Journal of Mobile Multimedia, , Vol. 9, No.3&4 (2014) 214-229

GOAL DRIVEN APPROACH TO MODEL INTERACTION BETWEEN VIEWPOINTS OF A MULTI-VIEW KDD PROCESS

EL MOUKHTAR ZEMMOURI (1), HICHAM BEHJA (1,2), BRAHIM OUHBI (1), BRIGITTE TROUSSE (2), ABDELAZIZ MARZAK (3), YOUSSEF BENGHABRIT (1)

(1) Ecole Nationale Supérieure d'Arts et Métiers – Meknes, Morocco
e.zemmouri@ensam-umi.ac.ma, h_behja@yahoo.com
ouhbib@yahoo.co.uk, you_benghabrit@yahoo.fr

(2) Equipe-projet AxIS, INRIA Sophia Antipolis - Méditerranée, France
brigitte.trousse@inria.fr

(3) Faculté des Sciences Ben Msik, Casablanca, Morocco
marzak@hotmail.com

Knowledge Discovery in Databases (KDD) is a highly complex, iterative and interactive process, with a goal-driven and domain dependent nature. The complexity of KDD is mainly due to the nature of the analyzed data (which are massive, distributed, incomplete, and heterogeneous) and the nature of the process itself (since the process is by definition interactive and iterative). Given this complexity, a KDD user faces two major challenges: on the one hand, he must manipulate prior domain knowledge to better understand data and business objectives. On the other hand, he must be able to choose, configure, compose and execute tools and methods from various fields (e.g., machine learning, statistics, artificial intelligence, databases) to achieve goals. Furthermore, in the business real world, a data mining project is usually held by several actors (domain experts, data analysts, KDD experts …), each with a different viewpoint. In this paper we propose to tackle the complexity of KDD process, and to enhance coordination and knowledge sharing between actors of a multi-view KDD analysis through a goal driven modeling of interactions between viewpoints. After a brief review of our approach of viewpoint in KDD, we will first develop a semantic Model of Goals that allows identification and representation of business objectives during the business understanding step of KDD process. Then, based on this Goal Model, we define a set of semantic relations between viewpoints of a multi-view analysis; namely equivalence, inclusion, conflict and requirement.

Key words: KDD process, Viewpoint, Goal analysis, Ontologies, SWRL

1 Introduction

In recent decades, enterprises' information systems become more and more flooded by all kind of data: structured (databases, data warehouse), semi-structured (XML, server log files), and unstructured data (raw text, multimedia data). This has created new challenges for companies and for the scientific community. Including, how to understand and analyze such a mass of data to extract knowledge. Hence, KDD (Knowledge Discovery in Databases [1]) have rapidly changed from a research area into an industry standard (i.e. CRISP-DM: Cross-Industry Standard Process for Data Mining [2]). In [3], Kurgan and Musilek draw up a comprehensive survey of knowledge discovery and data mining standard process models.

But, available data mining tools (commercial as well as free tools) support end-users only with graphical and/or manual construction of KDD execution plans. This is done without taking into account analyst viewpoint and coordination within organization (coordination between domain experts, data analysts, and KDD experts). But this is time consuming, cost sensitive, and requires good experience and prior knowledge of data mining domain to successfully conduct a KDD project.

Our objective in this paper is to model and formalize interactions between different viewpoints of a multi-view KDD process (which we define as a KDD process held by several experts who analyze the same data with different viewpoints). This is done by introducing a set of semantic relations between viewpoints. Our formalization is based on a goal driven approach and will allow us to enhance coordination, knowledge sharing and mutual understanding between different actors of a multi-view analysis, and reusability in terms of viewpoint of successful data mining experiences within an organization.

In the reminder of this introductory section we briefly recall our approach of viewpoint in KDD to fix definitions and context of our work. This approach was initiated by Behja et al. [4] [5] as part of Behja's doctoral thesis performed at Inria Sophia Antipolis – Méditerranée (AxIS team) in the context of France-Morocco cooperation (Software Engineering Network)[a] which was motivated by previous works on viewpoint management in cooperative design [6].

1.1. Viewpoint in KDD

A multi-view KDD process is usually held by one or more users who consequently manipulate several types of knowledge and know-how. They will have different objectives and preferences, different competences, and different visions of analyzed data, KDD methods and functions. In brief, they have different viewpoints. In this context, the KDD process will be guided by the analyst viewpoint [4] and several types of knowledge and expertise are incorporated (that is analyzed domain knowledge and analyst domain knowledge).

Figure 1 shows an example of a multi-view analysis of data from an e-learning system (that consists especially of HTTP log files, platform database, and courses material). These data can be analyzed by different actors of the system (teachers, learners, administrator, marketing, KDD expert, data miner). The objective of a teacher (e.g. improve quality of a course) is not the same as the administrator's one (e.g. ensure system reliability). Attributes used for evaluating a course are different from those used for studying the reliability. Similarly, chosen data mining methods, techniques and tools will be different, and interpretation/deployment of data mining results depends on the analyst's viewpoint. Therefore, it is fundamental to take into account the viewpoint of each analyst and interaction between different viewpoints.

