Michael Mielke (Bahn AG), Heiko Müller und Felix Naumann (Humboldt-Universität zu Berlin)

Ein Data Quality Wettbewerb

Daten und Informationen sind in den letzten Jahren zum wichtigsten Produktions- und Erfolgsfaktor geworden. Als Väter dieser Wissenschaft gelten heute Richard Y. Wang und Stuart E. Madnick vom Massachusetts Institute of Technology. Im Mai 1992 starteten sie ein Forschungsprogramm zum Total Data Quality Management, das u.a. von Unternehmen wie Fujitsu Personal Systems Inc. unterstützt wird. Mit der ersten Deutschen Information Quality Management Konferenz 2003 (GIQM´03) begann sich unter Führung von R. Wang, Michael Mielke und Marcus Gebauer die deutsche IQ-Community zu formieren. Es entstand u.a. eine branchenübergreifende Arbeitsgruppe (AG IQM) in der zahlreiche Unternehmen aus Deutschland, den Niederlanden und der Schweiz sich regelmäßig zum Erfahrungsaustausch in Frankfurt treffen. Im November 2003 entstand die Idee, im Zusammenhang mit einer Konferenz einen Datenqualitäts-Wettbewerb auszurichten. In einem solchen Wettbewerb sollten die IQ-Experten ihr Wissen und ihre Methoden an einem IQ-Problem aus der Praxis anwenden und veranschaulichen. Die teilnehmenden Teams sollten dabei interdisziplinär und nach Möglichkeit unternehmensübergreifend zusammen arbeiten. Das zu lösende Problem sollte aus der Praxis kommen, keine manipulierten Daten zur Grundlage haben und Lösungswege zeigen, die neue Erkenntnisse liefern. Den Teams stand es dabei frei, eigene oder am Markt etablierte Tools zur Lösung der Aufgabe zu verwenden.

Der IQ Contest 04 begann mit Ausgabe der Aufgabe selbst, einigen Datenstrukturen und einem kleinen Auszug der Daten im September 2004 und wurde parallel zur 2. Deutschen IQ Konferenz im November 2004 erfolgreich durchgeführt. Am IQ-Contest 2004 haben die folgenden Teams teilgenommen:

DQ-Tigers

- K. Libor und
- M. Weitz, Dresdner Bank AG

DQ-Dragons

- Dr. U. Windheuser und
- P. Caspers, WestLB AG

DQ-Knights

- M. Dobrat und
- L.Ritter, QlikTech Deutshland GmbH

AAA-Quality

- E. Bogner, und
- R. Kämmerer, Evoke Software
- T. Drummond, Similarity Systems
- J. Hüfner und
- R. Becher, TIQ Solutions GmbH

Aufgabenstellung

Die Aufgabenstellung des Wettbewerbs entstammte einem Problem aus dem Bereich der Datenintegration. Vorlage waren zwei sich überlappende Datenquellen mit dem übergeordneten Ziel, diese zu einer Datenquelle zu vereinen. Die Daten folgten dem gleichen Schema und repräsentierten eine identische bzw. überlappende Menge an Objekten.

Überlappende Datenquellen bergen die Gefahr sich widersprechender Informationen, d.h., die Repräsentation ein und desselben Objekts mit unterschiedlichen Werten. Gründe für diese Datenkonflikte sind u.a. Modifikation oder Transformation replizierter Informationen, unterschiedliche Aktualität der Daten oder ungenaue Messungen bzw. systematische Fehler im Rahmen der Datengenerierung. Im Zuge der Integration müssen diese Konflikte aufgelöst werden. Insbesondere ist die Entdeckung von Regelmäßigkeiten in den sich widersprechenden Daten interessant. Diese können einem Domänenexperten Hinweise auf mögliche Ursachen der Konflikte im Rahmen der Datenproduktion und -verarbeitung liefern und so zur qualitativen Bewertung der sich widersprechenden Daten und zur Konfliktlösung verwendet werden.

Den Wettbewerbsteilnehmern wurden zwei Datenbestände aus dem Bereich der Proteinstrukturforschung zur Verfügung gestellt. Beide Datenbestände resultieren aus einer Transformation der Einträge im zentralen Datenbestand zur Aufklärung der 3D-Struktur von Proteinen (PDB). Einer der Datenbestände entstammt dem Integrationsprojekt COLUMBA[1]; der zweite dem manuellen Datenbereinigungsprojekt OpenMMS[2]. Ausschnitte beider Datenquellen wurden für diese Aufgaben auf jeweils eine relationale Instanz desselben Schemas abgebildet. Dieses Schema umfasste insgesamt 15 Attribute, deren Semantik im Folgenden kurz dargelegt wird:

Schema der Wettbewerbsdaten

A_1 Eindeutige Bezeichnung des Eintrags.
A_2 Kurze Beschreibung der primären Eigenschaften der repräsentierten Proteinstruktur.
A_3 Einreichungsjahr des Eintrags.
A_4 Einreichungsdatum des Eintrags.
A_5 Veröffentlichungsdatum des Eintrags.
A_6 Autorenliste.
A_7 Experimentelle Methode zur Aufklärung der Struktur.
A_8 Maximale verwendete Auflösung.
A_9 Zur Nachbearbeitung verwendete Software.
A_{10} Anzahl möglicher Strukturen, die sich bei der Verwendung von NMR (*nuclear magnetic resonance*) zur Strukturaufklärung berechnen lassen.
A_{11} Anzahl unterschiedlicher Aminosäureketten der dargestellten Proteinstruktur.
A_{12} Gesamtzahl an Aminosäuren in den Ketten.
A_{13} Gesamtzahl an Atomen in den Ketten.
A_{14} Anzahl an Fremdmolekülen, sog. Heterogruppen, in der Proteinstruktur.
A_{15} Gesamtzahl an Atomen in den Heterogruppen.

