Hasso Plattner Institut
IT Systems Engineering
Universität Potsdam

FUZZY! Informatik AG

Bachelorprojekt für 2007/2008
Fachgebiet Informationssysteme – Prof. Dr. Felix Naumann

Datenfusion – Konsolidierung widersprüchlicher Daten

Hintergrund und Projektbeschreibung

In fast allen größeren Datenbeständen befinden sich so genannte Dubletten oder Duplikate, also mehrfache Repräsentationen des gleichen Realweltobjekts. Insbesondere im **Kundendatenmanagement** (CRM) haben Duplikate negative Auswirkungen sowohl für die Kunden als auch für die Unternehmen.

Die FUZZY! Informatik AG bietet Werkzeuge an, mit deren Hilfe solche **Duplikate** effizient gefunden werden können. Es bleibt hernach jedoch dem Unternehmen überlassen, wie systematisch mit solchen Duplikaten umgegangen werden soll. Jede der verschiedenen Repräsentationen des Kunden kann wertvolle aber auch **widersprüchliche Informationen** bergen. Deshalb ist das einfache Löschen aller bis auf einer Repräsentation nicht ratsam. Andererseits erlaubt die meist große Menge an Duplikaten keine manuelle Bearbeitung jedes einzelnen Falls.

Im Rahmen des Projekts soll eine Anwendung entwickelt werden, die Unternehmen bei der **Datenfusion** unterstützt. Die Anwendung sollte folgendes leisten

- Partitionierung von Duplikatgruppen in Gruppen vergleichbarer Konfliktsituationen
- Bereitstellen einer Vielzahl von teils selbst entworfenen Konfliktlösungsfunktionen und deren effizienter Implementierung
- Visuelle Darstellung von Duplikatgruppen und manuelle Konfliktlösung darin
- Automatische Auswahl von Konfliktlösungsfunktionen pro Attribut
- Effiziente Fusion großer Datenbestände
- Interaktion mit anderen FUZZY! Produkten, insbesondere FUZZY!Double

Rahmenbedingungen

Das Projekt wird vom Hasso Plattner Institut, vertreten durch das Fachgebiet Informationssysteme, in **Zusammenarbeit mit der FUZZY! Informatik AG** durchgeführt. Die FUZZY! Informatik AG kooperiert bereits seit Jahren erfolgreich mit der Arbeitsgruppe in Forschung und Lehre. Es werden beispielsweise regelmäßig gemeinsame Studenten-Workshops zur Duplikaterkennung durchgeführt und FUZZY! setzt von der Arbeitsgruppe entwickelte Data Profiling Methoden erfolgreich in Produkten ein.

Die Teilnehmerzahl in diesem Bachelorprojekt ist auf 8 Mitglieder beschränkt. Die technische Umsetzung erfolgt mit Java.

Projektvorbereitung

In der Vorbereitungsphase werden Grundlagen der **Informationsintegration** vorgestellt. Die Teilnehmer erlernen Techniken des Data Profiling, der Standardisierung und der Duplikaterkennung. Ein Schwerpunkt bilden Techniken der Datenfusion, der Umgang mit großen Datenmengen und zudem die Einarbeitung in die Produkt- und Entwicklungswelt von FUZZY!. Diese Themen werden im Rahmen regelmäßiger Treffen im Wintersemester 2007/08 durch die Teilnehmer erarbeitet und vorgestellt.

Projektbeginn: 15.10.2007

Kontakt

Für weiterführende Informationen steht Prof. Dr. Felix Naumann zur Verfügung. Eine Terminabsprache ist über das Sekretariat von Prof. Naumann möglich: office-naumann@hpi.uni-potsdam.de

http://www.hpi.uni-potsdam.de/naumann/
http://www.fazi.de/