[a] http://raweb.inria.fr/rapportsactivite/RA2003/axis/uid81.html

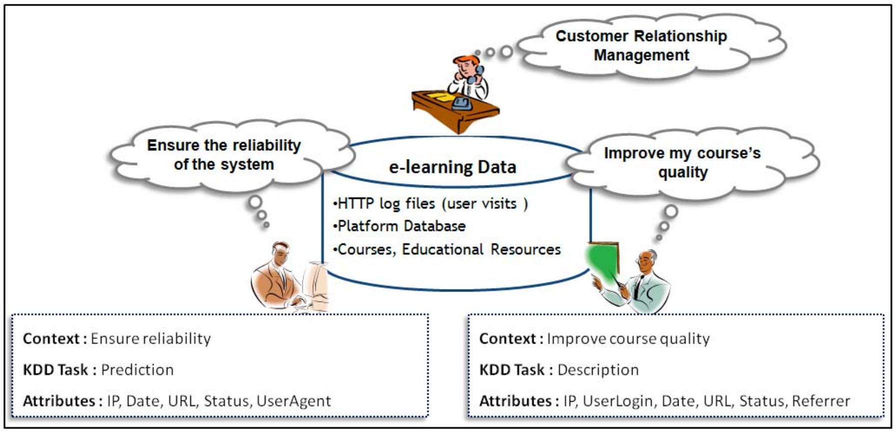

Figure 1 Multitude of viewpoints to analyze data from an e-learning system [7].

1.2. Knowledge Model for Multi-View KDD Process

In previous works we have developed a Knowledge Model for multi-view KDD process [5] which is "*a specification of the information and knowledge structures and functions involved [during a multi-view analysis]*" [8]. As shown in figure 2, our Knowledge Model integrating the viewpoint notion consists of four hierarchical sub-models structured in domain knowledge and strategic knowledge. Domain level describes the domain concepts and their relationships. The strategic level is based on the domain level and expresses how a task will be achieved.

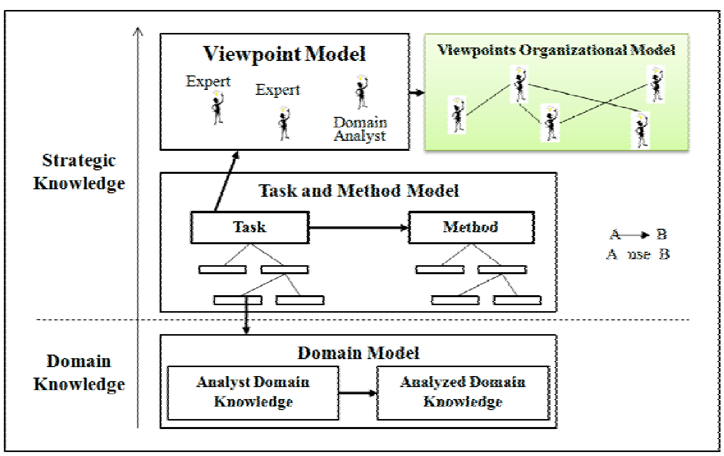

Figure 2 Knowledge Model for a multi-view KDD process [5].

The *Domain Model* concerns the studied domain concepts and the various relations between them. It describes the analyzed domain knowledge in terms of manipulated data. In our applicative context, that contains an e-learning ontology, platform database and HTTP log file structure.

The *Task and Method Model* describes the KDD process in terms of tasks and methods. Tasks are performed by methods. A task is a description of what must be done in the application in terms of goals and sub-goals. The methods describe how a goal can be achieved in terms of a series of operations and an order of execution. In [9], we have formalized this model as a generic semi-formal ontology OntoECD (Ontology for KDD process and data mining).

The *Viewpoint Model* is a conceptualization of a set of generic criteria that characterize a viewpoint in KDD [7]. These generic criteria were identified based on CRISP-DM reference model and formalized as a viewpoint ontology in OWL (Web Ontology Language). They allow modeling the vision of the analyst on manipulated data, the objective of analysis, and part of the expertise required for decisions made during the analysis.

In this paper we will mainly focus on the development and formalization of the *Viewpoints Organizational Model*. This model emphasizes the interaction and interdependence between various analyses according to different viewpoints

The remainder of the paper is organized as follows. In section 2 we will give a state of the art of works related to supporting users of KDD and managing knowledge in such process especially using ontologies. In section 3, we will develop a semantic model of goals that represents business objectives of different actors of a multi-view analysis, and formalize two types of relations between goals: AND/OR decomposition, and positive/negative influence. We will also illustrate our goal model within an example in e-learning. Based on this goal model, in section 4, we will define a set of semantic relations between viewpoints in KDD, especially: equivalence, conflict, and requirement. As stated before, the main purpose of these relations is to provide a means of coordination and mutual understanding between viewpoints' actors. In section 6, a case tool implementation to validate our approach is presented. Finally, we will draw some learned lessons in conclusion.