Die Existenz eines Primärschlüssels ermöglicht die einfache Identifikation der sich entsprechenden Einträge (Duplikate) aus den beiden Datensätzen. Insgesamt existierten 23.614 solcher Paare von Duplikaten zwischen den beiden Quellen. Jede Quelle für sich ist frei von Duplikaten.

Im Rahmen des Wettbewerbs sollte zunächst ermittelt werden, wie stark sich die Datenbestände unterscheiden. Hierfür war pro Attribut die Anzahl der Widersprüche (zwei unterschiedliche Werte) und die Anzahl der Unsicherheiten (ein Wert und ein NULL-Wert) innerhalb der Duplikatpaare zu bestimmen. Zusätzlich sollte die Anzahl der widerspruchsfreien Duplikat~~en~~paare angegeben werden. Die wesentliche Aufgabe bestand dann im Auffinden von Regelmäßigkeiten in den sich widersprechenden Daten. Die Beschreibung dieser systematischen Widersprüche erfolgt dabei pro Attribut, d.h. jede Regel beschreibt mögliche Ursachen für Konflikte und de-

[1] http://www.columba-db.de/
[2] http://openmms.sdsc.edu/

ren Ausprägungen in dem jeweils betrachteten Attribut. Einfache Beispiele solcher Regeln besagen, dass bei widersprüchlichen Werten in einem Attribut (z.B. Autoren) der Wert der einen Quelle immer vollständig in dem der anderen Quelle enthalten ist. Es existieren aber auch komplexere Regeln, die mehrere Attribute in Beziehung setzen. Die gesuchten Regeln konnten beispielsweise in Form von Assoziationsregeln mit fester Conclusio dargestellt werden, d.h.

WENN *Bedingung* DANN *Konflikt in* A_i

Die gefundenen Regeln sollten sowohl quantitativ, z.B. mit Hilfe der Maße *support* und *confidence*, als auch qualitativ, d.h. hinsichtlich ihrer Nützlichkeit, von den Teilnehmern bewertet werden. Der *support* einer Assoziationsregel wie oben beschrieben bezeichnet deren statistische Relevanz, d.h., die relative Häufigkeit ihres Auftretens bezogen auf die Gesamtmenge der Daten. Die *confidence* wird gerne als die Verlässlichkeit einer Regel bezeichnet. In unserem Fall handelt es sich um die bedingte Wahrscheinlichkeit dafür, dass ein Paar von Duplikaten einen Konflikt in einem Attribut A_i aufweist, unter der Voraussetzung, dass der Bedingungsteil erfüllt ist.

Zum Abschluss des Wettbewerbs war von jedem der teilnehmenden Teams eine Präsentation vor den Konferenzteilnehmern über max. 15 Minuten zu halten. Dabei sollte sowohl die gewählte Vorgehensweise als auch die zehn am höchsten bewerteten Regeln vorgestellt werden.

Von einer Jury wurden die Vorgehensweise, die Ergebnisse selbst und die Präsentation bewertet. Zusätzlich gingen die Ergebnisse einer Publikumsumfrage in die Gesamtbewertung ein

Ergebnisse des Wettbewerbs

In den einzelnen Kästen stellen die jeweiligen Teams ihre Lösungen vor. Die Ergebnisse zu den einzelnen Aufgaben wurden anhand zweier Merkmale bewertet: Die quantitative Note ergab sich aus der Korrektheit der Ergebnisse, der Menge der gefundenen Regeln etc. Die qualitative Note ergab sich aus der Nützlichkeit der gefundenen Regeln, der jeweiligen Interpretation der einzelnen Teams und der Kreativität bei der Bearbeitung der Aufgaben. Das Team der WestLB erhielt den ersten Preis gefolgt vom Triple-A Team auf Platz zwei. Den dritten Platz teilten sich das QlikTech Team und das Team der Dresdner Bank. Die ersten drei genannten Teams stellen ihre Ergebnisse und Lösungswege in den folgenden Kästen vor.

Alle vier Teams beteiligten sich mit Verve und stellten zum Abschluss ihre Ergebnisse einem großen und interessierten Publikum vor. Nicht zuletzt brachten die Ergebnisse der Teilnehmer neue Erkenntnisse, die in die weitere Forschungsarbeit an der Humboldt-Universität eingehen werden. So sollen verstärkt Eigenschaften der in Konflikt stehenden Werte, wie z.B. deren Abstand, als eigenständige Attribute in die Suche nach systematischen Unterschieden einbezogen werden. Einige Eigenschaften können zur Klassifikation der Konflikte verwendet werden, wodurch zusätzliche Informationen generiert werden und eine stärkere Fokussierung der Regelsuche auf einzelne Konfliktklassen ermöglicht wird.

Im kommenden Jahr wird der Wettbewerb mit neuen Daten und neuen Aufgaben wiederum im Rahmen der GIQM Konferenz ausgerichtet. Interessenten können sich bei Michael Mielke melden (Michael.Mielke@bahn.de).