Data Management Research at the Knowledge and Database Systems Lab (NTU Athens)

Timos Sellis, Yannis Vassiliou
Computer Science Division
School of Electr. and Comp. Engineering
National Technical University of Athens
Athens, Greece
{timos,yv}@dblab.ece.ntua.gr

1. INTRODUCTION

The Knowledge and Database Systems Lab (KDBSL) of the Electrical and Computer Engineering Dept. in the National Technical University of Athens was founded in 1992 by Prof. Timos Sellis and Prof. Yannis Vassiliou. Its activities involve theoretical and applied research in the area of Databases and Information Systems. The lab employs three postdoc researchers (Dr Theodore Dalamagas, Dr Alkis Simitsis, Dr Yannis Stavrakas), several PhD students and many graduate students. It has been involved in many research projects supported by the EU, international institutions, Greek organizations, the Greek Government and industrial companies.

This report presents a brief description of the major research activities of KDBSL. Those activities involve (a) data stream management, (b) modelling and operational issues in databases and data warehouses and (c) Web and P2P data management. More details about the activities, as well as the full list of KDBSL publications, are available at the site of the lab: `http://www.dblab.ece.ntua.gr`.

2. DATA STREAM MANAGEMENT

Stream processing must keep up with the fluctuating arrival rate of high-volume data sets. Hence it is not expected that fast in-memory computation can be applied over the entire stream. We consider that approximate query processing over compact, precomputed data synopses is sometimes the only viable option. The reason is that users can often tolerate small imprecision in query results, as long as these results are quickly generated and accompanied with accuracy guarantees. Besides, window specifications can be used to limit the scope of processing items as a means of providing incremental response to continuous queries. Finally, our interest also involves the special case of trajectory streams generated by traces of objects moving in a multidimensional space.

2.1 Constructing Optimal Wavelet Synopses

The wavelet transform is a powerful tool for constructing concise synopses of massive multi-dimensional data sets and rapidly changing data streams, while at the same time providing fast, approximate answers with accuracy guarantees. This transformation results in a set of wavelet coefficients offering a multi-resolution view of the data. Yet it requires only linear computational time and space. Due to its hierarchical decomposition, only a handful of the produced coefficients suffice to achieve a good summary of the data. As a result, an effective synopsis can be obtained by keeping those important coefficients and explicitly setting others to zero.

The work in [1] provides efficient I/O methods to create and maintain wavelet synopses over large multidimensional data sets. Results and space-time trade-offs are also presented in case of multi-dimensional data streams, when one of the dimensions (e.g., the time dimension) is continuously growing. For the more general and demanding case of aggregating streaming data, sketching techniques (i.e., randomized linear projections) need to be employed. The work in [2] solves the problem of estimating the most significant coefficients on the fly, as updates are streaming at high rates. The suggested method relies on two key technical ideas: (i) work directly on the wavelet domain by translating stream items; and (ii) a novel stream synopsis, termed Group-Count Sketch, which can be employed over hierarchically grouped coefficients to enable quick identification of important coefficients using a binary-search-like technique.

Our ongoing research focuses on constructing optimal synopses for more meaningful error metrics to measure the synopsis approximation quality. Improving space utilization, e.g., by employing adaptive quantization of coefficient values, can lead to significantly more accurate wavelet synopses.

2.2 Window Semantics for Streams

Since the total size of a stream is unknown, *windows* have been suggested as a means of limiting the amount of data given for processing and, thus, providing real-time responses to *continuous queries*. In [3] we have developed a foundation with formal semantics for specifying windows over data streams, and proposed a careful taxonomy of the most significant variants (count-based and time-based sliding, partitioned, landmark, and tumbling windows).

On the basis of rigorous algebraic specifications, we attempt to integrate rich windowing constructs with principal relational operators (selection, projection, and join). Our aim is to gain expressiveness for a wide range of SPJ continuous queries over streams.

