dbTrento: The Data and Information Management group at the University of Trento

Themis Palpanas
University of Trento, Italy
themis@disi.unitn.eu

Yannis Velegrakis
University of Trento, Italy
velgias@disi.unitn.eu

1. INTRODUCTION

The dbTrento group was established in 2006 by Profs Themis Palpanas and Yannis Velegrakis. Since then it has steadily grown into a fully functioning group with (currently) 17 members. It is located in Trento, a beautifully preserved historic town in the Dolomite mountains, which hosts one of the 6 ICT Labs of the European Institute of Innovation and Technology (EIT), and aims to become a reference research and technological center in Europe. The VLDB 2013 conference is being organized by dbTrento, its founders serving as the General Chairs. The mission of the group is to conduct high quality research on different aspects of large scale data and information management. The following sections provide a high level description of the work in these areas, which has also led to 3 Best Paper awards.

2. ADVANCED QUERY ANSWERING

[Semantic-based Keyword Search] Keyword search is becoming the de-facto mechanism for querying data [19], since it does not require knowledge of the full semantics or their organization in the repository, neither knowledge of some complex query language. For this reason, there has been enough work on querying structured (mostly relational) data through keywords. These works are typically based on an index that is built in advance, and which supports at run time the mapping of keywords to database structures. This index requirement makes these approaches difficult to apply when there is no prior access to the data, a common situation that occurs in integration system or on the web where the sources are autonomous and allow access to their data only through web interfaces or wrappers.

Collaboration with researchers from the University of Modena and the University of Zaragoza, has resulted into Keymantic [7], an engine for answering keyword queries over relational data that uses only the metadata provided by the database and some auxiliary information that is freely available on the web. Keymantic is using this information in order to understand the semantics of the keywords [6] and discover the best matching of these keywords to database tables, attributes or values. It does so by using an adapted version of the Hungarian algorithm. The discovered matches are combined to form SQL queries and returned to the user ranked in decreasing older of the likelihood that they represent the intended user query semantics. The KEYRY [9] is a version of Keymantics that uses a Hidden Markov Model, instead of the Hungarian algorithm, to make the predictions [8]. Apart from query answering, Keymantic can be used in data exploration. In particular, given a keyword query it can return structured queries exposing the structures of the data repository that may be related to the keyword query semantics. This can be applied not only on relational data but also on data of graph structure [10].

A lot of work within the group has been devoted to the management of entities that form the basis of Dataspaces and of Semantic Web Data [19]. One of these works is the discovery of the entities that best match a specification expressed in a keyword query. By exploiting attribute frequencies found in query logs, a classifier is trained to predict the intended semantics of the keyword query and to construct the answer as a ranked set of entities, with the most prominent at higher positions [35].

[Approximate Query Answering] Documents, or semi-structured data in general, provide a great deal of information, but their lack of schema makes information discovery a challenging tasks. Together with the University of Bozen-Bolzano and the University of Alberta, we have created TASM [3], a system that allows approximate query answering on large XML documents. It is based on the prefix ring buffer that allows the pruning of all the subtrees in the document above a threshold in a single postorder scan of the document, leading into an algorithm that depends only on the size of the query [4]. This work won the Best Paper award in ICDE 2010.

[Managing Evolution] The dynamic nature of the data has been realized from the very first days of data management. However, work has mainly focused on value

updates, offering the ability to query the data at different points of time. These approaches did not support the representation of relationships between different structures that model the same real world object at different points in time, which in turn made hard the realization of the evolution phases through which the object has gone through. Furthermore, even if such a modeling is made, it is not guaranteed that it will match the evolution model on which a query is constructed. For instance, a database may contain different instances of Germany from different phases of its history, i.e., as an empire, as the pre-war country, as East and West, and as a unified country. If these are different entities, a query asking for leaders of Germany, where Germany refers to the general concept of the German nation, will not be possible to answer.