2 Related Works

"*An analyst is usually not a KDD expert but a person responsible for making sense of the data using available KDD techniques. Since the KDD process is by definition interactive and iterative, it is a challenge to provide a high-performance, rapid-response environment that also assists users in the proper selection and matching of appropriate tools and techniques to achieve their goals*" [1]. This was a statement of Fayyad et al. in 1996 highlighting the challenge of supporting both expert and novice users of KDD due to its complexity. We notice also that during a KDD process (a data mining project according to CRISP-DM vocabulary) there will be several actors (domain experts, data analysts, data miners …). The question we address in this paper is how to provide a framework allowing these actors to collaborate, share knowledge (about business domain and data mining techniques), and reuse each others' experiences.

Several works and standards have addressed the complexity of KDD with different approaches with the aim of supporting both expert and novice users. Most of these approaches are based on ontologies (either domain ontologies, or data mining ontologies). They offer the user the advantage to explore the large space of valid data mining processes [10][11], to discover and access distributed data mining services [12][11], to reuse successful data mining experiences [13], etc. But without taking into account the multi-view aspect of a KDD analysis as proposed by Behja et al. [4] for annotating a multi-view KDD process in order to support the reuse of KDD processes.

One of the first ontologies proposed to support users of KDD is DAMON (Data Mining ONtology) [14] that is designed to simplify the development of distributed KDD applications on Grids. DAMON ontology concerns only the data mining phase of a KDD process, and offers a taxonomy for discovering tasks, methods and tools deemed more suitable for a given data mining goal.

In MiningMart project [13] a case-based reasoning (CBR) system to support end users during data preprocessing is proposed. This system is based on a meta-model (called M4) of KDD preprocessing chains that contains ontology for describing conceptual domain knowledge. In the same project, Euler [12] proposes a web-based platform (which is a case base containing MiningMart successful experiences) to publicly display preprocessing models in a structured way, together with descriptions about their business domains, goals, methods and results.

Bernstein et al., [10] propose an Intelligent Discovery Assistant (IDA) for valid data mining processes enumeration and ranking. IDA focuses mainly on preprocessing and data mining phases of the KDD process. It is based on a formal ontology that contains input/output, preconditions constraint, and performance (accuracy, complexity, and comprehensibility) of each data mining operator. This ontology allows selection and composition of data mining operators suitable for user's data and goal. The systematic enumeration and ranking of valid data mining processes is based on IA planning techniques. A similar approach is proposed by Diamantini et al., [11] in a project called KDDVM (KDD Virtual Mart), which is a web services based system that aims to support users in the design of valid KDD process. It represents KDD operations as services which can be "annotated, introduced, accessed, described, composed and activated". KDDVM is based on KDDONTO ontology and concerns only data preprocessing and data mining steps.

A recent European project (e-LICO[b]: e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science) deals with the problem of supporting users of KDD in a collaborative way [15]. One of the products of this project is eProPlan [16], an ontology based environment for planning KDD workflows. It is based on two ontologies DMWF-DMOP and uses IA planning techniques to automatically generate KDD execution plan for solving data mining problems. DMWF (Data Mining Work Flow Ontology) formalizes IO-objects, operators, goals, tasks and methods as well as the decomposition of tasks into methods and operators (this ontology is equivalent to our OntoECD ontology described in [9]). DMOP (Data Mining Optimization Ontology) provides a unified conceptual framework for analyzing data mining tasks, algorithms, models, datasets, workflows and performance metrics, and their relationships.

Our approach of supporting KDD users focuses on the reusability, coordination, and knowledge sharing between multi-users of a KDD process, rather than the automatic generation of KDD execution plans. In addition, we cover the whole phases of the KDD process (cf. 6 phases suggested by CRISP-DM reference model [2]).

In previous works we have addressed mainly three issues: integration of domain knowledge [7], annotation of KDD in terms of viewpoint [17], and viewpoint characterization and modeling [7].

[b] e-LICO Project, http://www.e-lico.eu/

In this paper we propose to focus on modeling interactions between viewpoints of multi-view KDD process. For this, our proposed model will be based on a goal driven approach [18].

3 Goal Model for Multi-View KDD Process

3.1 Goal-Driven Analysis

Goal driven analysis consists of the identification, modeling and analysis of goals and objectives that guide decisions and strategies of an organization at different levels. Goals are also defined to be a logical mechanism for identifying, organizing and justifying requirements [18].

Goal oriented modeling approaches are not new in the field of information systems engineering, they date from the early 90s [19]. The Goal oriented modeling has been practiced complementarily to object oriented modeling, the former focusing on the early stages of requirements analysis and on the rationalization of the development process, the latter on late stages of requirements analysis [20].

Goal models have been used in Software Engineering [19] [21] and Data Warehousing [22] [23] [24] in order to represent users requirements, business objectives and design qualities. Indeed, eliciting requirements (as high level goals) early in the development process is crucial [21]. Requirements can be functional or non-functional. Non-functional requirements (also known as quality requirements) are defined as attributes or constraints of the system such as performance, security, and reliability [25].