We have developed a prototype stream processing engine to verify the correctness of windowing semantics and to investigate perspectives of syntactic equivalences and query

optimization. Overall, we consider windows as first-class citizens in stream processing. Windows should be treated as an indispensable ingredient of a stream algebra and query language towards a stream-enabled SQL.

2.3 Managing Trajectory Streams

Positioning applications are becoming extremely popular, e.g., for sensor monitoring or location-based services, due to the recent advances in telecommunications and location-enabled facilities like GPS, RFIDs, PDAs etc. *Trajectory data* from tracing movement of people, vehicles, animals etc. is constantly collected from multiple sources. Then, data is transmitted to a central processor as continuous, time-varying and possibly unbounded data streams of timestamped locations. In [4] we introduce a model that captures (a) the multidimensional nature of trajectories, (b) the interaction between space and time, and (c) frequent positional updates and interrelationships of moving objects with stationary entities. In such a dynamic setting, we believe that a generalized notion of windowing operators that combine temporal and spatial properties would be appropriate. Such a notion can effectively capture the evolving characteristics of trajectories, especially with respect to *kinetic operations*.

Moreover, if approximate query answering is accepted as a trade-off to achieve real-time responses, then probably the amount of trajectory data should be reduced. In [5] we introduce two novel compression techniques based on *sampling* that take advantage of features pertinent to trajectories and yield reliable approximation quality at affordable computation overhead. The first technique is based on spatiotemporal thresholds that examine speed and orientation of moving objects. The second one attempts to preserve the shape of each trajectory, while maintaining a minimal sequence of point locations.

Moreover, we aim to exploit other synopsis techniques (like wavelets or sketches) for streaming trajectories, taking into account their multidimensional properties to efficiently support spatiotemporal continuous queries, such as range aggregates or similarity search.

3. MODELLING AND OPERATIONAL ISSUES IN DB'S AND DW'S

Research on modelling and operational issues of databases and data warehouses is another area of interest for our group. We study modelling problems such database schema evolution, pattern warehouses, the design and optimization of the internal processes of a data warehouse, and context-aware databases. We also study operational problems in databases. We propose a framework for a novel type of queries, precis queries, i.e. free-form queries that generate entire multi-relation databases. Such databases contain not only items directly related to the query selections but also items implicitly related to them. The query output may be transformed in narrative form providing much greater insight into the original data. Finally, our group also studies novel indexing schemes for containment queries.

3.1 Modelling & Optimization of ETL processes

Data warehouse operational processes normally compose a labor intensive workflow and constitute an integral part of the back-stage of data warehouse architectures. To deal with this workflow and to facilitate and manage the data warehouse operational processes, specialized processes are used under the general title *Extraction-Transformation-Loading (ETL) processes*. ETL processes are responsible for the extraction of data from several sources, their cleansing, their customization and transformation, and finally, their loading into a data warehouse.

We have studied the design, development and optimization of ETL processes [6]. The work in [7] proposes a novel *conceptual model* for the early stages of a data warehouse project. In those stages the time constraints of the project require a quick documentation of the involved data stores and their relationships rather than an in-depth description of a composite workflow.

Moreover, we have presented a formal *logical model* for the ETL environment. The model concentrates on the flow of data from the sources towards the data warehouse through the composition of transformations and data stores. It has two main characteristics: *genericity* and *customization* [8, 9]. The work in [9] also presents a palette of several templates to represent frequently used ETL activities along with their semantics and their interconnection. In [10] we build upon existing graph-based modelling techniques that treat ETL workflows as graphs [11]. We (a) extend the activity semantics to incorporate negation, aggregation and self-joins, (b) complement querying semantics with insertions, deletions and updates, and (c) transform the graph to allow zoom-in/out at multiple levels of abstraction. In [12] we exploit our modelling framework to introduce rigorous techniques to measure ETL workflows, and formally prove their applicability and correctness. We have complemented these two models with a method for the semi-automatic transition from the conceptual to the logical model for ETL processes [13].

