Data and Web Management Research at Politecnico di Milano

Stefano Ceri, Cristiana Bolchini, Daniele Braga, Marco Brambilla, Alessandro Campi, Sara Comai, Piero Fraternali, Pier Luca Lanzi, Marco Masseroli, Maristella Matera, Mauro Negri, Giuseppe Pelagatti, Giuseppe Pozzi, Elisa Quintarelli, Fabio A. Schreiber, Letizia Tanca

Dipartimento di Elettronica e Informazione (DEI), Politecnico di Milano
Piazza L. da Vinci, 32 – 20133 Milano, Italy
first_name.last_name@polimi.it

1. INTRODUCTION

Research in data management at Politecnico di Milano has a long and solid tradition; forefront books on distributed databases, conceptual database design, logical databases, and active databases contributed to shape the foundations of this discipline in the last two decades. Historically, our work has addressed both all aspects of innovation in the technology of modern data management systems and the consequent support of design methods and tools.

Recently, a large fraction of the group's efforts has been dedicated to the Web, considered as the ubiquitous infrastructure for making access to distributed and heterogeneous data sources. Our work for the Web concentrated on the models, methods, languages, and tools for supporting the design and automatic generation of modern, data-intensive Web applications.

Although classifying our work in data-driven vs Web–driven is rather arbitrary – and to some extent misleading, because in many projects we use both technologies – for ease of organization we will use this classification. Our report focuses on the last five years of activity.

2. DATA-DRIVEN RESEARCH

In recent years, work addressed the enhancement of database technology in several directions, including active, temporal, spatial, and mobile/very small databases; we also focused on query and mining languages for XML repositories and on supporting effective usage of genomic information. In these fields, emphasis was placed on formally defining the new features required by each specific data management extension, and then inferring properties descending from those definitions, which lead to improved system implementations or to better understanding of the system behaviour. While supporting the technological advances in data management is important, it is perhaps even more important to support developers and users in taking full advantages of the technology; therefore, database research in the group has always been characterized by an emphasis on innovative languages, methods, design support environments and tools, which could bridge technology to its use in real-life applications.

2.1 Active Databases

Research in active databases has been active since fifteen years, with very visible results worldwide. Recent work focused on defining the formal properties of active rules, enabling the development of more powerful rule analysis tools, and on defining strategies for improving the performances of rule sets under a characterization of the load due to passive and active computations. Active rule analysis is important for rule usability: thanks to rule analysis, rule interaction can be tested and properties such as termination and confluence can be proved. In [1], techniques descending from logic programming are proposed as a new paradigm for rule verification. In [2], the new property of event-trace independence is defined; it guarantees that rule executions are indistinguishable even when we consider an arbitrary sequence of their triggering events. In [3], the scheduling of detached rules is optimized under a characterization of the load due to passive and active computations.

2.2 Temporal Databases

Work in temporal databases focused on the introduction of process modelling aspects within temporal information management, and specifically on the impact of conceptual aspects (temporal constraints for process modelling exceptions that may occur during process execution) upon architectural

issues and choices [4][1]. Work [5] analyzes the potential applications of process modelling and temporal databases to medicine.

2.3 Context-Aware And Mobile Databases

Work on context-aware and mobile databases initially focused on data structures and access methods for improving the performance of data management on small, mobile devices. New efficient logical and physical data structures have been defined for DBMSs running on very small, portable, mobile devices [6]; performance, power consumption, and endurance parameters were optimized using an EEPROM Flash device as storage medium. Starting from the needs of the mobile scenario, we developed a comprehensive methodological framework for integrating, tailoring and delivering context-aware data [7]. The method, supported by a prototype tool – *Context-ADDICT,* demonstrated in [8] –, can be seamlessly applied to large information bases, in order to provide users and applications with the appropriate share of data, tailored to the current context [9]. The research on very small devices has recently drifted towards data management in embedded and pervasive systems, generating a research line on query languages for Wireless Sensors Networks [10] and an SQL-like language for pervasive system.

2.4 Spatial Data

Work on spatial data focused on solving the interoperability problems encountered in building a spatial data infrastructure (SDI), consistent with the INSPIRE directive, and in the development of an integrated interoperability architecture capable of dealing with the semantic mapping and geometric harmonization issues raised by the design of a strongly integrated SDI at regional level. The main achievement in this field has been the development of the GeoUML Spatial Conceptual Model [10-13] and its application in the development of a regional geographic database for "Regione Lombardia". The model has been adopted by the national committee for standards in geographic data (equivalent to FGDC in USA) at CNIPA (equivalent to NIST in USA), http://www.cnipa.gov.it/.