Goal models are recognized also to be useful for knowledge management systems which focus on strategic knowledge representation and reasoning [26][27]. In this case, goals are used to represent strategic objectives of an organization and to analyze and keep track of events and trends that may influence positively or negatively these objectives.

3.2. Generic Model of Goals for Multi-View KDD Process

In the context of multi-view KDD process we propose to use the goal-driven approach to identify and represent business objectives (this is recognized to be the first task of the first step of CRISP-DM reference model) of a multi-view KDD process and to model interactions between viewpoints of stakeholders of these objectives. This will provide users with a methodological assistance during the business understanding step of KDD and make business objectives persistent for the other steps (especially evaluation and deployment of data mining results). In fact, as is the case for requirements traceability and persistence in data warehousing [23] and software development [28], we support here the necessity of storing early defined business objectives. Their presence can guide many decisions and solve many problems encountered at different stages of a KDD analysis. The persistence of goals, in our approach, is intended also to be a complement of metadata schema introduced by Behja et al. [17][5] to annotate and keep track of multi-view analysis.

Our proposed generic model of goals for multi-view KDD process, represented as OWL ontology, is depicted in figure 3. This model will be detailed and instantiated through an example in the reset of this section.

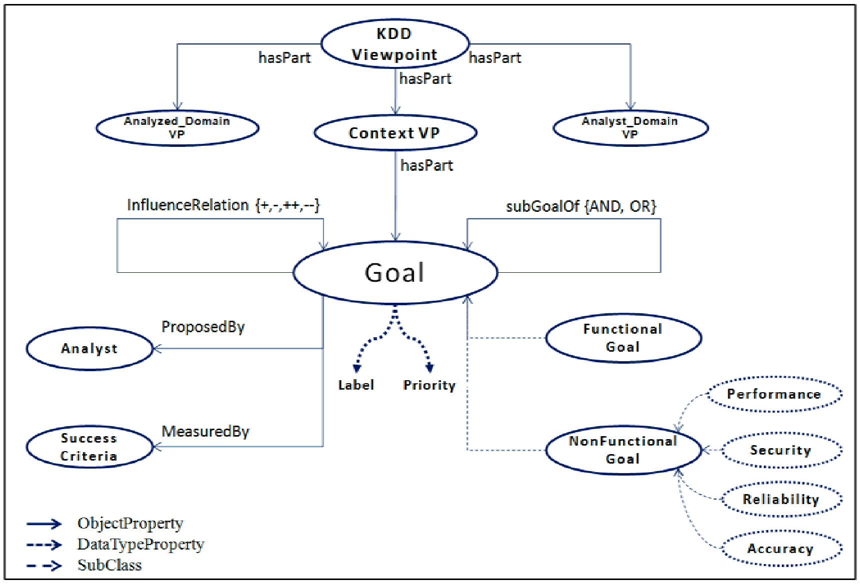

Figure 3 Generic Model of Goals for multi-view KDD process.

During a data mining project, an analyst (actor of the system) may have one or more goals (these goals are considered as a part of the context associated with his/her viewpoint [7]). A goal can be either functional or non-functional. Functional goals reflect business objectives, while non-functional goals reflect constraints of the system (such as security, reliability, and performance) or technical constraints on the execution of KDD methods (such as accuracy, cost sensitivity, comprehensibility, runtime, number of techniques that form the execution plan, etc.). As an example: "*Improve quality of a course*" is a functional goal, "*ensure reliability of the e-learning platform*" is a non-functional one.

A goal may have two attributes (Data Type Property in OWL vocabulary) priority and label. Label indicates if a goal is satisfied or denied [26]. Priority indicates if a goal is mandatory or optional [22].

A goal is proposed by an actor (KDD expert, domain expert, and data miner are example of actors) and is measured by some success criteria.

Thereby, a goal is characterized as:

Goal **<***Objective, Actor, Result, Functional / non Functional***>**

For example, "*ensure reliability of the e-learning platform*" is a non-functional goal, having the administrator of the platform as actor and can be measured by the probability to detect and block an http attack.

A Goal Model is represented by a directed graph G(E,V). Where E is a set of goals identified to represent operational objectives of different actors in a multi-view analysis, and V is a set of semantic relationships between the different goals of E. Figure 4 shows a simple example of such representation:

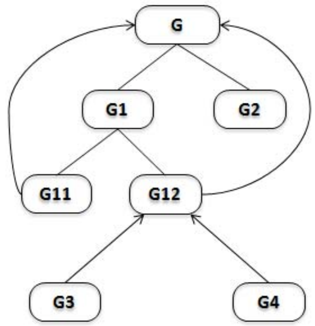

Figure 4 Graph representation of a Goal Model

In goal-driven approach, two types of reflexive relations (Object Property in OWL vocabulary) between goals are defined (especially for reasoning purposes) [26]: *sub-goal relation* and *influence relation*.