Additionally, we have delved into the *logical optimization* of ETL processes to find the optimal ETL workflow [14]. We have modelled this problem as a state-space search problem. Each ETL workflow is considered as a state and the state space is fabricated through a set of correct state transitions [15]. Moreover, we provide algorithms towards the minimization of the execution cost of an ETL workflow.

Finally, we have implemented ARKTOS II, a prototype ETL tool to facilitate the design, the (re-)use, and the optimization of ETL workflows [16].

Our ongoing work involves the physical optimization of ETL workflows taking into account physical operators and access methods. Another challenge is to apply our findings in more general workflows (not just ETL ones). Future plans for ARKTOS II involve the extension of data sources to more sophisticated data formats, other that relational domain, like object-oriented or XML data.

3.2 Database Schema Evolution

Current database systems are continuously evolving environments, where design constructs are added, removed or updated rather often. Even trivial changes (e.g., addition or deletion of an attribute, modification of a constraint) in the schema of the database greatly affect a wide range of applications and queries built around it, which may become syntactically or semantically invalid. The problem of database schema evolution has been addressed under several contexts, e.g., OODBMS, RDBMS, XML. Nevertheless, problems persist, mainly due to the fact that in most cases, the proper

attention is not given to the role of queries as integral parts of the environment and therefore queries are not designed to handle database evolution. Our research introduces techniques to adapt queries and views to schema changes in the underlying database, in such way that the human interaction is minimized.

Specifically, our approach provides a mechanism to perform what-if analysis for potential changes of database configurations [17]. A graph model that uniformly models queries, views, relations and their significant properties (e.g., conditions) is introduced. Apart from the simple task of capturing the semantics of a database system, the graph model allows us to predict the impact of a change over the system. Furthermore, we provide a framework to annotate database constructs (including queries and views) with policies concerning their behavior in the presence of hypothetical changes occurring in the database schema. Rules that dictate the proper adaptation actions to *additions* or *deletions* performed to *relations*, *attributes* and *conditions* (all treated as first-class citizens of the model) are provided. Finally, a tool to visualize and performe what-if analysis for several evolution scenarios is being implemented [18].

3.3 Logical Subsets of Databases

The wide spread usage of database systems nowadays has brought the need to provide inexperienced users with the ability to easily search a database with no specific knowledge of a query language. Several recent research efforts have focused on supporting keyword-based searches over relational databases.

We have presented an alternative proposal and we have introduced the idea of *précis queries* [19]. These are free-form queries whose answer (a précis) is a synthesis of results, containing not only items directly related to the given query selections but also items implicitly related to them in various ways with the purpose of providing to the user much greater insight into the original data. Précis queries include two additional novelties: they do not generate individual relations but entire multi-relation databases, and query results are customized to a user or group of users and/or domain requirements [20].

Currently, we are working on the development of a précis query answering prototype, named PRÉCIS, with the following characteristics: (a) support of a keyword-based search interface for accessing the contents of the underlying database, (b) generation of a logical subset of the database that answers the query, which contains not only items directly related to the query selections but also items implicitly related to them in various ways, (c) customization of the logical subset generated and hence the précis returned according to the needs and preferences of the user as an individual or as a member of a group of users, and (d) translation of the structured output of a précis query into a narrative synthesis of results. This output is an English presentation of short factual information précis [21].

3.4 Indexing of Set-Valued Attributes

Containment queries on set-values emerge in a variety of application areas ranging from scientific databases to XML documents. Examples of set valued data can be found in market basket analysis, production models, image and molecular databases. Moreover, as RDBMS's and IR come closer (often in the interest of storing and handling XML and Web data), containment queries on set values become a use case of high significance for an RDBMS.

Despite the fact that set values are supported by the object relational model and by the most popular commercial RDBMS's, indexing techniques for set values and containment queries are limited. The basic alternatives for indexing set values still remain the inverted file and signature based methods, with the former offering superior performance in most cases.

