Data Science and Decision Support at ERIC

Fadila Bentayeb, Julien Velcin, Stephane Bonnevay and Jerome Darmont
Universite de Lyon (Laboratoire ERIC)
Universite Lumiere Lyon 2 – 5 avenue Pierre Mendes-France
69676 Bron Cedex – France
http://eric.univ-lyon2.fr

1. OVERVIEW

ERIC is the French acronym for *Entrepôts, Représentation et Ingénierie des Connaissances*, which literally translates into "Warehouses, Representation and Knowledge Engineering". The ERIC laboratory was created in 1995 and is a joint research unit of two of the three universities in Lyon, both ranked among top universities in France and Europe in their respective fields: Université Claude Bernard Lyon 1, a university of science and medicine, and Université Lumière Lyon 2, a university of humanities and social sciences. ERIC is also a member of the *Institut des Sciences de l'Homme*, a federative institute related to the French National Center for Scientific Research (CNRS).

Research at ERIC aims at extracting value from huge, complex databases, especially (but not exclusively) in the fields of humanities. Our expertise lies in the following domains (Figure 1): 1) data warehousing: intelligent integration of complex data, multidimensional modeling of complex objects, personalized on-line analysis processing (OLAP), data warehouse security; 2) data mining and decision: machine learning, graph study and graph mining, complex data analysis, multicriteria aggregation, opinion mining, data mining software.

ERIC is constituted of two research teams: Decision support Information Systems (SID in French) and Data Mining and Decision (DMD), which address the two aforementioned domains, respectively. Overall, about 50 people regularly work at ERIC, including 6 full professors, 16 associate professors, 25 Ph.D. students and 1 to 3 postdoctoral fellows, depending on the period.

Eventually, the ERIC laboratory is involved in numerous academic and industrial partnerships all around the world, and promotes its research fields by steering and organizing both domestic (e.g., EGC and EDA, both referenced in DBLP) and international events (e.g., the ALT/DS 2012 joint conferences or the VLDB Cloud Intelligence workshop).

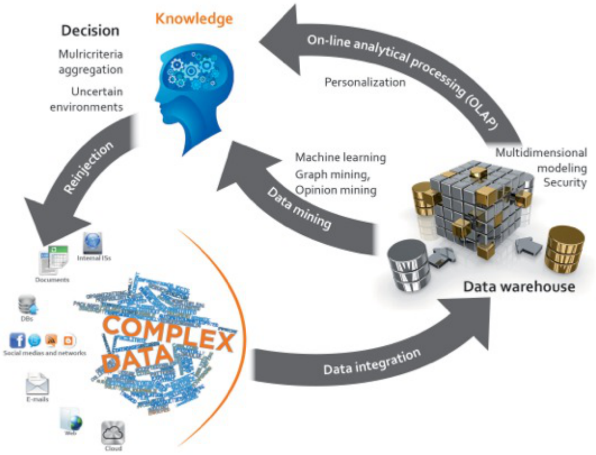


Figure 1: ERIC's Research Topics

In the remainder of this paper, we detail the research topics, some representative results and perspectives of both our research teams.

2. SID TEAM

The expertise of the SID team spans a wide range of areas and applications in data warehousing and OLAP. We aim at developing novel models and methods for designing, storing and analyzing very large-scale data by providing adequate big data warehouses. We address a number of aspects of decision support systems, but mostly focus on the extract-transform-load (ETL) process, multidimensional modeling, data analytics and data/knowledge management infrastructures. We contribute prominently to cloud and big data analytics, by dealing with data from any source, including the Web and social networks.

2.1 Complex Data Warehousing

Traditional systems are very successful in integrating and warehousing structured data for analysis. However, structured data represent only a small subset of interesting data that could be warehoused by many organizations. Moreover, the classical star schema [12] and its derivatives (snowflake and constellation schemas) are actually relational logical schemas, and prove limited for handling complex data. Thus, we propose methods and tools for integrating and warehousing complex data.

First, to answer to the increasing demand for handling complexity in data sources, we propose an active ETL framework [17] that allows: 1) integrating complex data into an Active XML (AXML) repository [2]; 2) exploiting active rules and mining logged events to self-manage, automate and activate integration tasks.