Sub-goal relation: consists of the decomposition of a general goal into more specific sub-goals. For this relation two main sub-relations are used: AND-decomposition and OR-decomposition. Definitions and semantic of these relations are as follow:

- **If** *a goal G is AND-decomposed into sub-goals* $G_1, G_2, \ldots, G_n$,

  Then *if all of the sub-goals are satisfied so is goal G.*

- **If** *a goal G is OR-decomposed into sub-goals* $G_1, G_2, \ldots, G_n$,

  Then *if at least one of the sub-goals is satisfied so is goal G.*

To formalize these relations, in the case of binary decomposition, we define the following SWRL[c] rules:

- AND decomposition:

```
satisfied(?G1) ∧ satisfied(?G2) →  satisfied(?G)
denied(?G1) →  denied(?G)
denied(?G2) →  denied(?G)
```

- OR decomposition:

```
satisfied(?G1) →  satisfied(?G)
satisfied(?G2) →  satisfied(?G)
denied(?G1) ∧ denied(?G2) →  denied(?G)
```

Notice that binary decomposition is not a restriction since it can be easily generalized into n-ary decomposition.

[c] SWRL: A Semantic Web Rule Language Combining OWL and RuleML, http://www.w3.org/Submission/SWRL/

Influence relation: this type of relation models situations where a goal contributes positively or negatively towards the satisfaction/denial of another goal. It is described to be more qualitative relation than AND, OR-decomposition [26].

Influence relation has 4 sub-relations that are labeled (+, -, ++, and --) depending on whether influence is positive or negative, full or partial. The definitions and semantic of these relations are as follow:

- +(G, G') : **if** *G is satisfied* **then** *G' is partially satisfied*.
- -(G, G') : **if** *G is satisfied* **then** *G' is partially denied*.
- ++(G, G') : **if** *G is satisfied* **then** *G' is fully satisfied*.
- --(G, G') : **if** *G is satisfied* **then** *G' is fully denied*.

To formalize these relations, we define the following SWRL rules:

```
posInfluence(?x,?y) ∧ satisfied(?x) → partSatisfied(?y)
negInfluence(?x,?y) ∧ satisfied(?x) → partDenied(?y)
posPosInfluence(?x,?y) ∧ satisfied(?x) → fullSatisfied(?y)
negNegInfluence(?x,?y) ∧ satisfied(?x) → fullDenied(?y)
```

Where (posInfluence, negInfluence, posPosInfluence, and negNegInfluence) represent respectively influence relations (+, -, ++, and --) in OWL.

In the case where there is no influence between two goals, we note an *unknown* (or *undetermined*) influence. The following figure shows the representation of sub-goal relation and influence relation on the graph of goals:

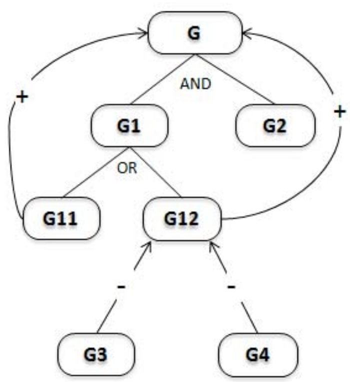

Figure 5 AND/OR-decomposition and positive/negative influence representation

3.3. An Example

In this subsection we will illustrate the goal model through an example in our applicative context e-learning. This example will be considered as an instance of the generic model of goals presented in the previous sub-section.

Assume that we want to conduct a data mining project to analyze data from a web based e-learning platform. According to CRISP-DM methodology [29], the first and critical step of the project is *business understanding*, where business objectives must be identified, analyzed and then transformed into data mining tasks. In our case study, examples of such business objectives are: "*improve course quality*" form the **teacher viewpoint**, "*manage users' relationship*" from the **marketing viewpoint**, and "*ensure reliability of the platform*" from the **admin viewpoint** (as depicted in figure 1).

These business objectives are modeled using goals and goal relations (AND/OR decomposition, +/- influence) as illustrated in figure 6 below.