We work on the design and implementation of novel indexing schemes that do not suffer from the drawbacks of the inverted file, and outperform it for containment queries [22]. Moreover, we investigate the evaluation of more complex queries on set values, involving ranking and similarity. We focus on proposing indices to support such queries, giving efficient IR functionality in an RDBMS.

3.5 Pattern Management

The vast volumes of information produced nowadays, makes the handling and storing of data increasingly harder. Still, the huge volumes of data do not constitute knowledge per se, but special extraction methods have to be employed in order to produce knowledge artefacts. We call these artefacts *patterns*. Till now, patterns had not been treated as first class citizens in most data models. The challenge from the database perspective is to develop models and tools for storing, querying and linking patterns with the underlying data.

Patterns are usually quite heterogeneous and require adhoc processing techniques. So far little emphasis has been given on developing an overall integrated environment for uniformly representing and querying different types of patterns. In recent research work [23] we have proposed such a framework: the *Pattern Warehouse*. The conceptual and logical model for the Pattern warehouse [24] defines an architecture of three modules: the *Pattern-Base*, where the pattern entities reside, the *Database* that is populated with raw data and the *Intermediate Mappings*, which is an indexing mechanism for tracing the connection between the patterns and the underlying data. Our late research interests focus on designing an efficient solution for the last module.

3.6 Context-Aware Databases

Information today is accessed and used in a global environment, where assumptions about data become less evident. Users with different backgrounds or viewpoints may interpret the same data in a different way. Moreover, the interpretation and suitability of data may depend on unpredictably changing conditions, like for example the current position of the user or the media he is using (laptop, mobile, PDA). To avoid such ambiguous situations the information provider needs to specify the context under which information becomes relevant. Conversely, information users can specify their own current context when requesting for data in order to denote the part that is relevant to their specific situation.

In our view the management of context should take place at the level of database systems in a uniform way, and context should be treated as a first-class citizen in data models and query languages. By incorporating context in semistructured data [25] and XML, we have shown that it is possible to have: a) management of data according to the interpretation frame, b) ability to pose *cross-world* queries that combine

information that hold under different contexts [26], c) personalization in a flexible and uniform manner and d) direct support for managing data and schema histories [27, 28].

Currently, we study the problem of incorporating context in the relational database model, which offers a strong formal background. We proposed the *ContextRelational* model (CR model) [29] that can be viewed as an extended relational model able to accommodate information entities that manifest different *facets* whose contents can vary in structure and value. Each facet of such a multi-faceted information entity is associated with a context, stating the conditions under which this facet holds. The operations of the CR model are context-aware extensions of the corresponding relational operations. We have developed a prototype implementation of the CR Model that demonstrates its basic principles.

4. WEB AND P2P DATA MANAGEMENT

We study models and techniques for managing Web data, and we explore alternative ways of interaction between databases and Web pages. Specifically, we work on querying methods for tree-structured data management based on partially specified queries. Also, we design effective architectures for P2P databases, and develop efficient query indexing and routing techniques for P2P spatial data. Finally, we design and implement new proxy architectures for dynamic Web pages.

4.1 Tree-structured Data Management

Huge volumes of Web data are organized in tree-structured form. Even if data is not stored natively in tree structures, export mechanisms make data publicly available in that form to enable its automatic processing by programs, scripts, and agents on the Web. XML data is by far the most prominent example. The recent proliferation of XML-based standards and technologies for managing data on the Web demonstrates the need for effective and efficient management of tree-structured data.

Querying tree-structured data sources usually requires resolving structural differences. Those differences appear because of the different possible ways of organizing the same data in tree-structures. Current tree-structured data query languages handle this issue in a procedural way. In this sense, the user should explicitly specify structural differences as part of the query itself. In fact, we identify a more general difficulty stemming from the fact that query formulation in current tree-structured data query approaches is strictly dependent on the structure of the data. Users cannot easily form queries without explicitly specifying the structural constraints.

Our ongoing research focuses on query languages that allow for the partial specification of the structure on which they are applied. Also, we study how syntactic and semantic information in tree-structured data can assist query evaluation.

