# A UML-EXTENDED APPROACH FOR MINING OLAP DATA CUBES IN COMPLEX KNOWLEDGE DISCOVERY ENVIRONMENTS

Alfredo Cuzzocrea
*ICAR-CNR and University of Calabria, Cosenza, Italy*

Keywords: Mining OLAP data cubes, UML, Complex knowledge discovery environments.

Abstract: In this paper, we propose theoretical assertions and practical instances of an innovative *UML-extended approach for mining OLAP data cubes in complex knowledge discovery environments*. This analytical contribution is further extended by means of a comprehensive set of case studies that clearly demonstrate the feasibility and the benefits of the proposed approach in the context of *next generation Data-Warehousing/Data-Mining platforms*.

## 1 INTRODUCTION

The problem of *mining OLAP data cubes* (e.g., Inmon, 1996; Han, 1997; Cuzzocrea, 2009) plays a leading role in *next-generation complex Data Mining systems and applications* such as *social networks*, *analytics over very large distributed repositories*, *stream and sensor data analysis*, and so forth. This maily because of data repositories stored, processed and mined in real-life systems and applications are ineherently *multidimensional, multi-level and multi-resolution in nature* (Cai et al., 2004; Han et al., 2005; Cuzzocrea et al., 2009) hence executing Data Mining algorithms over *multidimensional views* computed on top of original data sources leads to more expressive and powerful mining results (Inmon, 1996; Han, 1997; Rizzi et al, 2006; Cuzzocrea, 2007; Cuzzocrea, 2009).

The main idea of mining-OLAP-data-cube models and algorithms consists in formulating novel versions of traditional Data Mining approachs over flat data sources (e.g., relational databases) (Frawley et al., 1992; Fayyad et al., 1996) hence stirring-up towards innovative achievements that fully consider the multidimensionality and the multi-resolution of *data and schema cubes* (Gray et al., 1997).

This conceptual setting becomes more and more difficult in *complex knowledge discovery environments*, where knowledge discovery processes and routines are subjected to complex *constraints* and *external conditions*, and algorithms naturally become harder and harder as well. Relevant instances of such environments are stream and sensor data analysys tools, as mining stream and sensor data (Gaber at al., 2005) is a very challenging task due to a plethora of aspects ranging from *unbounded length* of streams to the need for *single-pass* (mining) algorithms, from *multi-rate arrivals* to *uncertainity and imprecision*, and so forth.

Another important aspect that adds complexity in mining OLAP data cubes is represented by the fact that, as recognized in (Zubcoff & Trujillo, 2006; Zubcoff & Trujillo, 2007; Zubcoff et al., 2007a; Zubcoff et al., 2007b), *Data Mining is still an artifact-like and solution-driven task*, which heavily depends on the particular application context. In practice, given a certain Data Mining problem, each data miner is likely to develop a proper Data Mining solution (with proper Data Mining models and algorithms), by referring-to and extending state-of-the-art results. Also, the solution usually turns to be very different from solutions proposed by other data miners. While this is reasonable, it has been recognized that *solution-driven approaches seriously limit the integration of software engineering paradigms* (Heineman & Councill, 2001), like conceptual modeling, project-sharing, re-engineering strategies, and so forth, *within Data Mining* (Zubcoff & Trujillo, 2006; Zubcoff & Trujillo, 2007; Zubcoff et al., 2007a; Zubcoff et al.,

Cuzzocrea A..
A UML-EXTENDED APPROACH FOR MINING OLAP DATA CUBES IN COMPLEX KNOWLEDGE DISCOVERY ENVIRONMENTS.
DOI: 10.5220/0003512302810289
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 281-289
ISBN: 978-989-8425-53-9
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

2007b). Furthermore, while this is critical in conventional Data Mining settings, it plays an even more challenging role in mining multidimensional structures like OLAP data cubes stored in Data Warehosue architectures (Zubcoff & Trujillo, 2006; Zubcoff & Trujillo, 2007; Zubcoff et al., 2007a; Zubcoff et al., 2007b), due to the complexities introduced by multidimensional data and schemas.

Starting from these considerations, in (Cuzzocrea et al., 2011) an innovative *model-driven Data Mining engineering framework* for *moving Data Mining models/algorithms from solution-oriented artifacts to "composable" Data Mining conceptual models* has been proposed. This framework allows us to enable more sophisticated theoretical settings leading to *software-engineering-inspired Data Mining approaches*, hence fully incorporating well-understood *software-component design paradigms* (Heineman & Councill, 2001).

A particularity of this framework is represented by the fact that it completely focuses on multidimensional data structures, due the above-mentioned benefits deriving from mining multidimensional cubes and views computed over the original data sets with respect to the more conventional, baseline case of mining flat data sources (Inmon, 1996; Han, 1997; Rizzi et al, 2006; Cuzzocrea, 2007; Cuzzocrea, 2009).