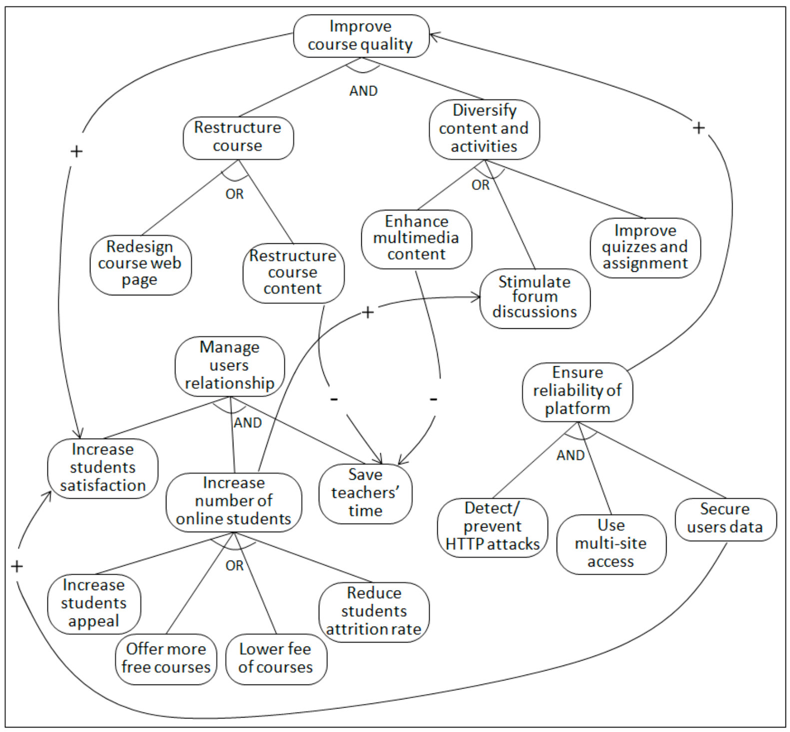

Figure 6 Example of a goal model for analyzing a web based e-learning system according to tree different viewpoints.

The teacher goal "*improve course quality*" is AND-decomposed to "*restructure the course*" and "*diversify content and activities*", and is positively influenced by the admin goal "*ensure reliability of the platform*" (i.e., satisfaction of the latter contribute positively towards the satisfaction of the former). The admin goal "*ensure reliability of the platform*" is AND-decomposed to "*secure users' data*", "*use multi-site access*", and "*detect and prevent HTTP attacks*".

The decomposition and refinement of goals continue until we have sub-goals that are tangible [27] (i.e., they can be analyzed using a simple KDD task).

One of the main issues in goal-driven analysis is goals identification. Even if there exist some attempts to guide the identification of goals (like eliciting goals from scenarios, eliciting goals by refinement, and eliciting goals by reuse [30]). We argue that, in the context of KDD, the assistance of domain experts is necessary to guide and validate the identification of goals and influences between them.

3.4. Reasoning on the Goal Models

Reasoning on goal models has already been the subject of several studies [27] especially to determine the satisfaction or denial of a given goal, and to study goals' conflicting situations. Our objective in this subsection is to exploit ontological reasoning (based on reasoners and rule engines like FaCT++[31], Pellet [32], and Apache Jena [33]) in order to check and ensure the consistency of a goal model of a multi-view analysis, and to infer new relationships between goals by propagating existing ones. Thus, we use the two reasoning mechanisms described below:

- Checking the consistence of the goal model (formalized as an OWL ontology: classes and instances) and inferring subsumption relationships. This reasoning mechanism is supported by most of existing reasoners, and allows the detection of design errors. We used FaCT++ reasoner which is integrated as a plug-in in the Protégé Ontology Editor.
- Propagation of influence relationships between goals: based on existing influence relationships in a goal model, we can deduce new relations using SWRL rules as follows:

```
posInfluence(?x,?y) □ posInfluence(?y,?z) → posInfluence(?x,?z)
posInfluence(?x,?y) □ negInfluence(?y,?z) → negInfluence(?x,?z)
negInfluence(?x,?y) □ negInfluence(?y,?z) → posInfluence(?x,?z)
negInfluence(?x,?y) □ posInfluence(?y,?z) → posInfluence(?x,?z)
```

For example, the first rule is a rule of transitivity. It states that if a positive influence relationship exists between goals x and y, and a positive influence relationship exists between goals y and z, then a positive influence relationship is inferred between goals x and z.

The following figure highlights an example of propagation and inference situation (dashed influence relations in the goals graph are inferred relations).

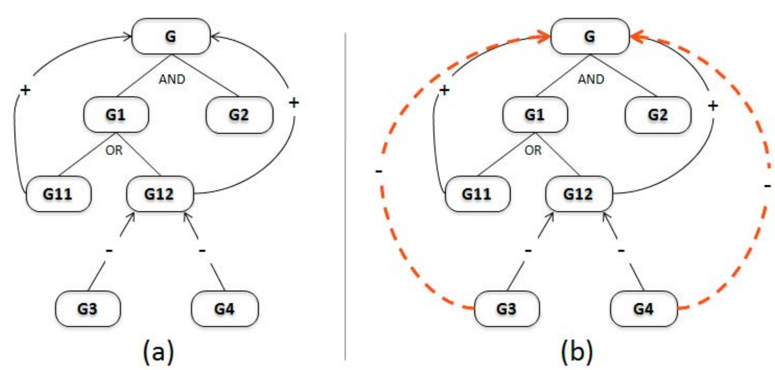

Figure 7 Example of influence propagation and inference.
(a) – Identified goals and relations.
(b) – Inferred influence relations (dashed lines).

4 Viewpoints Interaction

In most modeling paradigms for information systems (such as Entity Relationship modeling, Object Oriented modeling, and Conceptual Graphs), relations between objects are recognized to have a predominant role. In particular, relations express dependency between entities of a conceptual model. In addition, during a multi-view analysis, it is important to emphasize the interaction and interdependence between the various analyses according to different viewpoints. It is therefore necessary for our approach of knowledge representation in a multi-view KDD process to provide opportunities for expressing such interaction and dependency in terms of relations.