In [30, 31, 32], we studied *dimension graphs*, which are semantically rich constructs that assist query formulation and evaluation on tree structured data. A key feature in our approach is that queries are not restricted by the structure of the trees. We applied our approach on querying effectively and efficiently multiple trees in the presence of structural differences and inconsistencies.

We further elaborated on the issue of *partially specified queries*, and we designed a full-fledged query language with *partially specified tree patterns* [33]. In this language, users can flexibly specify the structure fully, partially, or not at all in the queries. Since a query language needs to be complemented with query processing and optimization techniques, we have already started working on query containment and minimization issues for partially specified tree structured queries [34].

4.2 Peer-to-peer Database Systems

In the last few years, there has been a growing interest in the Peer-to-Peer (P2P) paradigm. In contrast to data integration architectures, P2P data sharing systems do not assume a mediated schema to which all sources of the system should conform. Sources store and manage their data locally, revealing only part of their schemas to the rest of the peers. Due to the lack of global schema, they express and answer queries based on their local schema. Peers also perform local coordination with their their one-hop neighbors, creating mappings semi-automatically.

In unstructured P2P systems, peers that join the network get their one-hop neighbors randomly, without taking into account that there might be available peers that better meet their need for information. Therefore, they have to direct queries not only to their neighbors, but to a greater part of the system. Query processing in unstructured P2P systems involves the propagation of the query on paths of bounded depth in the network. At each routing step, the query is rewritten to the schema of its new host based on the respective acquaintance mappings. A query may have to be rewritten several times from peer to peer till it reaches peers that are able to answer it sufficiently in terms of quality but also quantity. However, the successive rewritings decrease or restrict the information that can be returned by a query and, thus, they reduce the possibility of accurate query answering. It is the case that peers may not be able to sufficiently answer received queries not because their local schema does not match the initial query adequately, but because the incoming rewritten version is too reduced compared to the initial query.

To deal with the above problems, we have studied and implemented techniques to provide peers that share relational data in unstructured P2P networks with accurate answers to locally posed queries without any global schema information [35]. Our approach employs a learning feature to gradually cluster nodes with semantically similar local schemas. This is performed by utilizing only regular query traffic. Based on learning results, mappings between remote peers are gradually built on their specific common interests.

Moreover, we are interested in P2P systems hosting multi-dimensional data. Until recently, research has focused mostly on P2P systems that host one-dimensional data (i.e. strings, numbers, etc). However, the need for P2P applications with multi-dimensional data is emerging. Yet, most existing indexing and search techniques are not suitable for such applications. Most indices for multi-dimensional data have been developed for centralized environments, while, at the same time, existing distributed indices for P2P networks aim at one-dimensional data. Our focus is on structured P2P systems that share spatial information. We suggest [36], a totally decentralized indexing and routing technique that is suitable for spatial data. Our technique can be used in P2P applications in which spatial information of various sizes can be dynamically inserted or deleted, and peers can join or

leave. The proposed technique preserves well locality, and supports efficient routing especially for popular and close areas.

We also investigate P2P and mobile agent architectures based on active database technology [37, 38]. We argue that employing ECA rules both for answering queries and deploying agents leads to an efficient query processing technique. Furthermore, the proposed mobile agent system architecture offers a number of advantages due to the performance and scalability that can be achieved using Active Databases.

As a new research direction, we study P2P architectures to support loosely coupled database communities. We have designed a flexible wrapping mechanism, based on RDFS schemas, for data sources that employ diverse local schema information. We use schema operators for such wrapping [39]. The operators are applied on RDF schema graphs available for the community, and produce new, integrated ones. Such integration is based on set-like semantics and gives an intuitive way in wrapping data sources. We have also implemented a query processing technique that does not require the existence of mapping rules during the propagation of the query in the sources. Under this technique, we are also able to retrieve answers, even in the case a query does not exactly match the schema of a local data source. Our ideas are implemented in SDQNET, a platform that supports semantic query processing in loosely coupled data sources [40].