With this in mind, our members have developed a mechanism that allows the evolution information to be included explicitly or implicitly [47, 45, 46, 42] in the data, forming an evolution graph. Apart from modeling and querying evolution [44], the system allows the answering of queries formed with a different evolution granularity in mind than the one of the repository [15]. This is achieved by performing at runtime merges of structures representing different evolution phases of an object. Achieving the minimum required such merges, boils down to efficient discovery of Steiner Forests which we solve using dynamic programming [14].

3. INFORMATION INTEGRATION

A long chapter in the dbTrento research agenda is related to Data and Information Integration from multiple disparate heterogeneous sources. Much of this work is based on entities as the basic data unit, driven by their recent popularity and expressiveness. Many of our developments were fundamental components in the OKKAM project that aims at the creation of an infrastructure for global identifiers for every web object [12, 5, 33, 27, 39] Our collaborations with colleagues from the Semantic Web community and our participation in many projects from that area has revealed numerous challenging data management issues, that were recently summarized in a related tutorial [24].

[Blocking Techniques] A basic task in every integration effort is Entity Resolution (ER), i.e., the ability to identify whether two pieces of information represent the same real world entity and then merge them into a single representation. This is inherently a quadratic task, requiring pair-wise comparisons among all objects in the collection. In order to scale ER to the volume of Web Data, blocking techniques are typically employed. However, most of the blocking techniques rely on schema information and are inapplicable to the highly heterogeneous settings of Web Data.

In collaboration with the L3S Research Center, we have proposed a novel framework consisting of two orthogonal layers (the effectiveness and efficiency layers), and showed how all blocking methods for highly heterogeneous data spaces map to this framework [41]. In addition, we have proposed several new techniques for improving the performance of ER [40, 41], namely, redundancy elimination, attribute-agnostic blocking, attribute clustering blocking, comparison pruning, and comparison scheduling, which all together offer significant time-performance improvements at a negligible cost on effectiveness. An interesting research direction would be to investigate the use of cloud computing and resource management technologies [28], in order to further improve the performance of the proposed techniques.

[On-the-Fly Entity Resolution] Traditional entity resolution approaches compute some similarity score between entities to decide whether they represent the same real world object. However, it is not clear what value of that score is high enough for such a decision, especially across different score computation methods. Furthermore, two representations may be seen as one in some cases and as independent in another, making the decision on the merging an application or query dependent.

With this is in mind, we have decided to take a radical approach and in collaboration with researchers from the L3S Research Center and the Technical University of Crete, we have developed LinkDB [26], a probabilistic linkage database system. We have thought that since it is hard to decide on whether a score is high enough to make a decision, we can postpone the decision until when needed. As such, we run entity linkage algorithms on our data but instead of doing any merging based on the results, we store the computed similarities in the repository alongside the data. At run time, when the user query is known, the system investigates what merges can be made and generate the answers the user query by evaluating it in a virtual database resulting from the merging of these entities [25]. An important feature of the system is that only the merges that affect the query results are taking place and not all the possible merges in the database. Another important feature is that the user may see results that are not in the database, but inferred from the stored data through on-the-fly merges.

[Schema and Data Mapping] A traditional topic of interest in the group is data mapping, i.e., the ability to associate data in one format with data in another. The work is rooted in our past participation in the development of Clio [20], a schema mapping tool from IBM Almaden. Clio was based on relational and semi-structured data, but we have ported this experience into

the development of mappings between entity repositories and ontologies. These new applications were developed in Papyrus [29], an EU project aiming at the creation of availability of content of one discipline to a different discipline, allowing people from the first, e.g., historians, to query data from the other, e.g., news, even if the latter uses a different terminology, structure [13], or language [52].

