Data Science approach for a cross-disciplinary understanding of urban phenomena

Application to energy efficiency of buildings based on physical measures and user behaviours

Servigne Sylvie[1], Gripay Yann[1], Deleuil Jean-Michel [2], Jay Jacques[3], Cavadenti Olivier[1], Mebrouk Radouane[1]

1. Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205
Laboratoire d'Informatique en Image et Systèmes d'Information
prenom.nom@insa-lyon.fr

2. Université de Lyon, CNRS, INSA-Lyon, EVS, UMR5600
Laboratoire Environnement, Ville et Société
prenom.nom@insa-lyon.fr

3. Université de Lyon, CNRS, INSA-Lyon, CETHIL, UMR5008
Centre de Thermique de Lyon
prenom.nom@insa-lyon.fr

1. Context and motivation

Urban-related phenomena are complex contextual phenomena. Researchers and institutions may have multiple ways to collect data to study these phenomena so as to build and validate modellings. Data are collected through dedicated campaigns: instrumentation for infrastructure monitoring like roads, buildings or urban networks; instrumentation for environmental monitoring like air, water and so on; surveys on users, inhabitants and citizens; historical and sociological studies... Those campaigns generate a large amount of heterogeneous and complex data, that are multi-sources, multi-dimensional (e.g., spatio-temporal data), and multi-media (e.g., numbers, texts, images, sounds, videos). Deep analyses of those data can lead to hypothesis validation and knowledge discovery. Such analyses requires advanced

skills to first understand raw data, to then discover their multi-scale properties, and to finally perform relevant aggregations and cross-sources comparisons.

In our project, we focus on the context of building-related phenomena. Nowadays energy efficiency is theoretically estimated for a new empty building. However, the practical efficiency is usually lower due to the complexity of the building process and due to behaviours of real occupants. If energy consumption can be directly measured through instrumentation, understanding the practical energy efficiency of a building requires a multi-disciplinary approach to understand energy consumption with regards to actual uses of the building. Cross-analyses of "instrumentation data" and "survey data" (and other studies data) are thus necessary to fully discover and then understand complex correlations between physical and human parameters.

Our objective is to develop theoretical and practical tools to model, explore and exploit heterogeneous data from various sources in order to understand a phenomenon of interest. We focus on the design of a generic model for data acquisition campaigns based on the concept of generic sensor. The concept of generic sensor is centred on acquired data and on their inherent multi-dimensional structure, to support complex domain-specific or field-oriented analysis processes. We consider that a methodological breakthrough, based on Data Science as a pivot for interdisciplinary dialog, may pave the way to deep understanding of voluminous and heterogeneous scientific data sets.

2. Methodology

Our approach revolves around a generic conceptual modelling of sensor data with a real multi-disciplinary approach named "Data Science approach" that involves researchers from computer science, human science and thermal science. The resulting generic model has to represent heterogeneous data sources. We first took a bottom-up strategy. We designed a model for heterogeneous physical sensors, from our real experimentation platform in occupied buildings, and another model for sociological surveys, to take into account opinions and feelings of occupants.

We then took a top-down strategy. Inspired by state-of-the-art abstract ontologies that describe sensor systems (Reed 2007, DUL 2010, Compton 2011), we designed the Virtual Generic Sensor model (VGS model) that encompasses (among others) data from physical sensors and surveys. This model focuses on data produced by those generic sensors, and on a common multi-dimensional structuring. The resulting structure is mainly based on time and space, but is designed to support specific field-oriented dimensions.

We designed a methodology for an agile multi-dimensional exploration of those data. Based on the VGS model and its multi-dimensional structure, we propose a language to finely define domain-specific or field-oriented indicators through successive aggregations along dimensions (in a similar way to Data Warehouses). We also designed a visualization framework linked to those dimensions that enables users to visually explore indicators using graphs. We further offer to visually

compare those graphs thanks to an interactive "matrix layout". The common multi-dimensional structure of data is then exploited at 3 levels: to structure data, to define indicators, and to explore data through these indicators.

