CARL VON OSSIETZKY universität OLDENBURG

Fakultät II – Informatik, Wirtschafts- und Rechtswissenschaften
Department für Informatik

Ad-hoc-Datentransformationen für Analytische Informationssysteme

Dissertation zur Erlangung des Grades eines
Doktors der Ingenieurwissenschaften

vorgelegt von

Dipl.-Inform. Christian Lüpkes

Gutachter:

Prof. Dr. Dr. h.c. H.-Jürgen Appelrath
Prof. Dr. rer. oec. Carsten Felden

Tag der Disputation: 26. Februar 2014

Zusammenfassung

Beim Betrieb von Data Warehouses kann es aufgrund ”äußerer Einflüsse” in der durch ein Data Warehouse modellierten Realwelt zu semantischen Inkonsistenzen (”Semantic Shift”) kommen, wenn sich in diesem Data Warehouse die Bedeutung von Dimensionselementen über die Zeit ändert. Bei Nichtbeachtung dieser Veränderungen kann es zu Informationsverlust und/oder inkorrekten Analyseergebnissen kommen. Im Rahmen der vorgelegten Dissertation wird ein Lösungsansatz zur Verhinderung semantischer Inkonsistenzen für auf einem multidimensionalen Modell basierende Analytische Informationssysteme vorgeschlagen. Dieser Ansatz macht die Änderungen der Daten für einen Analysten transparent und kommt ohne nachträgliche – potentiell mit Informationsverlust behaftete – Datenadaptionen aus. Dafür wird unter Berücksichtigung der analytischen Ausrichtung der Daten das multidimensionale Modell auf mögliche Änderungen untersucht und es werden Regeln aufgestellt, wie sich diese Änderungen abbilden und konsistent auswerten lassen. Die zentrale Fragestellung der beschriebenen Problemstellung lautet:

> Wie lassen sich Dimensionsänderungen in Analytischen Informationssystemen so fortschreiben, dass implizites Wissen sichtbar und für Auswertungen ohne zusätzlichen Informationsverlust nutzbar gemacht werden kann?

Der Vorschlag zur Beantwortung dieser Fragestellung ist ein *GrAHD – Graphenbasierte Ad-hoc-Datentransformation* genannter Ansatz, der Änderungen in den Dimensionen visualisiert und für Auswertungen verwendbar macht. Datenänderungen können dabei sowohl syntaktischer als auch semantischer Natur sein und werden als verbindende Kanten zwischen verschiedenen Versionen einer Dimension modelliert, wobei diese Dimensionen als Graphenstruktur aufgefasst werden. Durch die Interpretation der Verbindungen zum Zeitpunkt einer Analyseanfrage werden die möglichen Evolutionspfade identifiziert, die aus Mengen von verbundenen Dimensionselementen verschiedener Versionen bestehen. Die Evolutionspfade repräsentieren dabei domänenspezifisches Hintergrundwissen, wie z. B. die Bedeutungsänderung von Werten über die Zeit, den hier sogenannten *Semantic Shift*. Nutzer können dieses Hintergrundwissen visuell erfassen und sich für einen geeigneten Evolutionspfad entscheiden. Die Analyseanfrage wird dann zur Anfragezeit so umgewandelt, dass die Daten ”ad hoc”, d. h. zum Zeitpunkt der Anfrage und speziell für deren Zwecke, unter die gewählte Bedeutung des Evolutionspfads transformiert werden. Da die Evolutionspfade derart berechnet werden, dass sie inhaltlich vergleichbare Mengen repräsentieren, sind die Ergebnisse im Sinne der intendierten Anfrage ”akkurat”. Akkurat bedeutet hierbei, dass sie die modellierte, tatsächliche Entwicklung der Elemente wiedergeben und dabei im Vergleich zu bisherigen Ansätzen aus der Literatur auf Approximationen verzichten. Dies wird dadurch ermöglicht, dass die Daten in ihrem Originalformat und damit in ihrer originären Bedeutung gespeichert bleiben und die – normalerweise zur Datenadaption verwendeten – Transformationsregeln nur gespeichert, nicht aber direkt zur potentiell mit Informationsverlust behafteten Adaption der Daten verwendet werden.

Abstract

During the lifetime of a data warehouse, semantic inconsistencies can occur in the designed real-world-model of the data warehouse caused by changes to the meaning of dimension members due to external factors, the so-called Semantic Shift. Insufficient handling of these changes can lead to a loss of information and/or false analysis results. This thesis presents a solution to prevent semantic inconsistencies for analytical information systems based on a multidimensional data model. It makes the changes of the model discernable to the user without the need to convert the stored data which could cause a loss of information. For this purpose, the possible changes occurring in the multidimensional data model are identified and investigated, taking into account the analytical purpose of the data. Afterwards, rules are formulated as to how these changes can be modeled and utilized in a consistent manner. The central question is:

> How can changes in the multidimensional model of an analytical information system be maintained so that tacit knowledge about the changes can be visualized and employed to compute analysis results without a loss of information?

The described solution called *GrAHD – Graph Based Ad Hoc Data Transformation* is capable of visualizing changes to dimension data and generating comparable analysis results. Changes in the dimension data can be of syntactic or semantic nature and are represented as edges between versions of dimensions which are treated as graph structures. So-called evolution paths are constructed ad hoc, by computational interpretation of the connections between different versions at the time of an analysis query. These paths consist of dimension elements which originate from different dimension versions and represent domain specific knowledge (like the change to the semantics of elements over time, the so-called Semantic Shift). Users can interpret the knowledge visualization and decide which evolution path suits their analysis question best. The question is then transformed so that the stored data is adapted to match the semantics represented by the evolution path. The evolution paths are generated with the objective to provide semantically comparable sets of elements so that the results of the question accurately reflect the user's intention. Accurate means that the results represent the modeled factual development of the elements and, compared to previous approaches, do not require approximations. This is made possible by keeping the data stored in its original format and thereby preserving its original meaning. The transformation rules – normally used for data adaption – are just stored but not applied directly to the data, which could cause a loss of information.

Inhalt