Our ETL framework is implemented and deployed as a Web application and consists of three main modules: integration services, an event repository, and a management and reactivity module (Figure 2).

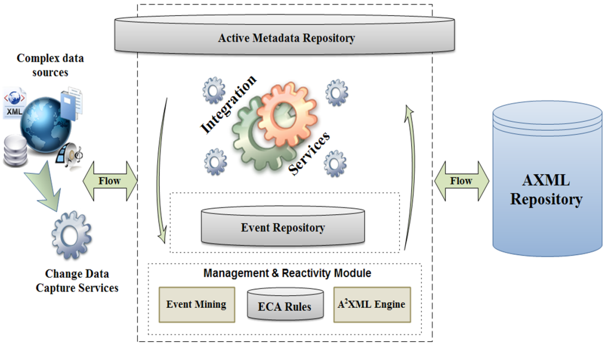


Figure 2: Complex data ETL framework

Integration services input source data and output a set of AXML documents, warehoused in a native XML-based repository. Employing XML and Web services for integrating data helps tackle data heterogeneity, interoperability, distribution and freshness. Moreover, integration services rely on metadata to ensure minimal user intervention for maintaining services. Moreover, services can also be invoked autonomously by triggering Event-Condition-Action (ECA) rules or by AXML calls.

The event repository logs all events related to data sources, integration services and AXML document querying. We apply data mining techniques onto logged events to discover rules [18], which are thereafter used to maintain, automate and reactivate data integration services.

Finally, we achieve system reactivity with a set of active rules that follow the ECA paradigm. ECA rules may be user-defined or automatically mined from event logs. Moreover, invoking embedded services in AXML documents refreshes the repository with up-to-date information. Such embedded services are managed by the AXML engine.

Once complex data are integrated into an AXML repository, which may be viewed as an operational data storage (ODS), we design multidimensional models by exploiting the object-oriented paradigm. We define a layered multidimensional model based on the concept of complex object, which encapsulates data and structure complexity and eases the creation and manipulation of complex data cubes. This model takes full advantage of the object-oriented paradigm to capture multidimensional concepts by symmetrically considering facts and dimensions [4]. In our model, facts and dimensions need indeed not be predefined at the conceptual level, but are designated at analysis time.

Our complex multidimensional model comprises four concepts: complex object, complex relationship, attribute hierarchy and object hierarchy. The package diagram layer models the universe as a set of complex objects, some of which being organized into hierarchies. Complex objects are linked by a set of complex relationships. The class diagram layer provides details about both the structure of each complex object and the origin of complex relationships, which allows defining attribute hierarchies.

This two-layer multidimensional modeling allows users designing complex cubes. To extract complex cubes from the object multidimensional schema, we propose an OLAP operator called cubic projection. Finally, a third layer is constituted of a metamodel for complex cubes that explicitly represents facts and dimensions. Moreover, we provide a translation algorithm that maps any conceptual schema into an XML logical schema, from which an XML physical schema is derived [5]. New analytical functions based on the nature of measure attributes are currently in the pipe.

2.2 Textual OLAP

Recent studies confirm that most data exploited in businesses and administrations are textual, e.g., reports, resumes, e-mails, social data, etc. While OLAP proves very useful for analyzing structured data, it faces several challenges in handling textual data. For example, aggregating numerical data is performed by using standard aggregation functions such as sum, average, etc. Such functions are obviously not suitable to analyze textual data. Thence, to achieve OLAP analyses over textual data, we combine OLAP and information retrieval (IR).

Our key idea is to use a data cube to represent relationships within textual documents. The benefits are twofold: easier representation and processing of text queries, and the creation of new contexts constructed by analyzing existing data. We propose a contextual text cube model, called CXT-Cube, which includes several semantic dimensions associated with a specific textual measure [15].