Our "agile" approach allows incremental and iterative data processes and analyses. At the data level, we consider incremental data sets, i.e., with raw data sets still being captured. The data set generation is also iterative: data can be progressively enriched with new interpreted data. At the analysis level, we consider an incremental exploration process, with new (aggregated) indicators that can be added when needed. The exploration process is also iterative when knowledge from exploration is used to refine further explorations, with refined or new dimensions, adjusted granularity, new points of view... This approach is designed to support and enrich current domain-specific approaches for complex and/or scientific data analyses.SPHINX iQ (2012) est un logiciel commercial qui permet de gérer des enquêtes et ana-lyser les données, quelle que soit leur nature, quantitative ou qualitative. SPHINX permet de créer des questionnaires composés de plusieurs types de questions : questions fermées (uniques, multiples ou ordonnées), ouvertes, échelles graduées, numériques, etc. Le logiciel propose des outils intégrés pour l'analyse des données qualitatives ou quantitatives et fournit différents indicateurs statistiques (effectifs, pourcentages, moyennes, médianes, écart-type, spécificités). La documentation sur le logiciel SPHINX nous a permis d'approfondir les spécificités liées à la conception des questionnaires à savoir les différents types de questions réponses. De plus nous nous sommes aperçus de l'importance de la phase d'analyse des données d'enquêtes qui constitue un des objectifs importants du recueil de données. Toutefois de nombreux utilisateurs d'enquêtes se contentent d'un tableur pour gérer les questionnaires (Deleuil 2010). En effet, les tableurs offrent une structuration basique accompagnée d'outils relativement accessibles. Pourtant, ces outils ne sont pas destinés à la gestion et l'exploitation de masses de données hétérogènes contrairement aux bases de données et aux entrepôts. Nous nous sommes donc inspirés d'une partie des concepts sous-jacents aux produits du marché pour concevoir un modèle conceptuel essentiellement centré sur les questionnaires et les données collectées.

3. Results

Our first result is the VGS (Virtual Generic Sensor) model (cf. Figure 1). It describes the static structure of a generic acquisition system, with a Sensor composed of several Detectors (further detailed by MeasureAttributes), and a dynamic structure with Samples, produced by a Sensor, composed of Measures (further detailed as Values). This model has been implemented for two Smart Building experimentation platforms managed by the LIRIS (SoCQ4Home, since October 2012 & MARBRE, since February 2014). A total of around 400 physical sensors are producing data: temperature, humidity, CO2/VOC, contact for doors/windows, weather station. A survey concerning 50 occupants of one of the buildings has been realised and results of questionnaires are going to be integrated, with questionnaires modelled as sensors, and answers to questions modelled as

measures. The current implementation of the VGS model is based on a MySQL database in particular to benefit from the expressiveness of the “golden standard” SQL language; moreover MySQL is a free open source tool.

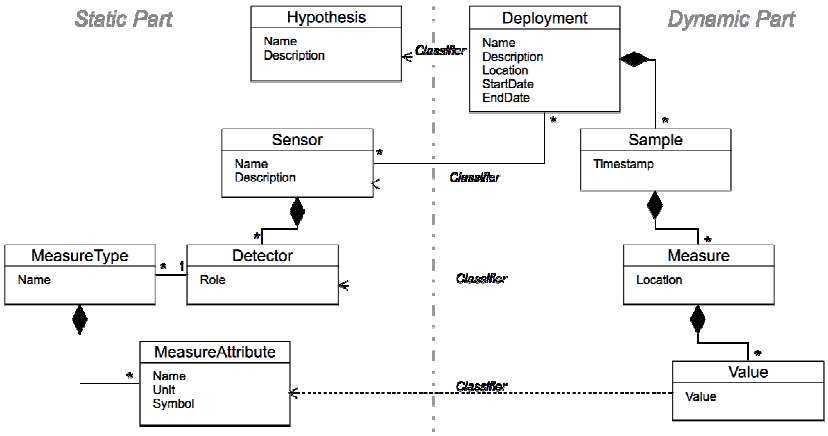

Figure 1. VGS Model with UML representation

In the context of the SoCQ4Home platform (about 20 instrumented rooms in a building), we illustrate data cross-visualisation of integrated heterogeneous data with an example based on simulated data. Figure 2 shows three temperature indicators aggregated by day: maximum, minimum and average of temperature measures (three curves, temperature values on left axe); and one temperature feeling indicator also aggregated by day, from a daily survey on users (vertical bars, values on right axe with: 3: too hot, 0: satisfying, -3: too cold). Those indicators are visualised for one month (31 days).