4.3 Web Caching

Proxy caching is a commonly accepted methodology used to reduce Internet traffic, decrease back-end related user delays and generally improve Web performance. Numerous Web caching approaches have been proposed concerning static Web pages. Given that the usage of dynamically generated Web pages increases continuously, there is need for Web caching algorithms for dynamic Web pages as well.

Front-end caching approaches implement caching outside the site infrastructure, e.g. a proxy or a cache that resides at the edge of a Content Delivery Network). However, front-end caching is inadequate when it comes to caching dynamic Web pages, due to their low degree of reusability and strong dependency on the back-end site infrastructure. This is because the request of a dynamic Web page depends on client-defined input parameters. Serving such a request using a cache could be possible only if a cached dynamic Web page had been produced from the same application with the same input parameters, a rather rare situation. Even in that case, the creation of dynamic Web pages depends on client-related information (e.g. through the use of cookies).

In order to overcome these obstacles, we have studied alternative front-end Web caching approaches. In particular, we have designed and implemented a new proxy architecture [41] for dynamic Web pages. In our approach, caching is performed on the generation process of the dynamic Web pages and not the pages themselves. Our current efforts focus on replacement policies based on group-oriented caching and pre-fetching algorithms that improve the performance of a Web caching system.

5. ACKNOWLEDGMENTS

The research achievements at KDBS Lab is the result of the work of many present and past members of the group. We would like to mention here all past and current PhD students who have greatly contributed into making KDBSL a great place to be: (past PhDs) D. Arkoumanis, T. Dalamagas, N. Karayannidis, M. Koubarakis, S. Ligoudistianos, A. Maniatis, D. Papadias, A. Simitsis, S. Skiadopoulos, Y. Stavrakas, E. Stefanakis, Y. Theodoridis, A. Tsois, P. Vassiliadis - (current PhD students) G. Adamopoulos, S. Athanasiou, P. Bouros, A. Dimitriou, K. Gavardinas, P. Georgantas, V. Kantere, Y. Kouvaras, G. Papastefanatos, K. Patroumpas, I. Roussos, D. Sacharidis, D. Skoutas, S. Souldatos, L. Stamatogiannakis, M. Terrovitis, M. Veliskakis, P. Xeros.

6. REFERENCES

[1] M. Jahangiri, D. Sacharidis, and C. Shahabi. Shift-Split: I/O Efficient Maintenance of Wavelet-transformed Multidimensional Data. In *Proceedings of the ACM SIGMOD'05 International Conference*, Baltimore, Maryland, Jun 14-16, 2005.

[2] G. Cormode, M. Garofalakis, and D. Sacharidis. Fast Approximate Wavelet Tracking on Streams. In *Proceedings of the EDBT'06 International Conference*, Munich, Germany, Mar 26-30, 2006.

[3] K. Patroumpas, T. Sellis. Window Specification over Data Streams. In *Proceedings of the ICSNW'06 International Conference*, Munich, Germany, Mar 30, 2006.

[4] K. Patroumpas, T. Sellis. Managing Trajectories of Moving Objects as Data Streams. In *Proceedings of the STDBM'04 Workshop*, Toronto, Canada, Aug 30, 2004.

[5] M. Potamias, K. Patroumpas, T. Sellis. Sampling Trajectory Streams with Spatiotemporal Criteria. In *Proceedings of the SSDBM'06 International Conference*, Vienna, Austria, Jul 3-5, 2006.

[6] A. Simitsis. Modeling and Optimization of Extraction-Transformation-Loading (ETL) Processes in Data Warehouse Environments. *PhD Thesis*, Athens, Greece, 2004.

[7] P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Conceptual Modeling for ETL Processes. In *Proceedings of the DOLAP'02 International Workshop*, McLean, USA, Nov 8, 2002.