[Benchmarking] Evaluation is a fundamental step of every scientific finding, since it allows comparison with similar products and leads to informative decisions. Unfortunately, for many relatively novel areas like schema mapping and entity management, there is no globally accepted evaluation methodology. Researchers or vendors use their own tests and metrics leading into a blurred environment among similar products, developments or findings. Our group has spent a significant amount of effort into studying the problem and developing complete, consistent and principled evaluation methodologies that were recently presented in a tutorial [11]. Among these works is STBenchmark [1], a benchmark co-developed with the University of California Santa-Cruz, for evaluating mapping systems. It consists of a collection of test cases that can be used to measure the capabilities and limitations of a mapping system in terms of expressiveness and flexibility. Furthermore, it can dynamically generate of test cases at different levels of complexity and size, allowing the evaluation of the mapping systems in terms of scale [2]. In the same spirit with STBenchmark we have developed EMBench[1]. It is based on the same principles but it is designed for evaluating entity matching systems and can be a useful complement to the test cases provided by the OAEI initiative.

[Updates] Integration Systems have traditionally been considered read-only, mainly due to the fact that the system had no control over the data stored in the individual sources. Nevertheless, many times the integration reveals information not original available by the individual sources, thus, it may require that updates be issued on the integrated data. These updates have to be propagated to the sources. In collaboration with the AT&T Research Labs have studied ways to implement this goal [30, 46] and overcoming the difficulties that the view update theory is posing on the aspect.

4. STRUCTURED DATA ANALYTICS

[Data Stream Processing] The availability and use of (various types of) sensor networks have generated a lot of research interest. A major part of this effort has concentrated on how to efficiently collect and analyze the streams of sensed data. The dbTrento group has been working on several problems related to streaming data, ranging from data collection and representation, to data management, and analysis. Much of this work is also relevant to wireless sensor networks, and has been described in a recent survey of the area [37].

We have recently proposed a data-driven acquisition technique based on a linear model, DBP [43], that reduces the communication costs while mitigating the problems of noise and outliers. Relative to previous approaches, it can be much more efficiently implemented on resource-scarce nodes, and provides accuracy guarantees on the reported sensor measurements. Our work has shown that in the case of wireless sensor network deployments, further advances of the data management techniques would have little practical impact on the system lifetime [43]. Instead, improvements are more likely to come from radical changes at the routing and MAC layers, where new, data-aware protocols need to be designed. This work won the Mark Weiser Best Paper award in PerCom 2012.

The sheer number and size of the data we need to manipulate in many of the real-world applications dictates in several cases the need for a more compact representation than the raw data. We have developed novel, *amnesic* data approximation techniques that represent the most recent data with low error, and are more forgiving of error in older data, for arbitrary user-specified amnesic functions [38]. These techniques are incrementally maintainable, and are applicable to both landmark and sliding windows.

Several of the applications that consume streaming data, possibly from multiple sources, have high processing requirements over a significant portion of these data. We have developed a framework targeted to such applications, which aims to approximate in an online fashion multi-dimensional data series distributions [50]. This framework is adaptive, requires no a priori knowledge about the distributions of the sensed values, and it operates in a distributed fashion. We have demonstrated the applicability of the above framework in addressing two diverse and demanding problems, namely, identification and tracking of homogeneous regions [49], and outlier detection [50].

In a similar setting of multiple data stream processing in a network of nodes, we have proposed a technique for processing continuous queries that optimizes for the *profiled input throughput* that is focused on matching the expected behavior of the input streams [48]. We have also separately considered the problem of streaming sub-space clustering for high-dimensional spaces [36].

The efficient detection of frequent items in data streams is another interesting problem with many applications across domains. In this context, we have

[1]http://db.disi.unitn.eu/pages/EMBench/

performed a comprehensive comparative analysis of the available solutions, leading to several insights [31], and we have proposed a solution to the problem of finding recent frequent items in *ad hoc* windows in the past [51].

[Learning in Data Streams] Data streams can also carry information (e.g., user preferences) useful for learning algorithms. In this context, we have proposed a novel approach that can combine the content (descriptive aspect) and the type (directly quantifiable, or binary aspects) of the information instances, and studied the learning curves of the algorithms under different random information shifts [21]. This work won the Best Paper award in ADAPTIVE 2009.