Our second result is a formal model and a declarative language to finely define indicators as aggregations along dimensions for VGS data. It is based on the relational algebra (a foundation concept for relational databases, like SQL databases). In our current prototype, we implemented this language by automatically translating it to complex nested SQL aggregation queries. It enables to easily define new domain-specific dimensions and/or to adapt existing dimensions (like time and space). We also implemented a proof-of-concept Web user interface to visually define user-specific aggregations.

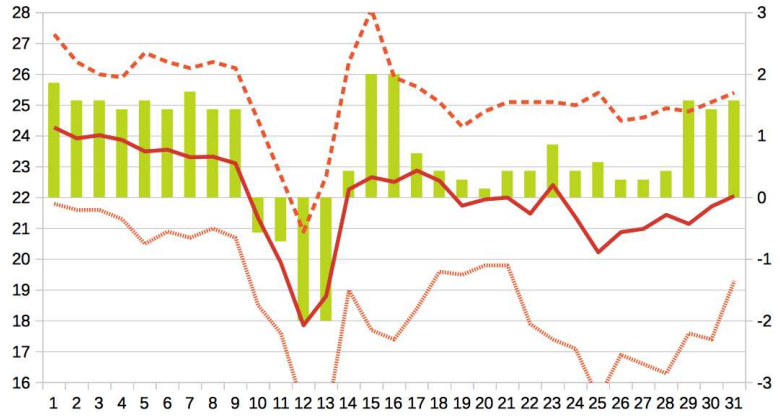

Figure 2.a. Cross-visualisation of temperature data and survey data with corresponding temporal dimension

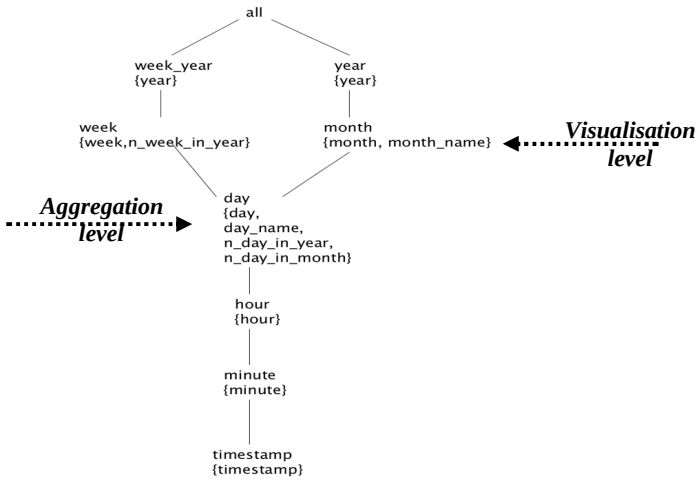

Figure 2.b. Dimension hierarchy: difference between aggregation level and visualisation level of Figure 2a.

Our third result is a proof-of-concept Web user interface to visualize data and/or indicators as a matrix of graphs. This style of visualisation is illustrated in Figure 3: one line per sensor in a graph, and one graph per room (rows) for Temperature and Humidity (columns); graph scales are identical within a column. The development of a full exploration Web interface, including definition of dimensions, indicators, and the dynamic navigation along dimensions is a work in progress.

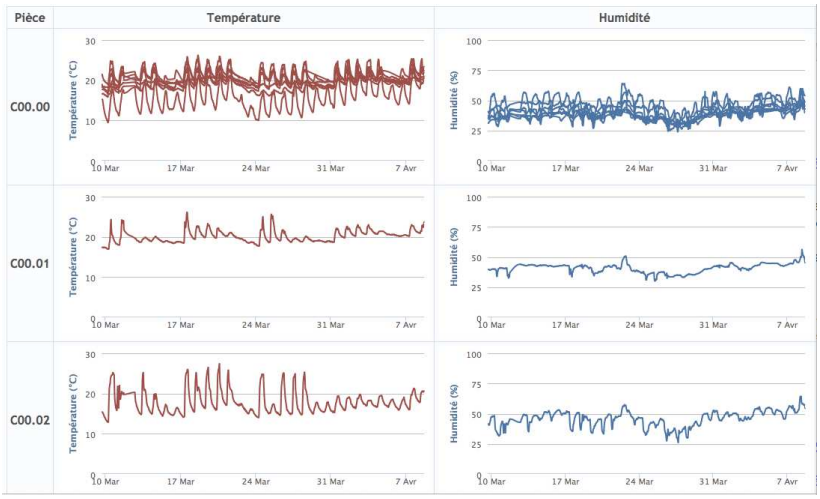

Figure 3. Web user interface to visualise data as a matrix of graphs (actual data from MARBRE platform)

