Allegato 3

CONSIGLIO NAZIONALE DELLE RICERCHE

RELAZIONE PROGRAMMA SHORT - TERM MOBILITY - Anno 2006

Proponente: *Prof. Domenico Saccà, Professore Ordinario, Università della Calabria*

Fruitore: *Ing. Angela Bonifati, ricercatrice a tempo indeterminato Icar-CNR*

Istituto di afferenza: *Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) CNR*

Dipartimento: nr. 09 codice

Titolo del programma:
Obiettivi (3-4 righe)

Title: CIX: Cleaning and Integrating XML Databases.
Goals: The aim of the project is to investigate the need of performing data cleaning operations on XML databases, prior to data integration.
An important aspect of integration of different XML sources is indeed to identify the similarities among the sources elements in both the instances and the structure. However, this task may become awkward if the data formats are not uniform. This project aims at working out new solutions for such a
a problem.

Relazione Finale del Programma svolto dall’Ing. A. Bonifati presso University of Saarland (Germania) dal 8 Maggio 2006 al 28 Maggio 2006:

The project *CIX* (pronounced “kicks”, stands for “*Cleaning and Integrating XML databases*”) aims at addressing the problem of XML data cleaning tailored to “ease” the integration task. This problem is very relevant to Information and Communication Technology, as most of the rough data today is in XML thus exhibiting a lot of heterogeneity.
Errors and inconsistencies among XML elements belonging to different sources typically hinder the data integration task. Many past works proposed declarative languages [*Galhardas*] for data cleaning or techniques based on interaction with the user [*Raman*] or on update propagation [*Labrinidis*].
In this project, we plan to study ad-hoc data cleaning techniques for new data formats, such as XML[*XML*], which are driven by the need of integration.
During the program, me and Prof.Koch and his students have been discussing many issues, regarding the development of the project. For instance, one can assume to operate in a distributed setting with sets of possible worlds, each world being an XML document or its corresponding relational table. Thus, the aim is to delete the pairs of worlds across the two sets that will never match.

A key example is that of dates which may appear in a whole element (date element) in a source and may be split into several elements in another source (day, month, year elements). These elements also identify a 1-to-many complex mapping among themselves which, being concatenations of one to another, is not easy to recognize. A recent project [*iMAP*] focused on designing a set of heuristics to find such complex correspondences. However, the problem of identifying uncorrect data prior to the integration process is not dealt with.
[*Cli*o] is a data integration tool for XML data that considers a simplified nested relational model and mappings among the elements that are not compositions of the latter through functions, such as concatenations or more complex arithmetic functions.
To the best of our knowledge, there is no project that handles the data cleaning and integration processes together.

We investigated numerous open-problems that concerns the data cleaning process of heterogeneous XML data and further explored how this impacts the quality of data integration. An interesting research direction is to focus on the kinds of mappings one wants to identify (especially the most complex 1-to-many and many-to-many which have received so far poor attention) and to design the data cleaning process accordingly.

Finally, as a proof of concept, we started to implement a research prototype which validates our approach. In particular, we want to show the improved quality of the integration process that derives from a better quality of the input data.

[*Galhardas*] *Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon, and Cristian-Augustin Saita* “Declarative data cleaning: language, model, and algorithms”, In Proc. of VLDB Conference 2001.
[iMAP] R. Dhamankar, Y.Lee, A.Doan, A.Y.Halevy, P.Domingos, “iMAP: Discovering Complex Mappings between Database Schemas”. In Proc. of SIGMOD Conference 2004: 383-394
[Labrinidis] A.Labrinidis and N.Roussopoulos “Update propagation strategies for improving the quality of data on the Web”. In Proc. of VLDB Conference 2001.
[Raman] V. Raman and J. M. Hellerstein “Potter's wheel: an interactive data cleaning system”, In Proc. of VLDB Conference 2001.
[Clio]L. Popa, Y.Velegrakis, R.J.Miller,M.A.Hernandez,R.Fagin,
"Translating Web Data", In Proc. of VLDB Conference 2002
[XML] http://www.w3.org/xml/
[XQuery] http://www.w3.org/xmlquery/

Firma del Proponente (Prof. Domenico Saccà)

Firma del Fruitore (Ing. Angela Bonifati)