Basically, the main idea underlying the framework (Cuzzocrea et al., 2011) consists in modeling both the multidimensional data reportiories and the mining algorithms by means of innovative *UML profiles* that allow us to build *conceptual models* rather than solution-driven impelmtations. This allow analysts to concentrate on the abstract and conceptual level of the target Data Mining problem, rather than wasting time on low-level aspects. Notice that, in complex knowledge discovery environments, this feature plays a leading role, as scalability (even at a conceptual/modeling) becomes a challenge in such environments. Another important characteristic of the framework proposed in (Cuzzocrea et al., 2011) consists in the fact that a set of suitable *model transformations* is able of generating both the data under analysis (in an apposite Data Warehouse platform) and the analysis model (in a proper Data Mining platform). Indeed, it should be clear enough that this approach keeps several points of research innovation in the context of Data-Warehousing/Data-Mining research.

This paper significantly extends the research results provided in (Cuzzocrea et al., 2011), and provides a a comprehensive set of case studies focused on the application of well-understood and popular Data Mining techniques to a large corporate database of a hypothetical electric-supply company, where we apply the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011). The final goal is that of demonstrating and validating the feasibility of this approach throughout significant examples, while also clarifying theoretical and methodological aspects correlated to the research results provided in (Cuzzocrea et al., 2011). The Data-Mining/Analysis platform taken as reference for platform-dependent implementations generated by our model-driven Data Mining engineering approach (see Section 4) is *MS Analysis Services* (Microsoft Research, 2009b), which makes use of *Data Mining eXtensions* (DMX) (Microsoft Research, 2009a) as mining/modeling language/formalism (see Section 3).

With these goals in mind, the rest of the paper is structured as follows. In Section 2, we describe the (large) corporate database of the hypothetical electric-supply company, by focusing the attention on most-distinctive characteristics.

We then move the attention on the application of two well-understood and popular Data Mining techniques over the electric-supply corporate database via following the methodological guidelines drawn by the model-driven Data Mining engineering framework (Cuzzocrea et al., 2011). This also implies, as a secondary task, the definition of a collection of suitable OLAP data cube models that aggregate multidimensional data within apposite multidimensional data cubes able of providing support for integrated, cleaned and aggregated multidimensional views over which appropriate Data Mining techniques can be applied with success (more than the case for traditional database-oriented conceptual KDD reference architecture – (Inmon, 1996; Han, 1997; Rizzi et al, 2006; Cuzzocrea, 2007; Cuzzocrea, 2009)). For each of these Data Mining techniques, we provide the platform-dependent implementation over MS Analysis Services as generated by the approach (Cuzzocrea et al., 2011), plus analysis and discussion of (mining) results still in the same platform. The Data Mining techniques we consider in this paper are the following: *clustering* (Tan et al., 2005a; Tan et al., 2005b), which is presented in Section 3, and *decision trees* (Quinlan, 1986), which is presented in Section 4.

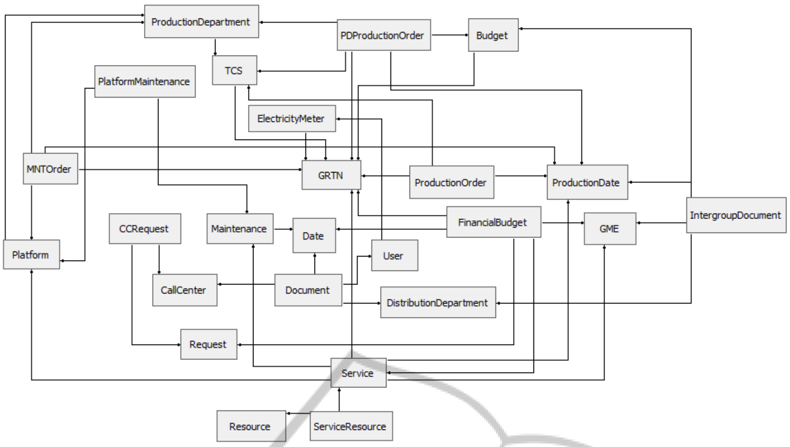


Figure 1: Schema of the electric-supply corporate database used as reference case study.

After that, in Section 5 we prove the *component-oriented capabilities* of the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011) via combining the so-obtained clustering conceptual mining model with the so-obtained decision-trees conceptual mining model into a novel clustering /decision-trees mining model, with the goal of enhancing the expressive power and the effectiveness of the overall Data Mining modeling process over the electric-supply corporate database.

Finally, in Section 6 we provide conclusions of the research presented in this paper and future research directions to be further investigated.

## 2 THE ELECTRIC-SUPPLY CORPORATE DATABASE

Figure 1 shows the schema of the electric-supply corporate database used as reference data source in the in-laboratory validation of the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011). The electric-supply corporate database stores all the information needed to support the activities of the electric-supply company, such as: (*i*) information needed to handle the main electric-supply network managed by the company; (*ii*) information needed to monitor and operate on the national electric-supply market; (*iii*) information needed to manage users of the company; (*iv*) information needed to manage the electric-supply distribution network across the country.