5. Conclusion

Our approach contributes to a better understanding of urban-related phenomena through cross-disciplinary analyses of large amount of data coming from phenomenon observations issued from multiple sources (sensors, surveys, various studies). This approach aims at being well integrated with user-specific and domain-specific analyses processes, in particular for scientific data analyses. Analyses may be multi-dimensional through the definition of dimensions: spatial dimensions, temporal dimensions, and field-specific dimensions. Dimensions dedicated to a given phenomenon have to be identified and co-built by computer scientists and scientists from other urban-related disciplines. Our methodology is agile, incremental, iterative, and interactive to allow knowledge discovery along the way by users.

Our VGS model is generic as it enables heterogeneous data handling. Some existing works focused on physical sensor and real-time systems (Bonnet 2001, Kassim 2011, Diallo 2012) while we focused on heterogeneous data sources (Noel 2005). The VGS model is semantically compatible with standard sensor ontologies defined by standardization organizations like Open Geospatial Consortium standards (Reed 2007, DUL 2010, Compton 2011). The DUL description is system-centred with low level of detail concerning data while our VGS model is data-centred.

Our conceptual VGS model has been built from a real multi-disciplinary approach and its generic design makes it easy to apply to other phenomena observation. This model, as well as our agile exploration approach, is moreover independent from a specific data management technology. Although it is currently

implemented on a SQL database, we aim at also implementing it on Big-Data-oriented databases like MongoDB or Cassandra.

Bibliography

Bonnet P., Gehrke J., Seshadri P. (2001). Towards Sensor Database Systems. Proceedings of Mobile Data Management. Conference. LNCS Springer. Pp.3-14

Compton M. et al. (2011). The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. 7p.

http://www.ict.csiro.au/staff/kerry.taylor/SSN-XG_SensorOntology.pdf

Diallo O. et al. (2012). Real-time datamanagement on wireless sensor networks: A survey. Journal of Network and Computer Applications. vol. 35, no3. Pp. 1013-1021

DUL (2010). The DOLCE+DnS Ultralite ontology. http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite.

Gripay Y., Laforest F., Petit J-M. (2010). A Simple (yet Powerful) Algebra for Pervasive Environments. EDBT 2010, 13th International Conference on Extending Database Technology, Lausanne, Switzerland. pp. 1-12.

Gutierrez C., Servigne S. (2013). Managing Sensor Data Uncertainty: A Data Quality Approach. International Journal of Agricultural and Environmental Information Systems. Pp. 35-54

Kassim M. et al. (2011). A Based Temperature Monitoring System. International Journal of Multidisciplinary Sciences and Engeneering. Vol. 2, N°. Pp.17-25.

Lumineau N., Laforest F., Gripay Y., Petit J-M. (2012). Extending Conceptual Data Model for Dynamic Environment. In 31st International Conference on Conceptual Modeling (ER 2012), Florence, Italy.

Mebrouk R. (2012). Modélisation conceptuelle générique pour un monitoring éco-informatique durable. Rapport de Master. Soutenu par le Labex IMU : Intelligence des Mondes Urbains.

Noel G., Servigne S., Laurini R. (2005). Spatial and Temporal information structuring for natural risk monitoring. Proceedings of the GIS Planet Conference. 10 p.

Noel G., Servigne S. (2005). Indexation multidimensionnelle de bases de données capteur temps-réel et spatio-temporelles. Ingénierie des Systèmes d'Information. Vol.10, n°4. pp.59-88

Patil N. S. et al. (2011). Data aggregation in wireless sensor network. International Journal Of Service Computing And Computational Intelligence, vol. 1, no 1, p. 7–10.

Reed C., Botts M., Davidson J., Percivall G. (2007). Ogc® sensor web enablement: overview and high level achitecture. Autotestcon, 2007 IEEE. pp. 372–380

SensorML (2007). OpenGIS® Sensor Model Language (SensorML) Implementation Specification. Open Geospatial Consortium. pp. 1–87. http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/.SensorML

SoCQ4Home platform (2012) & MARBRE platform (2014). SoCQ4Home Project, http://liris.cnrs.fr/socq4home/