2.5 Query Language Design

The contributions to query language design were focused upon XML and how to make XML repositories more usable, both in terms of user interface and of retrieval success. We designed XQBE (XQuery By Example) [14], a visual query language using examples of XML as a paradigm for querying XML repositories. XQBE is inspired by QBE invented by Moshe Zloof at IBM Watson Research and available on many products (e.g. MS Access). XQBE allows one to formulate simple queries on top of XML repositories, by drawing annotated trees; the language is formally defined and tools map XQBE to XQuery and XPath. The XQBE environment is referenced from the W3C site linking to XML query language implementations, and has been internationally used as a pedagogical tool for learning XQuery. XQBE is also inspiring recent joint research with IBM Almaden Research Center for enhancing the Clio System, an XML mapping research prototype already partially used by commercial products, by adding object-oriented concepts to it [15]. We also addressed the characterization of graph-based queries (for XML and for temporal databases) by means of model-checking based techniques [16].

2.6 Data Mining

Work on data mining focused on new paradigms, algorithms and execution environments for extracting association rules and sequential patterns from XML repositories, thus enabling classical mining operations for a new and important class of repositories. The research on mining XML repositories has given rise to the development of a rich tool environment, named XMINE, supporting several data mining patterns. The main XMINE operator, described in [17], is based on XPath; it can express complex mining tasks, by indifferently (and simultaneously) targeting both the content and the structure of the data.

Another data mining approach consists in the recognition of frequent patterns within XML documents, and on the use of such patterns as summarized representations of the data; these patterns can then be stored and queried, either when fast (and approximate) answers are required, or when the actual dataset is not available, e.g. it is currently unreachable [20].

Additionally, we worked on extracting unexpected patterns (*pseudo-constraints*) from relational databases. This method reveals properties on database states not declared as constraints, but whose violation instances are interesting facts, hence it considers data mining from a new, fully original perspective [19]. Finally, a complete survey of Web Usage Mining is presented in [18], which surveys about 200 papers published in this area between 2001 and 2005.

[1] The paper [4] had the highest number of downloads from the ACM digital library for a period of 28 months.

2.7 Genomic Data Management

Work on genomic data management has produced methodologies and algorithms to effectively use and mine genomic information in heterogeneous and distributed genomic databases. The work also generated a Web-enabled system, named GFINDer: Genome Function INtegrated Discoverer (http://www.bioinformatics.polimi.it/GFINDer/) [21-23], allowing scientists to select and evaluate efficiently and dynamically the most relevant functional and phenotypic information supporting knowledge discoveries in different biomedical experiments. It is a system for discovering, using, and mining a large amount of genomic information and knowledge retrieved from many heterogeneous and distributed databases accessible via the Internet for supporting the evaluation and biomedical interpretation of high-throughput biomolecular experiments. It has been actively used by international Research Centers and Universities: at the time of writing GFINDer Web site received nearly 97,000 accesses by more than 5,500 distinct IPs since its opening in 2004. We also developed, in collaboration with the National Institute of Health, Bethesda (USA), a novel heuristic strategy to filter semantic relations extracted from the scientific literature by using natural language processing [24]. The method allows extracting the valuable genomic functional information with enough quality for subsequent applications aimed at uncovering new biomedical knowledge.

3. WEB-DRIVEN RESEARCH

The growth of Web applications as the fundamental infrastructure for business and social activities has generated a strong interest in methods, environments, and tools supporting their design and deployment. The continuous evolution of technologies calls for a foundational, technology-independent approach, rooted in the tradition of information modelling. The main focus of this research is centred upon a conceptual modelling language, called **WebML** (Web Modelling Language)[2]. WebML describes a conceptual model of the Web application in which the various aspects of the specification (respectively the content, the hypertext, and the presentation) are orthogonally combined. The most innovative aspect of WebML is the modelling of hypertexts as collections of elementary units and links, where the units describe both the visualization of elementary elements of a Web page and the operations performed by the application, and links between units capture the user behaviour. The model was initially focused on the display and management of contents, but it has been progressively extended to incorporate other features of modern Web applications, including: process management, Web service invocation and publication, management of adaptive and reactive computations; management of collaborative applications; methods for improving the accessibility, and more in general the quality of Web applications; support for new media, technologies and architectures, including rich Internet applications, VOIP, and Web architectures for embedded systems. The WebML model and design method are patented in US [25] and described by an international book [26].