[8] P. Vassiliadis, A. Simitsis, S. Skiadopoulos. On the Logical Modeling of ETL Processes. In *Proceedings of the CAiSE'02 International Conference*, Toronto, Canada, May 27 - 31, 2002.

[9] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis. A Framework for the Design of ETL Scenarios. In *Proceedings of the CAiSE'03 International Conference*, Velden, Austria, Jun 16-20, 2003.

[10] A. Simitsis, P. Vassiliadis, M. Terrovitis, S. Skiadopoulos. Graph-Based Modeling of ETL Activities with Multi-Level Transformations and Updates. In *Proceedings of the DaWaK'05 International Conference* , Copenhagen, Denmark, Aug 22-26, 2005.

[11] P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Modeling ETL Activities as Graphs. In *Proceedings of the DMDW'02 International Workdhop*, Toronto, Canada, May 27, 2002.

[12] P. Vassiliadis, A. Simitsis, M. Terrovitis, S. Skiadopoulos. Blueprints and Measures for ETL

Workflows. In *Proceedings of the ER'05 International Conference*, Klagenfurt, Austria, Oct 24-28, 2005.

[13] A. Simitsis. Mapping Conceptual to Logical Models for ETL Processes. In *Proceedings of the DOLAP'05 International Workshop*, Bremen, Germany, Nov 4-5, 2005.

[14] A. Simitsis, P. Vassiliadis, T. Sellis. Optimizing ETL Processes in Data Warehouse Environments. In *Proceedings of the ICDE'05 International Conference*, Tokyo, Japan, Apr 5-8, 2005.

[15] A. Simitsis, P. Vassiliadis, T. Sellis. State-Space Optimization of ETL Workflows. In *IEEE TKDE*, 17(10), 2005.

[16] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, S. Skiadopoulos. A Generic and Customizable Framework for the Design of ETL Scenarios. In *Information Systems*, 30(7), 2005.

[17] G. Papastefanatos, P. Vassiliadis, Y. Vassiliou. Adaptive Query Formulation to Handle Database Evolution. In *Proceedings of the CAiSE'06 Forum*, Luxembourg, Grand-Duchy of Luxembourg, Jun 9, 2006.

[18] G. Papastefanatos, K. Kyzirakos, P. Vassiliadis, Y. Vassiliou. Hecataeus: A Framework for Representing SQL Constructs as Graphs. In *Proceedings of the EMMSAD'05 International Workshop*, Porto, Portugal, Jun 13-17, 2005.

[19] G. Koutrika, A. Simitsis, Y. Ioannidis. Précis: The Essence of a Query Answer. In *Proceedings of the ICDE'06 International Conference*, Atlanta, USA, Apr 3-7, 2006.

[20] A. Simitsis, G. Koutrika. Pattern-Based Query Answering. In *Proceedings of the PaRMa'06 International Workshop*, Munich, Germany, Mar 30, 2006.

[21] A. Simitsis, G. Koutrika. Comprehensible Answers to Précis Queries. In *Proceedings of the CAiSE'06 International Conference*, Luxembourg, Grand-Duchy of Luxembourg, Jun 5-9, 2006.

[22] M. Terrovitis, S. Passas, P. Vassiliadis, T. Sellis. A combination of trie-trees and inverted files for the indexing of set-valued attributes, 2006. *Submitted for publication.*

[23] S. Rizzi, E. Bertino, B. Catania, M. Golfarelli, M. Halkidi, M. Terrovitis, P. Vassiliadis, M. Vazirgiannis, E. Vrachnos. Towards a logical model for patterns. In *Proceedings of the ER'03 International Conference*, Chicago, USA, Oct 13-16, 2003.

[24] M. Terrovitis, P. Vassiliadis, S. Skiadopoulos, E. Bertino, B. Catania, A. Maddalena. Modeling and Language Support for the Management of Pattern-Bases. In *Proceedings of the SSDMB'04 International Conference*, Santorini Island, Greece, Jun 21-23, 2004.

