Automatische Mapping-Verarbeitung auf Webdaten

Andreas Thor <thor@informatik.uni-leipzig.de>

Das World Wide Web stellt eine Vielzahl von Informationen zur Verfügung. Neben dem Inhalt der Webseiten stellt ihre Verknüpfung durch Verweise zwischen ihnen eine wichtige Informationsquelle dar. So unterstützen z.B. Webseitenempfehlungen auf interessante andere Produkte Kunden großer E-Commerce-Websites bei der Suche nach dem passenden Produkt. Zusätzlich finden sich auf Seiten verschiedener Websites auch Informationen zum gleichen Objekt der realen Welt, so dass durch Links zwischen ihnen alle verfügbaren Informationen aus mehreren Datenquellen für entsprechende Analysen, z.B. einem Preisvergleich, erreichbar werden. In dieser Arbeit werden Verweise als sogenannte Mappings zusammengefasst, wobei ein Mapping eine Menge paarweiser Zuordnungen, sogenannter Korrespondenzen, zwischen Instanzdaten (Objekten) repräsentiert. Dabei wird mit dem Begriff Korrespondenz nicht nur die Verknüpfung gleicher Objekte gemeint, sondern allgemein eine semantische Beziehung zwischen zwei Objekten ausgedrückt. Innerhalb dieser Dissertation steht die Verarbeitung solcher Mappings, d.h. ihre Erzeugung, Optimierung und Verwendung, im Mittelpunkt.

Innerhalb einer Website können Korrespondenzen z.B. als Webseitenempfehlungen, sogenannte Recommendations, verwendet werden. Ausgehend von den vielen in der Literatur vorgeschlagenen Algorithmen zur Berechnung von Recommendations, sogenannte Recommender, stellt sich das Problem, welcher Recommender für welchen Nutzer unter welchen Umständen am nützlichsten ist. Durch eine gezielte und optimierte Auswahl sollen dabei die besten Recommendations bestimmt werden. Mit AWESOME [4, 5] wird dazu ein Ansatz vorgestellt, der eine solche adaptive Bestimmung von Recommendations zulässt. Durch die Aufzeichnung und Auswertung von Nutzer-Feedback können die Empfehlungen den Nutzerinteressen angepasst werden, so dass eine Steigerung der Recommendation-Qualität ermöglicht wird. Innerhalb der Arbeit werden insbesondere automatische Verfahren zur Verarbeitung des Nutzer-Feedbacks vorgestellt, so dass eine automatische Selbstoptimierung der Recommendations erzielt werden kann. Der AWESOME-Ansatz wurde innerhalb einer Website prototypisch implementiert und die Recommendation-Qualität bzgl. verschiedener Kriterien evaluiert. Dabei konnte insbesondere gezeigt werden, dass die automatische Adaption ähnliche Ergebnisse erzielt wie eine aufwändige, manuelle Optimierung der Recommendations.

Ein weiterer Schwerpunkt der Arbeit liegt in der Verarbeitung von Mappings zur Datenintegration. Dazu wird im zweiten Teil der Arbeit der Datenintegrationsansatz iFuice [2] präsentiert, der - im Gegensatz zu schemabasierten Datenintegrationsansätzen - auf instanzbasierten Mappings zwischen Datenquellen aufbaut. Datenquellen und Mappings werden dabei mit Hilfe eines Domänenmodells semantisch annotiert, das u.a. die Typen der Objektinstanzen sowie die Art der durch Mappings definierten Beziehungen charakterisiert. Im Gegensatz zu den meist anfragebasierten Datenintegrationsansätzen eröffnet iFuice dem Nutzer die Möglichkeit, mittels Skripten ausführbare Datenintegrationsprozesse zu definieren. Innerhalb der Skripte kommen Operatoren zum Einsatz, die Objektinstanzen und Mappings in einer generischen Art und Weise verarbeiten. Dadurch können sowohl neue Mappings generiert als auch bereits bestehende durch entsprechende Kombination effektiv wieder verwendet werden. Ein weiterer Aspekt von iFuice ist die Möglichkeit der Informationsfusion, bei der als gleich erkannte Objektinstanzen zu sogenannten aggregierten Objekten zusammengefasst werden können. Die Verwendung von iFuice zur Informationsfusion wird insbesondere am Beispiel einer Zitierungsanalyse wissenschaftlicher Publikationen detailliert vorgestellt [1, 3].

Abschließend wird im letzten Teil der Arbeit das MOMA-Framework präsentiert, das auf den wichtigen Aspekt des Object Matching, d.h. dem Erkennen von Abbildern der gleichen Objekte der realen Welt in (verschiedenen) Datenquellen, abstellt. Das MOMA-Framework [6] setzt auf dem iFuice-Ansatz auf und verwendet insbesondere dessen Operatoren und Datenstrukturen. MOMA unterstützt die Erstellung sogenannter Match-Workflows, die u.a. verschiedene Match-Verfahren, z.B. durch Berechnung syntaktischer Ähnlichkeiten von Attributwerten, integrieren können. Das Ergebnis eines Match-Workflows ist jeweils ein Mapping, das wiederum zur Informationsfusion innerhalb von iFuice genutzt werden kann. Wichtiges Kennzeichen des MOMA-Frameworks ist die Möglichkeit, existierende Mappings miteinander kombinieren und dadurch neue Mappings ableiten zu können. Mapping-Kombinationen können als sogenannte Match-Strategien definiert werden, welche flexibel für verschiedene Datenquellen eingesetzt werden können. Mit Hilfe einer prototypischen Implementation wurden verschiedene Match-Strategien unter Verwendung realer Datenquellen aus dem Bereich bibliografischer Daten evaluiert.

[1] Rahm, Thor: Citation Analysis of Database Publications. SIGMOD Record 34(4), 2005
[2] Rahm, Thor, et.al.: iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings. Proc. of WebDB, 2005
[3] Thor, Aumüller, Rahm: Data Integration Support for Mashups. Proc. of IIWeb, 2007
[4] Thor, Golovin, Rahm: Adaptive Website Recommendations with AWESOME. VLDB Journal 14(4), 2005
[5] Thor, Rahm: AWESOME - A Data Warehouse-based System for Adaptive Website Recommendations. Proc. of VLDB, 2004
[6] Thor, Rahm: MOMA - A Mapping-based Object Matching System. Proc. of CIDR, 2007