Based on the presented goal model and relationships between goals we introduce some relations between viewpoints of a multi-view analysis. The main purposes of these relations are to enhance coordination and mutual understanding between viewpoint stakeholders, and to allow reusability of KDD process in terms of viewpoint.

Let VP_1 and VP_2 be two different viewpoints during a multi-view analysis, $G_1=\{G_{11}, G_{12}, \ldots, G_{1n}\}$ goals associated with VP_1, $G_2=\{G_{21}, G_{22}, \ldots, G_{2m}\}$ goals associated with VP_2, $\{g_{11}, g_{12}, \ldots, g_{1i}\}$ a subset of G_1, and $\{g_{21}, g_{22}, \ldots, g_{2j}\}$ a subset of G_2.

We define and formalize (in SWRL language) *equivalence*, *inclusion*, *conflict*, and *requirement* relations between VP_1 and VP_2 as follow:

- **Equivalence:**

VP_1 is equivalent to VP_2 if:

Satisfaction of all of the goals associated with VP_1 implies satisfaction of all of the goals associated with VP_2 and vice versa.

```
satisfied(?G11) ∧ … ∧ satisfied(?G1n) →
                        satisfied(?G21) ∧ … ∧ satisfied(?G2m)
satisfied(?G21) ∧ … ∧ satisfied(?G2m) →
                        satisfied(?G11) ∧ … ∧ satisfied(?G1n)
```

- **Inclusion:**

VP_1 includes VP_2 if:

Satisfaction of some of the goals associated with VP_1 implies satisfaction of all of the goals associated with VP_2.

```
satisfied(?g11) ∧ … ∧ satisfied(?g1i) →
                        satisfied(?G21) ∧ … ∧ satisfied(?G2m)
```

- **Conflict:**

VP_1 is in conflict with VP_2 if:

Satisfaction of some of the goals associated with VP_1 implies denial of some of the goals associated with VP_2.

```
satisfied(?g11) □ … □ satisfied(?g1i) →
                                    denied(?g21) □ … □ denied(?g2j)
```

- **Requirement:**

VP_1 requires VP_2 if:

Satisfaction of some of the goals associated with VP_1 requires satisfaction of all of the goals associated with VP_2.

```
satisfied(?g11) □ … □ satisfied(?g1i)  ←
                                   satisfied(?G21) □ … □ satisfied(?G2m)
```

The equivalence relation serves mainly the purpose of reusability of KDD process in terms of viewpoint. In fact, it may be benefit within an organization to reuse successful data mining experiences to achieve different business objectives (goals) associated to different viewpoints (actors). This is done by comparing only the goal models associated with VP1 and VP2 without considering the technical details of the KDD execution plans.

The conflict, inclusion and requirement relations can be used to provide methodological assistance to a KDD user to achieve his/her goals. Especially tasks to be avoided in the case of conflict with another VP, or tasks to be partially reused in the case of inclusion and requirement.

5 Implementation Overview

Having surveyed our Knowledge Model for multi-view KDD process, our approach to model interactions between viewpoints based on goal analysis, and related inference mechanisms (i.e. SWRL rules), we proceed in this section to a brief overview of the prototype of a case tool that we propose to assist KDD users to design valid KDD process by reusing existing ones.

As illustrated in figure 5, our assistant consists of 5 main components: a Graphical User Interface, a Case Based Reasoner, a Rule Based Reasoner, a KDD Case Base, and OWL ontologies.

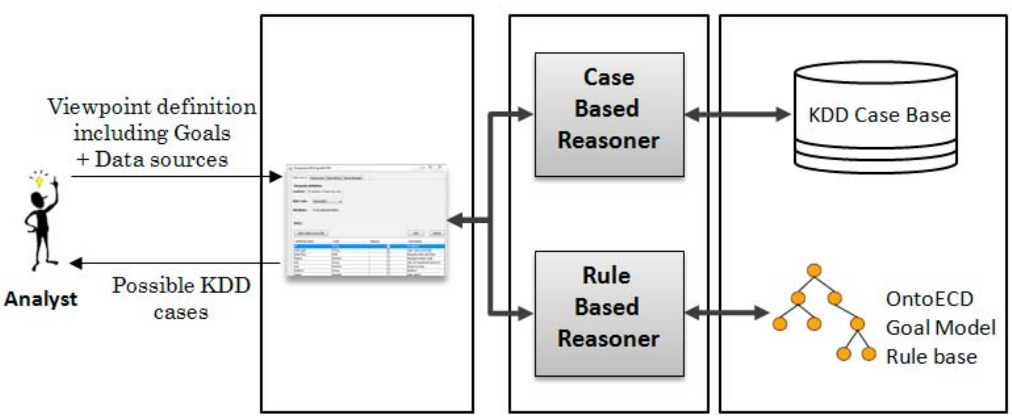

Figure 8 Architecture overview of our implementing tool.