[25] Y. Stavrakas, M. Gergatsoulis. Multidimensional Semistructured Data: Representing Context-Dependent Information on the Web. In *Proceedings of the CAiSE'02 International Conference*, Toronto, Canada, May 27-31, 2002.

[26] Y. Stavrakas, K., A. Efandis, T. Sellis. Implementing a Query Language for Context-dependent Semistructured Data. In *Proceedings of ADBIS'04 Conference*, Budapest, Hungary, Sep 22-25, 2004.

[27] Y. Stavrakas, M. Gergatsoulis, C. Doulkeridis, V. Zafeiris. Representing and Querying Histories of Semistructured Databases Using Multidimensional OEM. In *Information Systems*, 29(6), 2004.

[28] M. Gergatsoulis, Y. Stavrakas. Representing Changes in XML Documents Using Dimensions. In *Proceedings of the XSym'03 International Symposium*, Berlin, Germany, Sep 8, 2003.

[29] Y. Roussos, Y. Stavrakas, V. Pavlaki. Towards a Context-Aware Relational Model. In *Proceedings of the CRR'05 Workshop*, Paris, France, Jul 5-8, 2005.

[30] T. Dalamagas, D. Theodoratos, A. Koufopoulos, and V. Oria. Evaluation of Queries on Tree-Structured Data using Dimension Graphs. In *Proceedings of the IDEAS'05 Symposium*, Montreal, Canada, Jul 25-27, 2005.

[31] T. Dalamagas, D. Theodoratos, A.Koufopoulos, and I-Ting Liu. Semantic Integration of Tree-structured Data Using Dimension Graphs. In *Journal on Data Semantics*, IV, 2005.

[32] D. Theodoratos, and T. Dalamagas. Querying Tree-Structured Data using Dimension Graphs. In *Proceedings of the CAiSE'05 International Conference*, Porto, Portugal, Jun 13-17, 2005.

[33] D. Theodoratos, T. Dalamagas, A. Koufopoulos, and N. Gehani. Semantic Querying of Tree-Structured Data Sources Using Partially Specified Tree Patterns. In *Proceedings of the CIKM'05 Conference*, Bremen, Germany, Oct 31 - Nov 5, 2005.

[34] D. Theodoratos, T. Dalamagas, P. Placek, S. Souldatos and T. Sellis. Containment of Partially Specified Tree-Pattern Queries. In *Proceedings of the SSDBM'06 International Conference*, Vienna, Austria, Jul 3-5, 2006.

[35] V. Kantere, D. Tsoumakos, T. Sellis, N. Roussopoulos. GrouPeer: Dynamic Clustering of P2P Databases, 2006. *Submitted for publication.*

[36] V. Kantere, T. Sellis. Handling Spatial Data in P2P Systems, 2006. *Submitted for publication.*

[37] V. Kantere, A. Tsois. Using ECA Rules to Implement Mobile Query Agents for Fast-Evolving P2P Networks. In *Proceedings of the AAMAS'04 International Conference*, New York, USA, Aug 19-23, 2004.

[38] V. Kantere, I. Kiringa, J. Mylopoulos, A. Kementsientidis, M. Arenas. Coordinating P2P Databases Using ECA Rules. In *Proceedings of the DBISP2P'03 International Workshop*, Berlin, Germany, Sep 7-8, 2003.

[39] Z. Kaoudi, T. Dalamagas, T. Sellis. RDFSculpt: Managing RDF Schemas under Set-like Semantics. In *Proceedings of the ESWC'05 Conference*, Heraklion, Greece, May 29 - Jun 1, 2005.

[40] I. Spyropoulou, T. Dalamagas, T. Sellis. SDQNET: Semantic Distributed Querying in Loosely Coupled Data Sources, 2006. *Submitted for publication.*

[41] M. Veliskakis, Y. Roussos, P. Georgantas, T. Sellis. DOMProxy: Enabling Dynamic-Content Front-end Web Caching. In *Proceedings of the WCW'05 International Workshop*, Sofia Antipolis, France, Sep 12-13, 2005.