Building on this work, we subsequently proposed an analytic model that describes the effect of the memory window size on the average prediction performance of a learning system, regardless of its underlying algorithm [22, 23]. We have additionally identified simple criteria, some of which are tied to specific data characteristics, that can be used by our framework in order to compare the behavior of learning algorithms in the presence of varying levels of noise [34].

[Data Series] There is an increasingly pressing need, by several applications in diverse domains (ranging from astronomy and biology, to electrical grids and manufacturing), for developing techniques able to index and mine very large collections of data series, in the order of hundreds of millions to billions. Evidently, this requirement calls for novel approaches and techniques for management and processing data series.

In this line of work, we have developed iSAX 2.0, a data structure designed for indexing and mining truly massive collections of data series [16], in collaboration with the University of California at Riverside. We showed that the main bottleneck in mining such massive datasets is the time taken to build the index, and we thus introduced the first bulk loading mechanism specifically tailored to a data series index, and reported the first published experiments to index one billion data series. Even though these results are promising for the practitioners, we observe that the analysis step cannot start before the lengthy indexing step ends. Removing this restriction is an interesting research direction.

In this area, we have also proposed fast and scalable techniques for pattern identification in data series streams [32]. The observation that in several cases the values in the data series are uncertain, has guided us to investigate this parameter of the problem. This value uncertainty may be due to the inherent imprecision of sensor observations, data aggregations, privacy-preserving transforms, or error-prone mining algorithms. Our study suggests that a fruitful research direction is to develop models for processing uncertain data series that take into account the temporal correlations in the data [18], which has not been considered so far.

5. ANALYTICS ON NON-STRUCTURED DATA

[Subjectivity Analysis] In the past years we have witnessed Sentiment Analysis and Opinion Mining becoming increasingly popular topics in Information Retrieval and Web data analysis, allowing us to capture sentiments and opinions, expressed in online user-generated content, at a large scale. Tracking how opinions or discussions evolve over time can help us identify interesting trends and patterns, and better understand the ways that information is propagated. The dbTrento group has been involved in research work relevant to the areas of Sentiment Analysis and Opinion Mining, and has spearheaded the work on Contradiction Analysis, which has also led to collaborations with HP Labs, the Qatar Computing Research Institute, and the Al Jazeera news broadcasting network. We have recently presented a comprehensive survey on the research problems in the above areas [56].

Our main focus has been the problem of finding sentiment-based contradictions at a large scale [57]. We defined two types of contradictions, depending on the distributions of opposite sentiments over time, namely, synchronous and asynchronous (sentiment-shift) contradictions, and introduced a novel measure of contradiction that accounts for the variability within and across data collections. We also proposed a scalable method for identifying both types of contradictions at different time scales that employs sentiment values on a continuous scale. An interesting direction of research is to characterize (e.g., in terms of demographics) and explain (e.g., in terms of news events) the identified contradictions, as well as generalize the proposed model to arbitrary opinion data (i.e., not just numeric sentiments) [55].

[Facet Discovery] Advances in social-media and user-generated content technologies have resulted in collecting extremely large volumes of user-annotated media; for instance photos (flickr), urls (del.icio.us), and others. All these platforms provide users with the capability of generating content and assigning *ad hoc* tags to this content. Motivated by applications in the domain of collaborative tagging, we have introduced the problem of diverse dimension decomposition, which can be used for facet discovery, where a dimension is a set of mutually exclusive tag-sets. The information theoretic mining framework we have proposed together with Yahoo! Research can be interpreted as a dimensionality-reducing transformation from the space of all tags to the space of orthogonal dimensions [53, 54].

6. FUTURE DIRECTIONS

The group continues the work on data and information management and analysis, with an emphasis on the problems arising from the scale of the data, their non-structured, heterogeneous and uncertain nature, and from specific application requirements, such as privacy guarantees on public administration data, or particular analysis tasks on scientific data. More specifically, the main research directions of the group are the following.