In the following, we provide an overview of the main tables populating the electric-supply corporate database:

1. *GRTN*: it stores data related to the handling of the main electric-supply network managed by the company. Most significant attributes are the following: (*i*) *GRTN_NodeID*, which models the absolute identifier of the actual electric-supply network node; (*ii*) *Sector*, which models the network sector to which the actual node belongs-to; (*iii*) *SharedCapital*, which models the amount of electric-supply power the actual node manages; (*iv*) *Address*, which models the address where the actual node is located.
2. *GME*: it stores data related to the monitoring and the operating on the national electric-supply market. Most significant attributes are the following: (*i*) *GME_ NodeID*, which models the absolute identifier of that node whit respect to which supply and demand of the electric-supply market are managed; (*ii*) *Sector*, which models the network sector to which the actual node belongs-to; (*iii*) *SharedCapital*, which models the amount of electric-supply power the actual node manages; (*iv*) *Address*, which models the address where the actual node is located.
3. *User*: it stores data related to the management of users of the company. Most significant attributes are the following: (*i*) *UserID*, which models the absolute identifier of the actual user of the electric-supply company; (*ii*) *Type*, which models the class of the actual user; (*iii*) *ElectricityMeter*, which models the absolute identifier of the electricity meter in usage by the actual user; (*iv*) *FirstName*, which models the first name of the actual user; (*v*) *LastName*, which models the first name of the actual user; (*vi*) *Address*, which models the address of the

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

actual user; (*vii*) *Phone*, which models the phone number of the actual user.

4. *DistributionDepartment*: it stores data related to the management of the electric-supply distribution network across the country. Most significant attributes are the following: (*i*) *DistDeptID*, which models the absolute identifier of the actual distribution department; (*ii*) *ProductionCapacity*, which models an indicator representing the production capacity of the actual department; (*iii*) *SharedCapital*, which models the amount of electric-supply power the actual department produces; (*iv*) *CallCenter*, which models the absolute identifier of the call center associated with the actual department.

## 3 THE CLUSTERING CONCEPTUAL DATA MINING MODEL *USERCLUSTERING*

The first case study of our analysis focuses on the application of a clustering technique (Tan et al., 2005a; Tan et al., 2005b) over user data of the hypothetical electric-supply company, which we name as *UserClustering*, whose final goal is that of clustering users based on their electric-supply purchases (represented in terms of billing documents). Figure 2 shows the conceptual Data Mining model of *UserClustering* shaped according to the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011). As shown in Figure 2, the data layer of *UserClustering* is captured by the three-dimensional OLAP data cube *Sales*, whereas *Expectation Maximization* (EM) (Dempster et al., 1977) has been adopted as specific clustering algorithm. Looking into detail, *Sales* is characterized by the fact *Document*, which models information about the billing documents, and by the following dimensions: (*i*) *User*, which models users of the electric-supply company; (*ii*) *CallCenter*, which models call centers of the electric-supply company; (*iii*) *Date*, which is a standard (OLAP) temporal dimension. For what instead regards the algorithmic parameters of the target EM clustering algorithm, the minimum support has been set as equal to 1 and the maximum number of generated clusters has been set as equal to 10, respectively. Attributes *Address* and *Type* of the dimension *User*, and attributes *Tax* and *Taxable* of the fact *Document*, respectively, have been set as input model parameters of the EM clustering algorithm, whereas *User* has been set as *case dimension* (i.e., a dimension containing case parameters – see Section 4.2) for *UserClustering*.

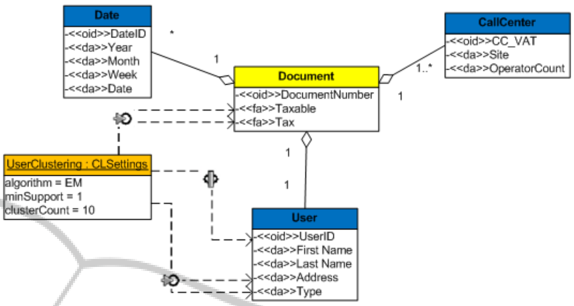


Figure 2: Conceptual Data Mining model of *UserClustering*.

We now move the attention on the platform-dependent implementations of *UserClustering* generated by the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011) on MS Analysis Services (Microsoft Research, 2009b). To this end, Figure 3 shows the implementation of the data cube *Sales*. Figure 4 shows the *mining structure* of *UserClustering* (i.e., the set of data and meta-data supporting the execution of *UserClustering* within the core layer of MS Analysis Services). Finally, Figure 5 shows the final platform-specific realization of *UserClustering* and Figure 6 shows the corresponding DMX-based implementation, respectively.