Inspired by IR, we use a space vector model, i.e., an algebraic model that allows representing text documents by a vector of terms in a multidimensional space. Each semantic dimension is extracted from an external knowledge source, i.e., a domain ontology related to the dimension area. Its hierarchy specifies the semantic levels and relationships among text terms in the CXT-Cube. A textual measure is then defined by several vectors of weighted concepts, i.e., one vector per dimension. Term weight is computed with respect to term occurrence frequency and a relevance propagation method, which allows reassigning term scores from leaf nodes to their own ancestors in the concept hierarchy.

One of the possible application domains that would benefit from our textual OLAP approach is recruitment. Better recruitment decisions may indeed be achieved by pre-selecting applicants based on their resumes. A CXT-Cube can be built with respect to resumes (Figure 3) with two semantic dimensions, namely Topic and Location. The Topic dimension includes a topical hierarchy representing domain skills. The Location dimension has three hierarchical attributes City, Region and Country. Another contextual dimension is Time, which represents the sending date of resumes as metadata.

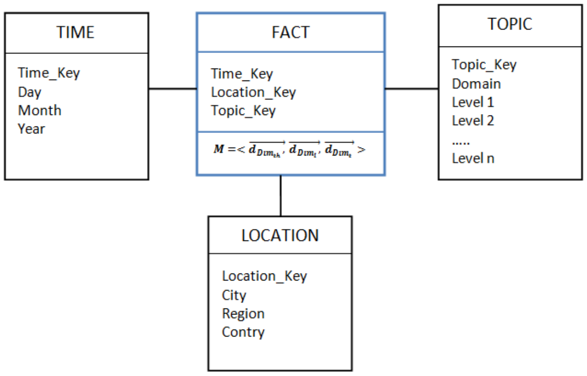
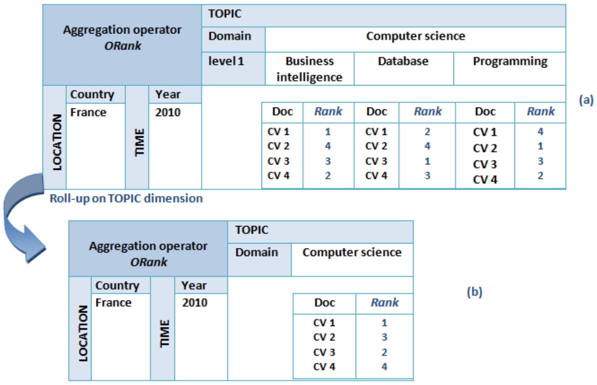


Figure 3: CXT-Cube of resumes

Of course, the main objective of this work is to allow users to easily interact with a CXT-Cube and execute textual queries. To this aim, we propose a new aggregation operator named ORank (OLAP-Rank). ORank's objectives are twofold: 1) aggregating and navigating through textual documents (e.g., resumes); 2) ranking such documents with respect to their relevance degree to a query (Figure 4).



Aggregation operator *ORank*				TOPIC						
				Domain	Computer science					
				level 1	Business intelligence		Database		Programming	
LOCATION	Country	TIME	Year							
	France		2010		Doc	*Rank*	Doc	*Rank*	Doc	*Rank*
					CV 1	1	CV 1	2	CV 1	4
					CV 2	4	CV 2	4	CV 2	1
					CV 3	3	CV 3	1	CV 3	3
					CV 4	2	CV 4	3	CV 4	2

(a)

Roll-up on TOPIC dimension

Aggregation operator *ORank*				TOPIC		
				Domain	Computer science	
LOCATION	Country	TIME	Year			
	France		2010		Doc	*Rank*
					CV 1	1
					CV 2	3
					CV 3	2
					CV 4	4

(b)

Figure 4: OLAP ranking on resumes

Experiments on resumes highlight the advantage of using ORank over traditional IR ranking functions with respect to recall, precision and precision@k, i.e., the number of relevant documents in the k first retrieved documents divided by k [15].

2.3 Future Research

The advent of cloud computing and the increasing exploitation of big data provide a natural application field to the research performed in the SID team. New opportunities, as well as new research challenges [6], pop up in this environment. We currently aim at developing a cloud computing environment for big data warehousing and OLAP.