[Smart Cities] The availability of data and information on several different aspects of everyday life in digital format allows us to form a clear picture about the workings of a city, or a community in general. This can help us design tools for better managing fundamental principles relevant to citizens (such as privacy [17]) in new unexplored contexts, e.g., Big Open Data, and applications, e.g., Data Journalism. They will also enable us to react, follow on, predict, and influence various societal situations. In collaboration with the public administration and relevant industries, we will investigate novel techniques and methodologies that can help us achieve the above goals, even in new fields, like e-crime.

[User-Generated Content] Complementary to the previous direction is that of analyzing user-generated content. Given the wealth of such data on the web, we aim to develop a subjectivity analysis toolset that will take into account social structures and events information, and will offer intuitive analytics functionalities for understanding, explaining, and predicting trends and behaviors on the social web. Furthermore, the toolset will be predicting goals, user intentions and will be building dynamic user profiles from user generated content and user actions. Finally, it will be able to evaluate the quality of the provided information using the history of the users and the reactions of the crowds.

[Scientific Data] Through the ongoing collaboration with scientists, e.g., biologists and neuroscientists, who need to analyze large collections of data, usually on commodity hardware, we are aiming at providing them with tools that can efficiently perform complex analytics that take into account the special nature of their data and their intended tasks.

Acknowledgments

We thank all our postdoc, PhD and MSc students for their dedication and hard work: S. Bykau, A. Camerra, A. Chiasera, A. Cordioli, M. Dallachiesa, V. Falletta, G. Giannakopoulos, E. Iori, M. Lissandrini, A. Marascu, K. Mirylenka, D. Mottin, B. Nushi, D. Papadimitriou, M. Zerega, F. Rizzolo, C. Tsinaraki, M. Tsytsarau, and K. Zoumpatianos. We would also like to acknowledge the contributions of our internal and external collaborators, who made this research possible.

References

[1] B. Alexe, W. C. Tan, and Y. Velegrakis. Comparing and evaluating mapping systems with STBenchmark. *PVLDB*, 1(2), 2008.

[2] B. Alexe, W. C. Tan, and Y. Velegrakis. STBenchmark: towards a benchmark for mapping systems. *PVLDB*, 1(1), 2008.

[3] N. Augsten, D. Barbosa, M. H. Böhlen, and T. Palpanas. Tasm: Top-k approximate subtree matching. In *ICDE*, 2010.

[4] N. Augsten, D. Barbosa, M. H. Böhlen, and T. Palpanas. Efficient top-k approximate subtree matching in small memory. *TKDE*, 23(8), 2011.

[5] B. Bazzanella, T. Palpanas, and H. Stoermer. Towards a general entity representation model. In *IRI*, 2009.

[6] S. Bergamaschi, E. Domnori, F. Guerra, R. Trillo Lado, and Y. Velegrakis. Keyword Search over Relational Databases: A Metadata Approach. In *SIGMOD*, 2011.

[7] S. Bergamaschi, E. Domnori, F. Guerra, M. Orsini, R. T. Lado, and Y. Velegrakis. Keymantic: Semantic Keyword based Searching in Data Integration Systems. *PVLDB*, 3(2), 2010.

[8] S. Bergamaschi, F. Guerra, S. Rota, and Y. Velegrakis. A Hidden Markov Model Approach to Keyword-based Search over Relational Databases. In *ER*, 2011.

[9] S. Bergamaschi, F. Guerra, S. Rota, and Y. Velegrakis. KEYRY: a Keyword-based Search Engine over Relational Databases based on a Hidden Markov Model. In *ER*, 2011.

[10] S. Bergamaschi, F. Guerra, S. Rota, and Y. Velegrakis. Understanding Linked Open Data through Keyword Searching: the KEYRY approach. In *LWDM*, 2011.

[11] A. Bonifati and Y. Velegrakis. Schema Matching and Mapping: From Usage to Evaluation. In *EDBT*, 2011.

[12] P. Bouquet, T. Palpanas, H. Stoermer, and M. Vignolo. A conceptual model for a web-scale entity name system. In *ASWC*, 2009.

