Zusammenfassung der Dissertation

Automatische Mapping-Verarbeitung auf Webdaten

vorgelegt von Diplom-InformatikerAndreas Thor

Neben dem Inhalt von Webseiten stellt die 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 verschiedenen Webseiten Informationen zum gleichen Realweltobjekt, so dass durch Verweise alle verfügbaren Informationen für entsprechende Analysen, z.B. einem Preisvergleich, erreichbar werden. In dieser Arbeit werden Verweise als sogenannte Mappings zusammengefasst, wobei ein Mapping semantische Beziehungen zwischen Instanzdaten (Objekten) mit Hilfe paarweiser Zuordnungen repräsentiert.

Innerhalb einer Website können Verweise z.B. als Webseitenempfehlungen, sogenannte Recommendations, verwendet werden. Die Vielzahl der Algorithmen zur Berechnung von Recommendations, sogenannte Recommender, führt zur Fragestellung, welcher Recommender für welchen Nutzer unter welchen Umständen am nützlichsten ist. Mit AWESOME wird dazu ein Ansatz vorgestellt, der eine gezielte und optimierte Auswahl der Recommender und damit eine 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 präsentiert, der auf instanzbasierten Mappings zwischen Datenquellen aufbaut. Datenquellen und Mappings werden dabei mit Hilfe eines Domänenmodells semantisch annotiert. iFuice eröffnet 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.

Abschließend wird im letzten Teil der Arbeit das MOMA-Framework präsentiert, das auf den wichtigen Aspekt des Object Matchings, d.h. dem Erkennen von Abbildern der gleichen Objekte der realen Welt in (verschiedenen) Datenquellen, abstellt. Das MOMA-Framework 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.