Essential issues also remain in parallelizing analytical algorithms for big data to benefit from massively parallel architectures in the cloud. Existing scalable and parallel cloud computing frameworks able to process big data, such as Hadoop, are not adapted to all kinds of processing, e.g., intrinsically interactive processes such as OLAP. However, precomputing expected query answers or aggregates, e.g., by building materialized views or OLAP cubes, can definitely be parallelized to make subsequent visualization and navigation efficient. Thereafter, queries have to be rewritten to run onto distributed chunks of data. In the cloud, with budget coming in as a new constraint in the pay-as-you-go paradigm, view materialization and query rewriting is a scientific lock that is seldom addressed as of today.

On a much smaller scale that we could term "small data", we also start developing a so-called personal intelligence [1] platform called BI4people, to allow very small businesses, organizations or even individuals accessing to simple, cloud-based business intelligence tools.

Finally, as security remains one of the top concerns of cloud users and would-be users, we address data security issues (privacy, availability and integrity) with the help of new secret sharing schemes for cloud data warehouses and OLAP [3]. However, we still work on demonstrating that the cost of our solution is lower than that of data loss or pilfering, which requires skills in management science.

3. DMD TEAM

DMD team members aim at creating new systems, models and algorithms for data mining and decision making. Complex data come from heterogeneous sources. They are semi-structured, since they can be embedded in a graph-like structure whose nodes and edges are attached to various contents (e.g., text, image, metadata). They are voluminous, imprecise and highly dynamic.

To handle such data, we build on techniques coming from statistics and artificial intelligence: machine learning, information retrieval, multiobjective optimization, multicriteria aggregation, reasoning under uncertainty, etc. Moreover, we claim that producing theoretical results and applying our models to concrete cases are of equal importance. Many applications can be foreseen, such as medical image reconstruction, country ranking, reputation management, social media analysis.

3.1 Structuring Complex Data

In the context of big, interconnected data, the techniques related to data science are especially promising. For instance, DMD team members work on topic modeling for dealing with textual datasets extracted from the Web.

We particularly propose a temporal-aware topic sentiment (TTS) model for dealing with both topic and sentiment over time [7]. Our approach has several important features that are not jointly addressed by other models of the literature, such as ASUM [11] and JST [13]. First, time is jointly modeled with topics and sentiments, which allows capturing the evolution of sentiment about a topic over time. Second, topic-specific sentiments are extracted from the whole data at once, and not from each single document, providing an overall view of topic-sentiment correlations. Finally, no post-processing is needed to match similar topics under different sentiment polarities.

TTS' graphical model is an extension to the classical LDA model (Figure 5). The important point is that sentiment polarity s is drawn given topic z, before drawing both word w and timestamp t given tuple (z, s). We have estimated hidden parameters by classical Monte-Carlo sampling.

We favorably compare our model to other state-of-the-art topic-sentiment models on two datasets extracted from social media: the MDS dataset (Amazon product reviews) and an original dataset of news articles about the Strauss-Kahn case. In particular, we show that our model achieves better results than JST and ASUM for catching the topic-sentiment correlation over time [7].

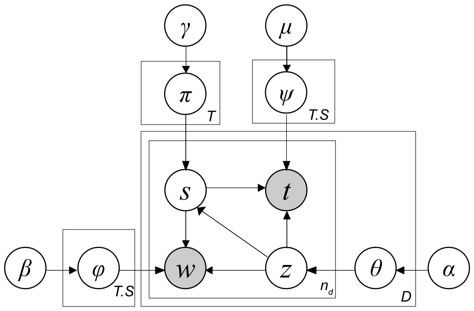


Figure 5: TTS model

We currently try to use this powerful holistic model to improve IR systems by adding new sentiment-oriented features built on TTS.

3.2 Ensemble Methods

Another contribution of the team leverages topological graphs to design metrics that are well-suited to machine learning. Relevant metrics are paramount for solving both unsupervised and supervised machine learning issues.

Obviously, if representing variables are well chosen, similar objects shall belong to the same class or cluster. However, similarity is a relative notion that heavily depends on the density of objects in the description space. This observation leads us to use regions of influence and neighborhood graphs. For instance, in Figure 6, X_{new}'s regions of influence is computed using a relative neighbor graph.