[13] S. Bykau, N. Kiyavitskaya, C. Tsinaraki, and Y. Velegrakis. Bridging the Gap Between Heterogeneous and Semantically Diverse Content of Different Disciplines. In *FLEXDBIST*, 2010.

[14] S. Bykau, J. Mylopoulos, F. Rizzolo, and Y. Velegrakis. Supporting Queries Spanning Across Phases of Evolving Artifacts using Steiner Forests. In *CIKM*, 2011.

[15] S. Bykau, J. Mylopoulos, F. Rizzolo, and Y. Velegrakis. On Modeling and Querying Concept Evolution. *Journal on Data Semantics*, 1, 2012.

[16] A. Camerra, T. Palpanas, J. Shieh, and E. J. Keogh. isax 2.0: Indexing and mining one billion time series. In *ICDM*, 2010.

[17] A. Chiasera, F. Casati, F. Daniel, and Y. Velegrakis. Engineering Privacy Requirements in Business Intelligence Applications. In *SDM*, 2008.

[18] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas. Uncertain time-series similarity: Return to the basics. *PVLDB*, 5(11), 2012.

[19] R. De Virgilio, F. Guerra, and Y. Velegrakis. *Semantic Search over the Web*. Springer, 2012.

[20] R. Fagin, L. M. Haas, M. A. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis. Clio: Schema mapping creation and data exchange. In A. Borgida, V. K. Chaudhri,

P. Giorgini, and E. S. K. Yu, editors, *Conceptual Modeling: Foundations and Applications*, volume 5600 of *Lecture Notes in Computer Science*. Springer, 2009.

[21] G. Giannakopoulos and T. Palpanas. Content and type as orthogonal modeling features: a study on user interest awareness in entity subscription services. *International Journal of Advances on Networks and Services*, 3(2), 2010.

[22] G. Giannakopoulos and T. Palpanas. The effect of history on modeling systems' performance: The problem of the demanding lord. In *ICDM*, 2010.

[23] G. Giannakopoulos and T. Palpanas. Revisiting the effect of history on learning performance: The problem of the demanding lord. *KAIS*, accepted for publication.

[24] O. Hassanzadeh, A. Kementsietsidis, and Y. Velegrakis. Data Management Issues on the Semantic Web. In *ICDE*, 2012.

[25] E. Ioannou, W. Nejdl, C. Niederee, and Y. Velegrakis. OntheFly Entity-Aware Query Processing in the Presence of Linkage. *PVLDB*, 3(1), 2010.

[26] E. Ioannou, W. Nejdl, C. Niederee, and Y. Velegrakis. LinkDB: A Probabilistic Linkage Database System. In *SIGMOD*, 2011.

[27] E. Ioannou, C. Niederee, and Y. Velegrakis. Enabling Entity-Based Aggregators for Web 2.0 data. In *WWW*, 2010.

[28] E. Iori, A. Simitsis, T. Palpanas, K. Wilkinson, and S. Harizopoulos. Cloudalloc: A monitoring and reservation system for compute clusters. In *SIGMOD*, 2012.

[29] A. Katifori, C. Nikolaou, M. Platakis, Y. Ioannidis, A. Tympas, M. Koubarakis, N. Sarris, V. Tountopoulos, E. Tzoanos, S. Bykau, N. Kiyavitskaya, C. Tsinaraki, and Y. Velegrakis. The Papyrus Digital Library: Discovering History in the News. In *TPDL*, 2011.

[30] Y. Kotidis, D. Srivastava, and Y. Velegrakis. Updates Through Views: A New Hope. In *ICDE*, pages 13–24, 2006.

[31] N. Manerikar and T. Palpanas. Frequent items in streaming data: An experimental evaluation of the state-of-the-art. *DKE*, 68(4), 2009.

[32] A. Marascu, S. A. Khan, and T. Palpanas. Scalable similarity matching in streaming time series. In *PAKDD*, 2012.