A WebML specification is a graph, therefore WebML specifications are supported by a visual design tool with extensible components; such tool, called *WebRatio*, has been initially developed at Politecnico as result of EU-funded projects in the fourth and fifth framework, then has been the core of the spin-off company *Web Models*[3].

Recent work in WebML addresses Web services invocation and publication, process management, model-driven design and translation, support of semantic Web services, development of rich Internet and of embedded Web applications.

3.1 Service Invocation and Publication

Web service invocation and publication is explored in [27], where WebML is extended to model complex interactions with Web services. The concepts presented in this paper were fully implemented, as described in [28]. Industrial applications are described in the joint work between our group and the spin-off Web Models [29].

3.2 Process Management

Process management was addressed in [30], by extending the WebML modelling language to describe process-enabled Web applications, i.e. applications where the navigation of the user is

[2] **WebML** is extensively used in research and teaching by about 50 national and international institutions, including Technion Haifa (Israel), ETH Zurich (Switzerland), Katholieke Universiteit Leuven (Belgium), University of California San Diego (USA), TIFR (India).
WebML was also independently extended by other research groups, as demonstrated by several research papers published at WWW, ICWE, ECBS, SEKE.

[3] **Web Models** is a spin-off company participated by the Politecnico having now about 20 employees, subdivided in the two locations of Milano – for commercial development – and Como – the software factory (http://www.webratio.com/).

driven by workflow constraints. The process modelling phase drives the hypertext generation phase, by automatically generating (low-level) hypertext skeletons from (high-level) process models, according to different styles of process enactment. In the above context, special emphasis is given to reverse engineering, i.e. the ability to reconstruct the process from the generated hypertext when the latter is subject to modifications and evolution. This work represents one of the first Web engineering proposals for modelling data- and process-centric applications. Due to the loose control on Web clients, exceptions occurring during execution of processes on the Web are a hard problem. Papers [31,32] present a high-level model enabling case-based exception resolution.

3.3 Model-Driven Design

Model-driven design has been the common driver in many research efforts. We have concentrated on collaborative applications in [33], on context awareness in [34] (with an application to the e-learning context in [35]), and on application quality in [36]; quality depends on conceptual properties of the designed applications instead of interface-specific properties. The method is supported by a tool automating the evaluation process. In [37] we address a general method for designing Web applications which uses the new notion of "Web mart", an extension of the notion of data mart which is suited to Web applications.

3.4 Model Transformations

Model transformations have adapted WebML to the Model-Driven Architecture (MDA) context. WebML has been generalized using UML Meta-Object Facility (MOF), to make it consistent with OMG's MDA. Code generation techniques are generalized into an abstract model-transformation framework, capable of addressing such tasks as: the generation of metric models for evaluating different size measures of a project (e.g., for automatically producing the Function Point count from a conceptual application model); or the generation of models for driving the automatic testing of applications.

3.5 Support of Semantic Web Services

Support of Semantic Web Services (SWS) *is* fundamental for spreading SWS. *A* joint team with CEFRIEL[4] has produced SWEET, a WebML-based environment for designing applications of SWS that automatically derives a large portion of SWS annotations from their high level models [38], thereby reducing the efforts required by experts. The environment was experimented for developing Web Service Meta Object (WSMO) components (services and mediators), and has participated to the Semantic Web Challenge [39], designed by Prof. Charles Petrie (Stanford University), which took place at Stanford, Budva, and Athens in 2006, and at Innsbruck and Stanford in 2007. This research received an IBM Faculty Award in 2006.

3.6 RIA and Embedded Applications

WebML conceptual model and WebRatio code generation technology were extended along the direction of rich Internet applications (RIA) in order to transfer more application logic from the server to the client. In addition, Web architectures and applications were downscaled to embedded systems, adapting the conceptual modelling primitives, runtime architectures, and code generation techniques to the space and time constraints of embedded architectures. Embedded applications are envisioned in domotics, intelligent buildings, cultural heritage, and industrial automation.

4. CONCLUSIONS

In October 2007, we held a one-day workshop dedicated to new research directions. We gathered about 40 people, including professors, researchers, and students from DEI, CEFRIEL and Web Models. We agreed that emphasis in the future will be dedicated to five major themes: bio & nano technologies, mediation & mapping, social analytics for the Web, new user experiences, and stream reasoning. The last topic, which is not covered in this paper, addresses the massive computation of reasoning (e.g., logical rules) on streaming data, and will be covered within FP7 by a joint research unit DEI-CEFRIEL. We are discussing each theme within a Wiki, and we will be glad to grant read access to anyone interested to contribute.