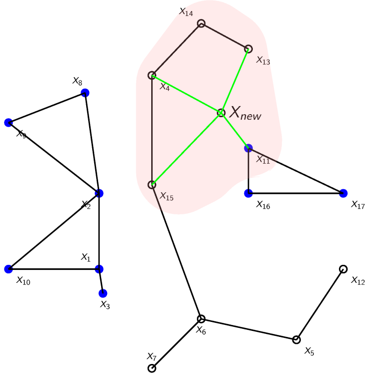


Figure 6: Sample relative neighbor graph

In the domain of supervised inductive learning, we particularly propose an original ensemble method based on neighborhood graphs [19]. This algorithm constructs several neighborhood graphs by using various projections and, for each of them, infers a label depending on their neighbors' labels. For instance, in Figure 6, this label may be chosen using either the direct neighbors or the two connected components linked to X_{new}.

Then, the different propositions are aggregated to achieve the final result. After conducting an extensive search for the best neighborhood graph and aggregation method, we selected the relative neighborhood graph, which achieves a good balance between too many and too few neighbors.

We compared our approach to state-of-the-art methods on 18 UCI datasets. Our experiments show that we challenge the most powerful techniques, such as Random Forests and SVM. Besides, our approach outperforms classical methods based on neighborhood, such has k-Nearest Neighbors. Our classifier is ranked first when using both the mean error and mean rank of all tested algorithms.

3.3 Multicriteria Decision

DMD team members who work in this field are interested in multicriteria decision, multiobjective decision and collective decision models. They conduct theoretical research to study the properties of various multicriteria analysis methods. Studying the properties of different models is essential to characterize them and understand the pros and cons of one specific model or method over another one.

In particular, we study the generation of Choquet optimal (C-optimal) solutions for biobjective combinatorial optimization problems. C-optimal solutions optimize a Choquet integral. The Choquet integral is used as an aggregation function, presenting different parameters and allowing to take interactions between objectives into account [10].

We propose a new property that characterizes C-optimal solutions [14]. From this property, we define a general method to easily generate optimal solutions in the case of two objectives. We apply our method to two classical biobjective optimization combinatorial optimization problems (BOCO): the biobjective knapsack problem and the biobjective minimum spanning tree problem.

We demonstrate that C-optimal solutions that are not weighted sum optimal (WS-optimal) solutions represent only a small proportion of all C-optimal solutions and are located in a specific area of the objective space, but are much harder to compute than WS-optimal solutions.

This work opens many perspectives. The property we introduce has to be generalized to problems with more than two objectives. It is also interesting to study and define what brings exactly C-optimal solutions that are not WS-optimal. Finally, some specific methods to compute all C-optimal solutions (branch and bound methods) or specific methods to optimize the single objective problems with additional constraints could be studied.

3.4 Future Research

Many exciting challenges lie at the crossroad of DMD members' expertise. To begin with, we continue manipulating graphs for detecting misclassified individuals in supervised learning [16]. Another line of research is to focus on the curse of dimensionality that considerably affects neighborhood graphs. In past studies, we fixed this issue by using random projections that were well-adapted to the general framework of ensemble methods. Now, we are investigating the use of dimension reduction techniques to solve the problem more generally.

In the global context of social media analysis, we recently paid attention to community detection and social role identification issues [8]. It was indeed showed that taking dynamics into account is crucial in this area [9]. A promising track of research consists in addressing both issues jointly, thanks to modern machine learning algorithms based on the description of people, what they write *and* their interactions.

Eventually, according to the multicriteria decision analysis (MCDA) literature, many methods have been developed in the field of multicriteria decision making for eliciting parameters when real preferences are observed. This (quite old) issue is called "preference learning" in machine learning and up to now, there has not been any effective convergence between machine learning and MCDA. One big issue would be to use the great experience of the machine learning community to improve MCDA elicitation methods.

We plan to explore many exciting issues, such as building a collaborative benchmark database in MCDA and studying the possibility of introducing experiment designs in MCDA. Another important issue will be to study how classical MCDA methods, and especially methods based on an outranking principle, can be efficiently used for enhancing machine learning techniques in the big data context.

4. ACKNOWLEDGEMENTS

This paper was written by the laboratory's board, whose members would like to acknowledge the dedication and hard work of all colleagues, Ph.D. students, postdoctoral fellows and administrative staff members. We would also like to thank our external

collaborators all around the world.

5. REFERENCES

[1] A. Abello, J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazon, F. Naumann, T.-B. Pedersen, S. Rizzi, J. Trujillo, P. Vassiliadis, and G. Vossen. Fusion Cubes: Towards Self-Service Business Intelligence. *International Journal of Data Warehousing and Mining*, 9(2):66–88, April-June 2013.

[2] S. Abiteboul, O. Benjelloun, and T. Milo. The Active XML project: an overview. *VLDB Journal*, 17(5):1019–1040, August 2008.

[3] V. Attasena, N. Harbi, and J. Darmont. fVSS: A New Secure and Cost-Efficient Scheme for Cloud Data Warehouses. In *17th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2014), Shangai, China*, pages 81–90, 2014.

[4] D. Boukraâ, O. Boussaid, and F. Bentayeb. Complex Object-Based Multidimensional Modeling and Cube Construction. *Fundamenta Informaticae*, 132(2):203–238, 2014.

[5] D. Boukraâ, O. Boussaïd, F. Bentayeb, and D. E. Zegour. Managing a fragmented XML data cube with Oracle and Timesten. In *15th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2012), Maui, HI, USA*, pages 97–104, 2012.

[6] J. Darmont, T.-B. Pedersen, and M. Middelfart. Cloud Intelligence: What is REALLY New? *1st International Workshop on Cloud Intelligence (Cloud-I 2012), Istanbul, Turkey*, August 2012. Panel.

[7] M. Dermouche, J. Velcin, S. Loudcher, and L. Khouas. A Joint Model for Topic-Sentiment Evolution over Time. In *IEEE International Conference on Data Mining (ICDM 2014), Shenzhen, China*, 2014.

[8] M. Forestier, A. Stavrianou, J. Velcin, and D. A. Zighed. Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems*, 10(1):117–133, 2012.

[9] W. Fu, L. Song, and E. P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *26th International Conference on Machine Learning (ICML 2009), Montreal, QC, Canada*, pages 329–336, 2009.

[10] M. Grabisch and M. Roubens. *Application of the Choquet integral in multicriteria decision making*, pages 348–374. Fuzzy Measures and Integrals – Theory and Applications. Physica Verlag, 2000.

[11] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *4th ACM International Conference on Web Search and Data Mining (WSDM 2011), Hong Kong, China*, pages 815–824, 2011.

[12] R. Kimball. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley, 1996.

[13] C. Lin, Y. He, R. Everson, and S. Ruger. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, 2012.

[14] T. Lust and A. Rolland. 2-additive Choquet optimal solutions in multiobjective optimization problems. In *15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014), Montpellier, France*, 2014.

[15] L. Oukid, O. Asfari, F. Bentayeb, N. Benblidia, and O. Boussaid. CXT-cube: contextual text cube model and aggregation operator for text OLAP. In *16th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2013), San Francisco, CA, USA*, pages 27–32, 2013.

[16] F. Rico, F. Muhlenbach, D. A. Zighed, and S. Lallich. Comparison of two Topological Approaches to deal with Noisy Labeling. *Neurocomputing*, 2015.

[17] R. Salem, O. Boussaid, and J. Darmont. An Active XML-based Framework for Integrating Complex Data. In *27th Annual ACM Symposium On Applied Computing (SAC 2012), Riva del Garda, Italy*, pages 888–892, 2012.

[18] R. Salem, J. Darmont, and O. Boussaid. Efficient Incremental Breadth-Depth XML Event Mining. In *15th International Database Engineering and Applications Symposium (IDEAS 2011), Lisbon, Portugal*, pages 197–203, 2011.

[19] D. A. Zighed, D. Ezzeddine, and F. Rico. Neighborhood Random Classification. In *Advances in Knowledge Discovery and Data Mining*, volume 7301 of *LNCS*, pages 98–108. Springer, 2012.