[33] Z. Miklos, N. Bonvin, P. Bouquet, M. Catasta, D. Cordioli, P. Fankhauser, J. Gaugaz, E. Ioannou, H. Koshutanski, A. Mana, C. Niederee, T. Palpanas, and H. Stoermer. From web data to entities and back. In *CAiSE*, 2010.

[34] K. Mirylenka, G. Giannakopoulos, and T. Palpanas. Srf: A framework for the study of classifier behavior under training set mislabeling noise. In *PAKDD*, 2012.

[35] D. Mottin, T. Palpanas, and Y. Velegrakis. Entity Ranking Using Click-Log Information. *Intelligent Data Analysis Journal*, 17:5, 2013.

[36] I. Ntoutsi, A. Zimek, T. Palpanas, P. Kröger, and H.-P. Kriegel. Density-based projected clustering over high dimensional data streams. In *SDM*, 2012.

[37] T. Palpanas. Real-time data analytics in sensor networks. In C. Aggarwal, editor, *Managing and Mining Sensor Data*. Springer, 2012.

[38] T. Palpanas, M. Vlachos, E. J. Keogh, and D. Gunopulos. Streaming time series summarization using user-defined amnesic functions. *TKDE*, 20(7), 2008.

[39] G. Papadakis, G. Giannakopoulos, C. Niederée, T. Palpanas, and W. Nejdl. Detecting and exploiting stability in evolving heterogeneous information spaces. In *JCDL*, 2011.

[40] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. In *WSDM*, 2012.

[41] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederee, and W. Nejdl. A blocking framework for entity resolution in highly heterogeneous information spaces. *TKDE*, accepted for publication.

[42] A. Presa, Y. Velegrakis, F. Rizzolo, and S. Bykau. Modeling Associations through Intensional Attributes. In *ER*, 2009.

[43] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco. What does model-driven data acquisition really achieve in wireless sensor networks? In *PerCom*, Lugano, Switzerland, 2012.

[44] F. Rizzolo, Y. Velegrakis, J. Mylopoulos, and S. Bykau. Modeling Concept Evolution: A Historical Perspective. In *ER*, 2009.

[45] D. Srivastava and Y. Velegrakis. Intensional Associations between Data and Metadata. In *SIGMOD*, pages 401–412, 2007.

[46] D. Srivastava and Y. Velegrakis. MMS: Using Queries As Data Values for Metadata Management. In *ICDE*, pages 1481–1482, 2007.

[47] D. Srivastava and Y. Velegrakis. Using Queries to Associate Metadata with Data. In *ICDE*, pages 1451–1453, 2007.

[48] I. Stanoi, G. A. Mihaila, T. Palpanas, and C. A. Lang. Whitewater: Distributed processing of fast streams. *TKDE*, 19(9), 2007.

[49] S. Subramaniam, V. Kalogeraki, and T. Palpanas. Distributed Real-Time Detection and Tracking of Homogeneous Regions in Sensor Networks. In *RTSS*, Rio de Janeiro, Brazil, 2006.

[50] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *VLDB*, 2006.

[51] F. I. Tantono, N. Manerikar, and T. Palpanas. Efficiently discovering recent frequent items in data streams. In *SSDBM*, 2008.

[52] C. Tsinaraki, Y. Velegrakis, N., and J. Mylopoulos. A Context-based Model for the Interpretation of Polysemous Terms. In *ODBASE*, 2010.

[53] M. Tsytsarau, F. Bonchi, A. Gionis, and T. Palpanas. Diverse dimension decomposition of an itemset space. In *ICDM*, 2011.

[54] M. Tsytsarau, F. Bonchi, A. Gionis, and T. Palpanas. Diverse dimension decomposition for itemset spaces. *KAIS*, accepted for publication.

[55] M. Tsytsarau and T. Palpanas. Towards a framework for detecting and managing opinion contradictions. In *ICDM Workshops*, 2011.

[56] M. Tsytsarau and T. Palpanas. Survey on mining subjective data on the web. *DMKD*, 24(3), 2012.

[57] M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable discovery of contradictions on the web. In *WWW*, 2010.