[4] **CEFRIEL** is a Center for ICT excellence, set up in 1988 as a consortium whose components are Academia (represented by Politecnico di Milano, Università degli Studi di Milano and Università degli Studi di Milano – Bicocca), Enterprises (including some of the most important ICT companies operating in Italy), and Public Administration, represented by the Lombardy Region. http://www.cefriel.it/.

5. REFERENCES

[1] S. Comai and L. Tanca. Termination and confluence by rule prioritization. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):257-270, 2003.

[2] A. Bonifati, S. Ceri, and S. Paraboschi. Event trace independence of active behaviour. *Information Processing Letters*, 94(2):71-77, 2005.

[3] S. Ceri, S. Paraboschi, G. Serazzi, and C. Gennaro. Effective scheduling of detached rules in active databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):2-13, 2003.

[4] C. Combi and G. Pozzi. Architectures for a temporal workflow management system. In Proc. of the 2004 ACM Symposium on Applied Computing (SAC), Nicosia, Cyprus, March 14-17, 2004. ACM Press, New York, NY, 659-666.

[5] K.P. Adlassnig, C. Combi, A.K. Das, E.T. Keravnou, and G. Pozzi. Temporal representation and reasoning in medicine: Research directions and challenges. *Artificial Intelligence in Medicine*, 38(2):101-113, 2006.

[6] C. Bolchini, F. Salice, F.A. Schreiber, and L. Tanca. Logical and physical design issues for smart card databases. *ACM Transactions on Information Systems* (TOIS) ACM, 21(3):254-285, 2003.

[7] C. Bolchini, C. Curino, F.A. Schreiber, and L. Tanca. Context integration for mobile data tailoring. In Proc. IEEE Int. Conf. on Mobile Data Management (MDM), Nara (Japan), 2006, 5.1-5.8.

[8] C. Bolchini, C.A. Curino, E. Quintarelli, F.A. Schreiber, and L. Tanca. Context-ADDICT: a tool for context modeling and data tailoring. In Proc. IEEE Int. Conf. on Mobile Data Management (MDM), (demo paper), May 2007.

[9] C. Bolchini, C.A. Curino, G. Orsi, E. Quintarelli, R. Rossato, F.A. Schreiber, and L. Tanca. And what can context do for data? *Communications of the ACM*, (in press).

[10] Schreiber F.A. Automatic generation of sensor queries in a WSN for environmental monitoring. In B. Van de Walle, P. Burghardt, and C. Nieuwenhuis, eds. Proc. 4th Int. ISCRAM Conference, Delft, May 2007, 245-254.

[11] A. Belussi, M. Negri, and G. Pelagatti. Modelling spatial whole-part relationships using an ISO TC 211 conformant approach. *Information and Software Technology*, 48: 1095-1103, 2006.

[12] A. Belussi, M. Negri, and G. Pelagatti. An ISO TC 211 conformant approach to model spatial integrity constraints in the conceptual design of geographical databases. In Advances in conceptual modeling theory and practice. LNCS 4231, Springer, 2006, pp.100-109.

[13] A. Belussi, M.A. Brovelli, M. Negri, G. Pelagatti, and F. Sansò. Dealing with multiple accuracy levels in spatial databases with continuous update. 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal, 5-7 July 2006.

[14] D. Braga, A. Campi, and S. Ceri. XQBE (XQuery By Example): a visual interface to the standard XML query language. *ACM Transactions on Database Systems*, 30(2):398-443, 2005.

[15] A. Raffio, D. Braga, S. Ceri, P. Papotti, and M.A. Hernandez. Clip: a visual language for explicit schema mapping. In Proc. IEEE Int. Conf. on Data Engineering, 2008 (in press).

[16] E. Quintarelli. Model-checking based data retrieval: an application to semistructured and temporal data. LNCS 2917, Springer Verlag, 2004.

[17] D. Braga, A. Campi, S. Ceri, P.L. Lanzi, and M. Klemettinen. Discovering interesting information in XML data with association rules. ACM-SAC 2003, Melbourne, USA, March 2003, pp. 1163-1167.

[18] E. Baralis, P. Garza, E. Quintarelli, and L. Tanca. Answering XML queries by means of data summaries. *ACM Transaction on Information System*, 25(3), 2007.

[19] S. Ceri, F. Di Giunta, and P.L. Lanzi. Mining constraints violations. *ACM Transactions on Database Systems*, 32(1):6, 1-32, 2007.

[20] F.M. Facca and P.L. Lanzi. Mining interesting knowledge from Weblogs: a survey. *Data & Knowledge Engineering*, 53(3): 225-241, 2005.

[21] M. Masseroli, D. Martucci, and F. Pinciroli. GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Research*, 32(Web Server issue):W293-W300, 2004.

[22] M. Masseroli, O. Galati, M. Manzotti, K. Gibert, F. Pinciroli. Inherited disorder phenotypes: Controlled annotation and statistical analysis for knowledge mining from

gene lists. *BMC Bioinformatics*, 6(Suppl 4):S18, 1-8, 2005.

[23] M. Masseroli. Management and analysis of genomic functional and phenotypic controlled annotations to support biomedical investigation and practice. *IEEE Transaction on Information Technology in Biomedicine*, 11(4):376-385, 2007.

[24] M. Masseroli, H. Kilicoglu, F-M. Lang, and T.C. Rindflesch. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics*, 7(1):291, 1-12, 2006.

[25] S. Ceri and P. Fraternali. Model for the definition of World Wide Web sites and methods for their design and evaluation. Patent US 6,591,271, July 2003.

[26] S. Ceri, P. Fraternali, A. Bongio, M. Brambilla, S. Comai, and M. Matera. Designing data-intensive Web applications. Morgan-Kaufmann Series in Data Management Systems, J. Gray ed., Morgan-Kaufmann, 2003.

[27] I. Manolescu, M. Brambilla, S. Ceri, S. Comai, and P. Fraternali. Model-driven design and deployment of service-enabled Web applications. *ACM Transactions on Internet Technology*, 5(3):439-479, 2005.

[28] M. Brambilla, S. Ceri, S. Comai, and P. Fraternali. A CASE tool for modelling and automatically generating Web service-enabled applications. *Int. Journal of Web Engineering and Technology (IJWET)*, 2(4):354-372, 2006.

[29] M. Brambilla, S. Ceri, P. Fraternali, R. Acerbis, and A. Bongio. Model-driven design of service-enabled Web applications. In ACM SIGMOD Conf. 2005, 851-856.

[30] M. Brambilla, S. Ceri, P. Fraternali, and I. Manolescu. Process modelling in Web applications. *ACM Transactions on Software Engineering and Methodology*, 15(4):360-409, 2006.

[31] M. Brambilla and C. Tziviskou. Fundamentals of exception handling within workflow-based Web applications. *Journal of Web Engineering*, 4(1):38-56, 2005.

[32] M. Brambilla, S. Ceri, S. Comai, and C. Tzivisko. Exception handling in workflow-driven Web applications. In Proc. WWW 2005, May 10-14, 2005, Chiba, Japan, pp. 170-179.

[33] M. Matera, A. Maurino, S. Ceri, and P. Fraternali. Model-driven design of collaborative Web applications. *Software-Practice & Experience*, 33:701-732, 2003.

[34] S. Ceri, F. Daniel, and F. Facca. Modelling Web applications reacting to user behaviours. Special issue on Web dynamics. *Computer Networks*, 50(10):1533-1545, 2006.

[35] S. Ceri, P. Dolog, M. Matera, and W. Nejdl. Adding client-side adaptation to the conceptual design of e-learning Web applications. *Journal of Web Engineering*, 4(1):21-37, 2005.

[36] P. Fraternali, P.L. Lanzi, M. Matera, and A. Maurino. Model-driven Web usage analysis for the evaluation of Web application quality. *Journal of Web Engineering*, 3(2):124-152, 2004.

[37] S. Ceri, M. Matera, F. Rizzo, and V. Demaldé, Designing data-intensive Web applications for content accessibility using Web marts. *Communications of the ACM*, 50(4):55-61, 2007.

[38] M. Brambilla, I, Celino, S. Ceri, D. Cerizza, and E. Della Valle. Model-driven design and development of semantic Web service applications. *ACM Transactions on Internet Technology (TOIT)*, 8(1), 2008 (in press).

[39] M. Brambilla, I. Celino, S. Ceri, D. Cerizza, E. Della Valle, and F.M. Facca. A software engineering approach to design and development of semantic Web service applications. In Proc. of the 5th Int. Semantic Web Conf. (ISWC 2006), Athens, GA, USA, November 5-9, 2006, LNCS 4273, 172-186.