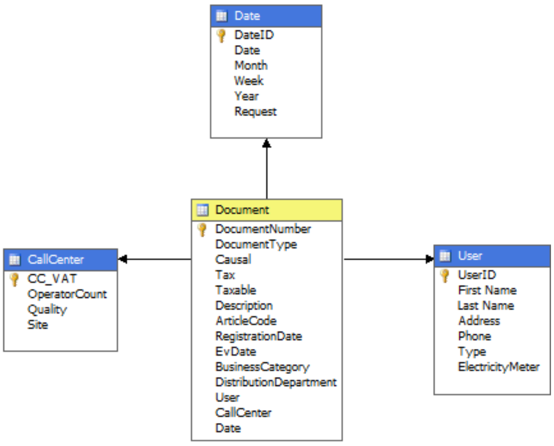
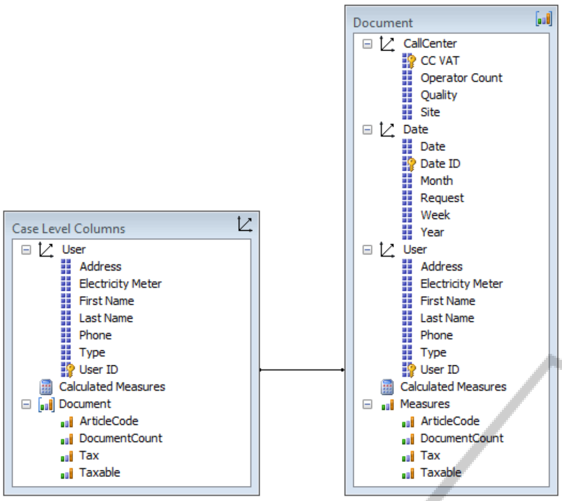


Figure 3: The data cube *Sales* of *UserClustering*.



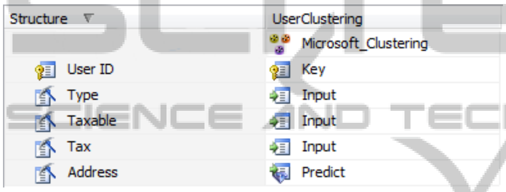
Case Level Columns

- User
  - Address
  - Electricity Meter
  - First Name
  - Last Name
  - Phone
  - Type
  - User ID
- Calculated Measures
- Document
  - ArticleCode
  - DocumentCount
  - Tax
  - Taxable

Document

- CallCenter
  - CC VAT
  - Operator Count
  - Quality
  - Site
- Date
  - Date
  - Date ID
  - Month
  - Request
  - Week
  - Year
- User
  - Address
  - Electricity Meter
  - First Name
  - Last Name
  - Phone
  - Type
  - User ID
- Calculated Measures
- Measures
  - ArticleCode
  - DocumentCount
  - Tax
  - Taxable

Figure 4: The mining structure of *UserClustering*.



| Structure | UserClustering |
|---|---|
| | Microsoft_Clustering |
| User ID | Key |
| Type | Input |
| Taxable | Input |
| Tax | Input |
| Address | Predict |

Figure 5: The final realization of *UserClustering*.

```
CREATE MINING MODEL
UserClustering{
  UserID text key,
  Type text discrete,
  Taxable long continuous,
  Tax long continuous,
  Address text discrete
} USING Microsoft_Clustering(
  MINIMUM_SUPPORT = 1, CLUSTER_COUNT = 10);
```

Figure 6: The DMX-based implementation of *UserClustering*.

As highlighted above, in our analysis we again make use of the underlying Data-Mining/Analysis platform (MS Analysis Services, in this case) to provide discussion of retrieved (mining) results (clusters, in this case). To this end, Figure 7 shows the results obtained from executing *UserClustering* over the target multidimensional data repository. Here, nodes represent clusters, arcs represent relations between clusters (e.g., *hierarchical Relations*), and darker clusters represent clusters grouping higher-in-cardinality user populations.

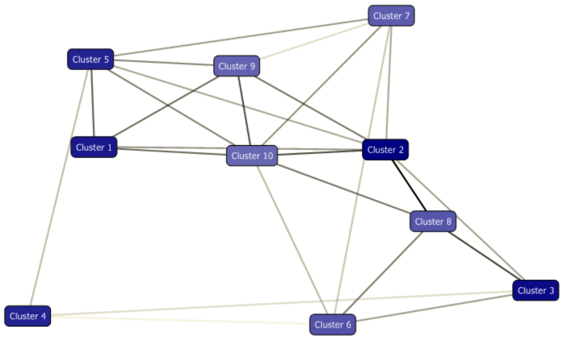


Figure 7: Mining results of *UserClustering*.

## 4 THE DECISION-TREES CONCEPTUAL DATA MINING MODEL *MAINTENANCECAUSAL*

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

The second case study of our analysis deals with the application of a decision-tree technique (Quinlan, 1986) over service data of the electric-supply company, named as *MaintenanceCausal*. *MaintenanceCausal* aims at discovering major causal of maintenance services performed in the main electric-supply network managed by the company. Figure 8 shows the conceptual Data Mining model of this second case study. As shown in Figure 8, the three-dimensional OLAP data cube *Planning* characterizes the data layer of *MaintenanceCausal*. In more detail, *Planning* introduces the fact *Service*, which models information about the services performed within the