- The Graphical User Interface allows a KDD user to select data sources and define his/her viewpoint by instantiating the generic criteria that characterize a viewpoint in KDD [7].
- OWL ontologies and rule base give a formal representation of knowledge involved during a KDD analysis. They consist mainly of the OntoECD ontology, the generic model of goals as well as related SWRL rules defining the semantic of relations between goals and mechanisms to infer influence relations.
- The KDD Case Base holds detailed annotations about successful data mining experiences (data sources characterization, goals instances, used data mining methods and their parameters, etc.). These annotations conform to the metadata schema proposed in [17] and extended by goals definition.
- The Case Based Reasoner provides a set of similar KDD cases based on the analyst definition of viewpoint (that includes goals identification and data sources) and using the previously defined relations between viewpoints (equivalence, inclusion, conflict, and requirement). Reusability/adaptation of a KDD case is carried out by the inference capabilities of the Rule Based Reasoner, which is based on FaCT ++ reasoner and SWRL.

6 Conclusion

To support users of multi-view KDD process within an organization, we have proposed a goal-based formal framework to model and formalize business objectives and interactions between viewpoints of a multi-view KDD process.

At first we have presented a goal model to identify and represent business objectives during the business understanding step of KDD. We defined also two reflexive relations within goals: AND/OR decomposition and influence relation. A precise semantic has been given for these relations as well as a formalization using SWRL language. This goal model has the benefit of assisting users during the early stage of the KDD process to clearly define business objectives of the data mining project, and then to make them persistent. This persistence will guide decisions at different steps of the KDD process, especially during the evaluation and deployment of data mining results.

Then, to formalize interaction and interdependence between various analyses according to different viewpoints, we have presented a set of semantic relations between viewpoints. We have defined equivalence, inclusion, conflict, and requirement relations. These relations allow us to enhance coordination, knowledge sharing and mutual understanding between different actors of a multi-view analysis, and reusability in terms of viewpoint of successful data mining experiences within an organization.

Finally we argue that our approach has the benefit of minimizing the cost of developing new KDD analyses (since we can develop new analyzes from earlier), and providing a methodological assistance during KDD steps (especially about tasks and choices to be reused or avoided). The evaluation of our proposal is conducted by implementing a case tool and analyzing data and viewpoints from an e-learning system.

References

1. Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 1996, 27-34.
2. Shearer, C. The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, Vol. l5(4), 2000, 13-19.
3. Kurgan, L. A., and Musilek, P. A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, Cambridge University Press, USA, Vol. 21(1), March 2006, 1-24.
4. Behja, H., Trousse, B., and Marzak, A. Prise en compte des points de vue pour l'annotation d'un processus d'Extraction de Connaissances à partir de Données. In S. Pinson and N. Vincent (Eds.), Revue des Nouvelles Technologies de l'Information (RNTI-E-3) : Vol. 1, Cépaduès-Editions 2005, 245-256.
5. Behja, H. Plateforme objet d'évaluation orientée point de vue d'un système d'information. Ph.D. thesis, Faculty of Science Ben M'sik, Casablanca, Morocco, February 2009.
6. Trousse, B. Viewpoint Management for Cooperative Design. In Proceedings of the IEEE Computationnal Engineering in Systems Applications (CESA'98), P. Borne, A. E. Kamel (editors), UCIS - Ecole Centrale de Lille - CD-Rom, april 1998.
7. Zemmouri, E., Behja, H., Marzak, A., and Trousse, B. Ontology-Based Knowledge Model for Multi-View KDD Process. International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 4(3), 2012, 21-33.
8. Schreiber, A., Akkermans, J. M., Anjewierden, A., Hoog, R., Shadbolt, N. Van de Velde, W., and Wielinga, B. J. Knowledge engineering and management: The CommonKADS methodology. MIT Press, Cambridge, Massachusetts, London, 2000.
9. Zemmouri, E., Behja, H., and Benghabrit, Y. OntoECD: A CRISP-DM Based Ontology for KDD Process and Data Mining. Journées Doctorales en Technologies de l'Information et de la Communication, JDTIC'2012, Casablanca, Morocco, 08-10 novembre 2012.
10. Bernstein, A., Provost, F., and Hill, S. Towards intelligent assistance for a data mining process: an ontology based approach for cost-sensitive classification. IEEE Transactions on Knowledge and Data Engineering, 17(4), (2005), 503-518.
11 Diamantini, C., Potena, D., and Storti, E. Ontology-driven KDD process composition. In N. M. Adams, C. Robardet, A. Siebes, and J.-F. Boulicaut (Eds.), Lecture Notes in Computer Science: Vol. 5772, Advances in Intelligent Data Analysis VIII, 285-296, Springer Verlag, 2009.
12 Euler, T. Publishing operational models of data mining case studies. In Proceedings of Workshop on Data Mining Case Studies at the 5th IEEE ICDM, 99-106, Houston, