instanzen „1a1“ und „1b2“ Spezialisierungen der Instanz „1“ des Konzeptes „T“. Die Instanz „1biv“ des Konzepts „N“ ist nur im Kontext mit „1“ valide. Wird während der Datenbereinigung das Tupel („1biv“, “1a1“) erkannt, so kann dieses automatisch mit der Generalisierung des fehlerhaften Wertes ersetzt werden, so daß das Tupel zu („1biv“, „1“) korrigiert werden kann.

Der letzte Ansatz ermöglicht die größte Automatisierbarkeit des Bereinigungsschrittes, da er keine Benutzerinteraktion erfordert. Dies ist in Szenarien sinnvoll, in denen der Ontologie kein Wissen mehr zugeführt werden wird. In anderen Situationen ist der zweite Ansatz sinnvoll, da hier die Ontologie während der Bereinigung optimiert werden kann.

Treten bei der Integration heterogener Datenquellen in ETL-Prozessen Konflikte auf, läßt sich deren Semantik bisher nur unzureichend erkennen, da beispielsweise kein Vergleich zweier Informationen aus verschiedenen Quellen möglich ist. Durch Einsetzung des in dieser Arbeit vorgestellten Ansatzes zur Datenbereinigung im Integrationsprozeß von Data Warehouses lassen sich diese Daten semantisch integrieren, da Unzulänglichkeiten in Datenquellen durch eine Domänenontologie aufgelöst werden können.

4 Zusammenfassung

In dieser Arbeit wurde ein Ansatz zur ontologiebasierten Datenbereinigung im ETL-Prozeß von DWS-Systemen vorgestellt, in welchem domänenspezifisches Wissen zum Data Cleaning genutzt werden kann. Die initiale Erstellung der Ontologie kann auf Grundlage der Basisdatenbank des DWS erfolgen; Instanzen der Ontologien können mit Data Mining Techniken identifiziert werden. Während der Datenbereinigung werden Konzepthierarchien zur Eleminierung von Fehlern genutzt und neu gefundene valide Tupel können in die Ontologie integriert werden. Diese Ontologie eignet sich ebenfalls zur Integration von Daten aus diversen Datenquellen, da diese einen Vergleich der Daten ermöglicht.

Literatur

1. Bauer, A., Günzel, H.: Data Warehouse Systeme - Architektur, Entwicklung, Anwendung. dpunkt.verlag (2004)
2. Kurz, A.: Data Warehousing Enabling Technology. MITP-Verlag (1999)
3. Inmon, W.: Building the Data Warehouse. ", John Wiley and Sons Inc., 2nd Edition, New York (2002)
4. Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen. PhD thesis, Universität Oldenburg (2002)
5. Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering **23(4)** (2000) 3–13
6. Heiko Müller, J.C.F.: Problems, methods, and challenges in comprehensive data cleansing. Technical report, Humboldt University Berlin (2003)
7. Brüggemann, S., Rohde, M.: Caramel: A plugin-architecture for the secure integration of standards in medical information systems. In Malina Jordanova, F.L., ed.: E-Health, Proceedings of Med-e-Tel 2006. (2006) 165–170
8. Galhardas, H., Florescu, D., Shasha, D., Simon, E.: Ajax: An extensible data cleaning tool. In: Proceedings of the ACM SIGMOD Conference. (2000)
9. Raman, V., Hellerstein, J.M.: Potter's wheel: An interactive data cleaning system. In: Proceedings of the 27th International Conference on Very Large Databases (VLDB 2001). (2001)
10. Milano, D., Scannapieco, M., Catarci, T.: Using ontologies for xml data cleaning. In Robert Meersman, Zahir Tari, P.H., ed.: On the Move to Meaningful Internet Systems 2005. Volume 3762., LNCS, Springer Berlin / Heidelberg (2005)
11. Kedad, Z., Métais, E.: Ontology-based data cleaning. In: Natural Language Processing and Information Systems: 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002. (2002)
12. Wang, X., Hamilton, H.J., Bither, Y.: An ontology-based approach to data cleaning. Technical report, Department of Computer Science, University of Regina (2005)
13. International Union Against Cancer (UICC): TNM Classification of Malignant Tumours, 6th edition. John Wiley & Sons, Hoboken, New Jersey (2001)
14. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5(2)** (1993) 199–220
15. Helbig, H.: Die semantische Struktur natürlicher Sprache. Springer, Berlin Heidelberg (2001)