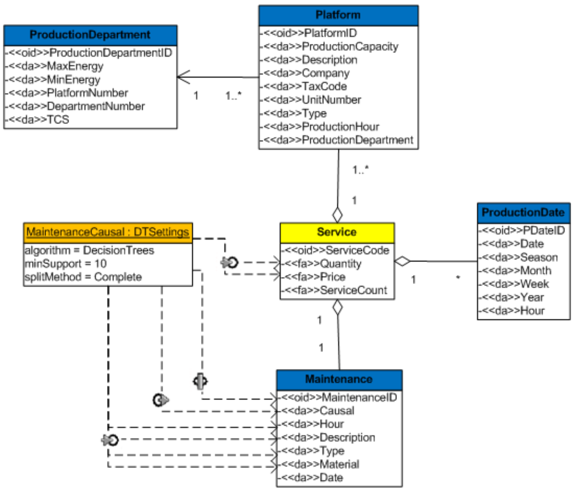


Figure 8: Conceptual Data Mining model of *MaintenanceCausal*.

electric-supply company, and the following dimensions: (*i*) *Maintenance*, which models information about specific maintenance operations of (maintenance) services; (*ii*) *Platform*, which models information about the platform where services are performed – *Platform* also comprises a normalized dimensional table, *ProductionDepartment*; (*iii*) *ProductionData*, which is a standard (OLAP) temporal dimension. For what instead regards the specific decision-trees algorithm, *MaintenanceCausal* makes use of the ad-hoc implementation provided by MS Analysis Services that consists of a *hybrid decision-tree algorithm* (Microsoft Research, 2009b) able of supporting both classification and regression functionalities during the mining phase. For this algorithm, the minimum support has been set as equal to 10 and the partitioning method has been chosen as the *binary* one, respectively. Furthermore, attributes *Quantity* and *Price* of the fact *Service*, and *Hour*, *Description*, *Type* and *Material* of the dimension *Maintenance*, respectively, have been set as input model parameters of the decision-trees algorithm, whereas attribute *Causal* of the dimension *Maintenance* has been set as output model parameter.

Figure 9 shows the platform-dependent implementation of the data cube *Planning*. Figure 10 shows instead the mining structure of *MaintenanceCausal*, whereas, to conclude, Figure 11 shows the final realization of *MaintenanceCausal* and Figure 12 shows the corresponding DMX-based implementation, respectively.

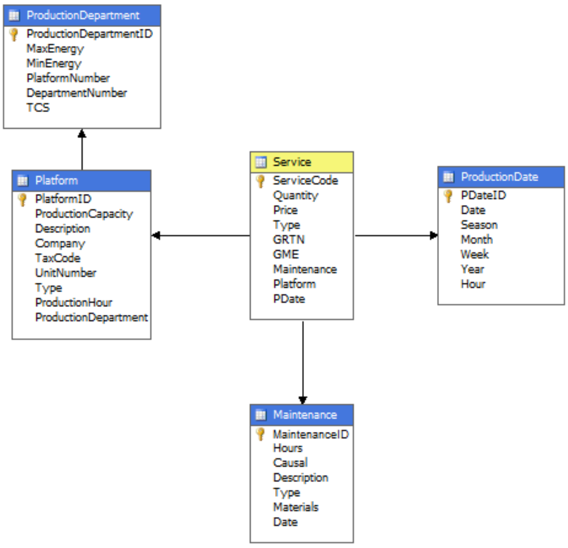
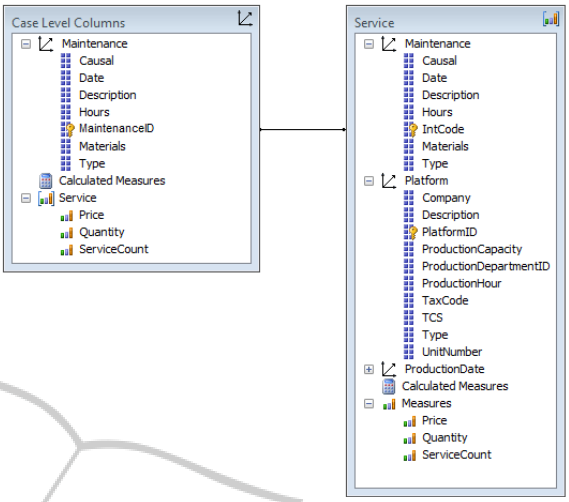


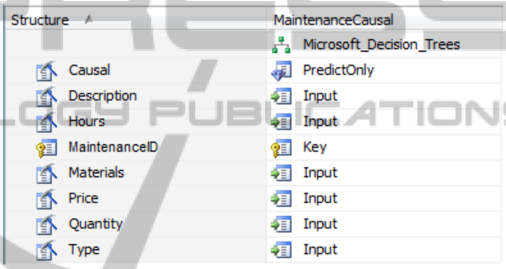
Figure 9: The data cube *Planning* of *MaintenanceCausal*.



Case Level Columns
- Maintenance
  - Causal
  - Date
  - Description
  - Hours
  - MaintenanceID
  - Materials
  - Type
- Calculated Measures
- Service
  - Price
  - Quantity
  - ServiceCount

Service
- Maintenance
  - Causal
  - Date
  - Description
  - Hours
  - IntCode
  - Materials
  - Type
- Platform
  - Company
  - Description
  - PlatformID
  - ProductionCapacity
  - ProductionDepartmentID
  - ProductionHour
  - TaxCode
  - TCS
  - Type
  - UnitNumber
- ProductionDate
- Calculated Measures
- Measures
  - Price
  - Quantity
  - ServiceCount

Figure 10: The mining structure of *MaintenanceCausal*.



| Structure | MaintenanceCausal |
|---|---|
| | Microsoft_Decision_Trees |
| Causal | PredictOnly |
| Description | Input |
| Hours | Input |
| MaintenanceID | Key |
| Materials | Input |
| Price | Input |
| Quantity | Input |
| Type | Input |

Figure 11: The final realization of *MaintenanceCausal*.

```
CREATE MINING MODEL
MaintenanceCausal{
  MaintenanceID long key,
  Causal text discrete predict_only,
  Description text discrete,
  Hours date continuous,
  Material text discrete,
  Price long continuous,
  Quantity long continuous,
  Type text discrete
} USING Microsoft_Decision_Trees(
  MINIMUM_SUPPORT = 10, SPLIT_METHOD = 2);
```

Figure 12: The DMX-based implementation of *MaintenanceCausal*.

Results obtained from executing *MaintenanceCausal* over the target multidimensional data repository are depicted in Figure 13. Here, darker nodes represent decision-tree nodes classifying higher counts of maintenance services, and the histogram stored by each decision-tree node captures the distribution of maintenance

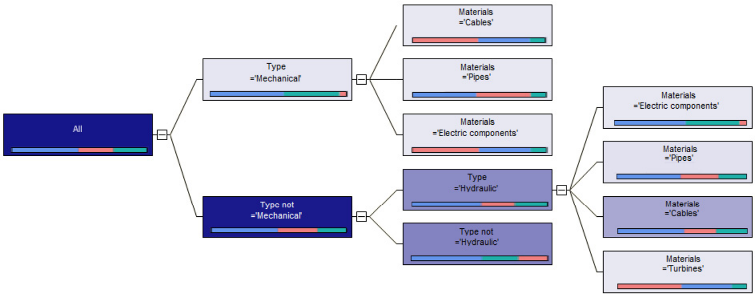


Figure 13: Mining results of *MaintenanceCausal*.

services with respect to the following causal: (i) "material breaches", represented by a blue bar; (ii) "technical assistance", represented by a red bar;, (iii) "ordinary assistance", represented by a green bar.

# 5 COMBINING CONCEPTUAL DATA MINING MODELS

A nice amenity supported by the model-driven Data Mining engineering approach (Cuzzocrea et al., 2011) consists in the fact that obtained conceptual Data Mining models can be further combined into a novel model, according to widely-understood software-component design paradigms (Heineman & Councill, 2001). This with the goal of enhancing the expressive power of the overall Data Mining modeling process. In line with this leading aspect, in the third (and last) case study we combine the clustering conceptual Data Mining model *UserClustering* and the decision-trees conceptual Data Mining model *MaintenanceCausal*, which have been presented and discussed in Section 3 and Section 4, respectively. This allows us to obtain a novel clustering/decision-trees mining model, named as *UserClusteredMaintenanceCausal*, which is depicted in Figure 14. *UserClusteredMaintenanceCausal* is a complex conceptual Data Mining model whose main goal consists in discovering major causal of maintenance services performed in the main electric-supply network managed by the company (which is the specific goal of *MaintenanceCausal* – see Section 4) by grouping per clusters of users generated on the basis of users' electric-supply purchases (which is the specific goal of *UserClustering* – see Section 3). It should be noted that the Data Mining analysis supported by *UserClusteredMaintenanceCausal* turns to be extremely useful for decision makers of the hypothetical electric-supply company, as, based on results derived from executing *UserClusteredMaintenanceCausal* over suitable repositories of multidimensional data, decision makers can determine electric-supply distribution policies different from the actual ones if clear relations among users' electric-supply purchases and maintenance service causal are discovered (e.g., peaks of users' electric-supply purchases may frequently involve in high numbers of request for technical assistance).

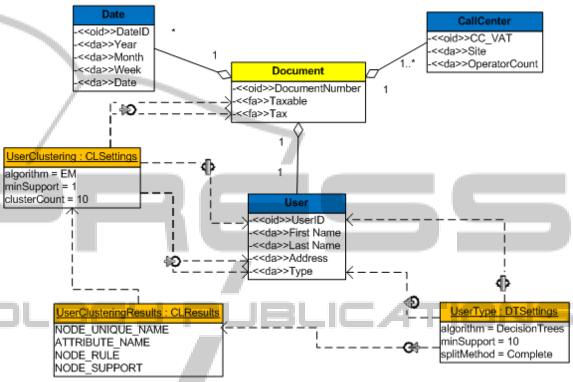
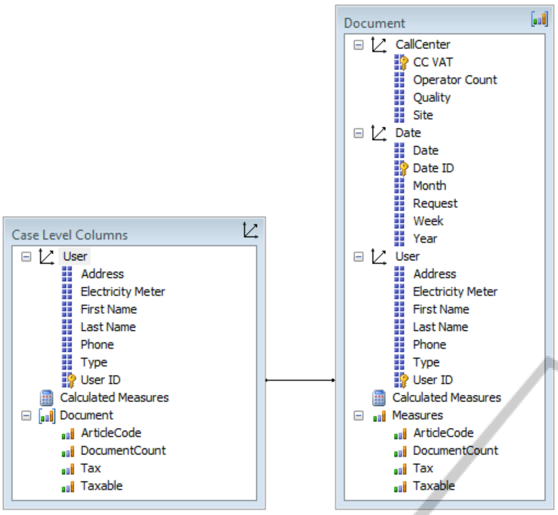


Figure 14: Conceptual Data Mining model of *UserClusteredMaintenanceCausal*.

As shown in Figure 14, in *UserClusteredMaintenanceCausal* the output provided by *UserClustering* (i.e., user clusters) is given as input to *MaintenanceCausal*, and the final result is still modeled in terms of a conventional decision tree, like in *MaintenanceCausal*. In particular, *UserClustering* is used like a sort of novel "dimension" of the three-dimensional OLAP data cube *Sales*, which originally was part of *UserClustering* (see Section 3), and the artificial attribute *NODE_UNIQUE_NAME*, which models a node-based representation of user clusters obtained from *UserClustering* (like the one shown in Figure 7), has been set as input model parameter for *MaintenanceCausal*.

Figure 15 shows the mining structure of *UserClusteredMaintenanceCausal* (recall that the

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS



Case Level Columns

- User
  - Address
  - Electricity Meter
  - First Name
  - Last Name
  - Phone
  - Type
  - User ID
- Calculated Measures
- Document
  - ArticleCode
  - DocumentCount
  - Tax
  - Taxable

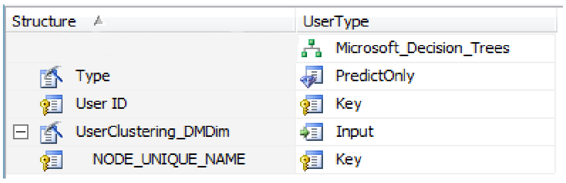
Document

- CallCenter
  - CC VAT
  - Operator Count
  - Quality
  - Site
- Date
  - Date
  - Date ID
  - Month
  - Request
  - Week
  - Year
- User
  - Address
  - Electricity Meter
  - First Name
  - Last Name
  - Phone
  - Type
  - User ID
- Calculated Measures
- Measures
  - ArticleCode
  - DocumentCount
  - Tax
  - Taxable

Figure 15: The mining structure of *UserClusteredMaintenanceCausal*.

implementation of *Sales* if the same of the one provided for *UserClustering*, which is depicted in Figure 3). Furthermore, Figure 16 shows the final realization of *UserClusteredMaintenanceCausal* and Figure 17 shows the corresponding DMX-based implementation, respectively.

Results obtained from executing *UserClusteredMaintenanceCausal* over the target multidimensional data repository are shown in Figure 18 (due to space reasons, only the first four *UserClustering* (see Figure 13), darker nodes represent decision-tree nodes classifying higher levels of the output decision tree are shown). Here, like for the case of the mining results of counts of maintenance services grouped by user clusters, and the histogram stored by each decision- tree node captures the distribution of maintenance services with respect to the following user clusters: (*i*) “home users”, represented by a blue bar; (*ii*) “partner users”, represented by a red bar; (*iii*) “Business users”, represented by a green bar.



| Structure | UserType |
|---|---|
| | Microsoft_Decision_Trees |
| Type | PredictOnly |
| User ID | Key |
| UserClustering_DMDim | Input |
| NODE_UNIQUE_NAME | Key |

Figure 16: The final realization of *UserClusteredMaintenanceCausal*.

```
CREATE MINING MODEL
UserType{
  UserID long key,
  Type text discrete predict_only,
  UserClustering_DMDim table (
    NODE_UNIQUE_NAME text key)
} USING Microsoft_Decision_Trees(
  MINIMUM_SUPPORT = 10, SPLIT_METHOD = 2);
```

Figure 17: The DMX-based implementation of *UserClusteredMaintenanceCausal*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

## 6 CONCLUSIONS AND FUTURE WORK

Starting from the research results provided in (Cuzzocrea et al., 2011), where an innovative model-driven Data Mining engineering framework that focuses on multidimensional data structures has been proposed, in this paper we have further and significantly extended these results by proposing a comprehensive set of case studies focused on the application of well-understood and popular Data Mining techniques to a large corporate database of a hypothetical electric-supply company. These case studies are meant to prove some relevant characteristics of the framework (Cuzzocrea et al., 2011), among which the suitability of this framework in providing conceptual Data Mining the

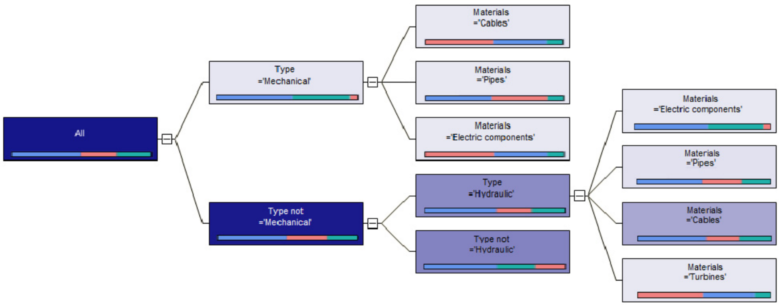


Figure 18: Mining results of *UserClusteredMaintenanceCausal*.

nice amenity of making these model "composable" in order to achieve the definition of models for Data Mining problems arising in complex knowledge discovery environments, and complex Data Mining models starting from simpler ones play the major roles.

Future work of this research is oriented towards integrating within the framework (Cuzzocrea et al., 2011) innovative aspects as to capture advanced features of Data-Warehouse/Data-Mining platforms, such as *security and privacy*, and *uncertainty and imprecision*.

## REFERENCES

Cai, Y. D., Clutterx, D., Papex, G., Han, J., Welgex, M., and Auvilx, L. (2004) 'MAIDS: mining alarming incidents from data streams', in *Proceedings of the 2004 ACM International Conference on Management of Data*, pages 919–920.

Cuzzocrea, A. (2007) 'An OLAM-based framework for complex knowledge pattern discovery in distributed-and-heterogeneous-data-sources and cooperative information systems', in *Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery, LNCS Vol. 4654*, pages 181–198.

Cuzzocrea, A. (2009) 'OLAP intelligence: meaningfully coupling OLAP and data mining tools and algorithms', *International Journal of Business Intelligence and Data Mining*, 4(3-4): 213–218.

Cuzzocrea, A. Furfaro, F., Masciari, E., and Saccà D. (2009) 'Improving OLAP analysis of multidimensional data streams via efficient compression techniques', in *A. Cuzzocrea (ed.), "Intelligent Techniques for Warehousing and Mining Sensor Network Data", IGI Global*, 17–49.

Cuzzocrea, A., Mazon, J.-N., Trujillo J., and Zubcoff, J. (2011) 'Model-driven data mining engineering: from solution-driven implementations to "composable" conceptual data mining models", *International Journal of Data Mining, Modelling and Management*, to appear.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1992) 'Knowledge discovery in databases: an overview', *AI Magazine*, 13(3): 213–228.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Widener, T. (1996) 'The KDD process for extracting useful knowledge from volumes of data', *Communications of the ACM*, 39(11):27–34.

Gaber, M. M., Zaslavsky, A. B., and Krishnaswamy, S. (2005) 'Mining data streams: a review', *SIGMOD Record*, 34(2): 18–26.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997) 'Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals', *Data Mining and Knowledge Discovery*, 1(1):29–53.

Han, J. (1997) 'OLAP mining: an integration of OLAP with data mining', in *Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics*, pages 1–9.

Han, J., Chen, Y., Dong, G., Pei, J., Wah, B. W., Wang, J., and Cai, Y. D. (2005) 'Stream cube: an architecture for multi-dimensional analysis of data streams', *Distributed and Parallel Databases*, 18(2): 173–197.

Heineman, G. T., and Councill, W. T. (2001) *Component-Based Software Engineering: Putting the Pieces Together*, Addison-Wesley Professional, Reading, MA, USA.

Inmon, W. H. (1996) 'The data warehouse and data mining', *Communications of the ACM*, 49(4)83–88.

Microsoft Research (2009a) *Data Mining eXtensions (DMX) Reference*, http://msdn.microsoft.com/en-us/library/ ms132058.aspx

Microsoft Research (2009b) *SQL Server Analysis Services – Data Mining*, http://msdn.microsoft.com/en-us/library/ bb510517.aspxQuinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, 1(1):81–106.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005a) *Introduction to data mining – Chapter 8: Cluster analysis: basic concepts and algorithms*, Addison-Wesley, Reading, MA, USA.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005b) *Introduction to data mining – Chapter 9: Cluster analysis: additional issues and algorithms*, Addison-Wesley, Reading, MA, USA.

Zubcoff, J., and Trujillo, J. (2006) 'Conceptual modeling for classification mining in data warehouses', in *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery, LNCS Vol. 4081*, pages 566–575.

Zubcoff, J., and Trujillo, J. (2007) 'A UML 2.0 profile to design association rule mining models in the multidimensional conceptual modeling of data warehouses', *Data & Knowledge Engineering*, 63(1):44–62.

Zubcoff, J., Pardillo, J., and Trujillo, J. (2007a) 'Integrating clustering data mining into the multidimensional modeling of data warehouses with UML profiles', in *Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery, LNCS Vol. 4654*, pages 199–208.

Zubcoff, J., Trujillo, J., and Cuzzocrea, A. (2007b) 'On the suitability of time series analysis on data warehouses', in *Proceedings of the 1st IADIS European Conference on Data Mining*, pages 17–24.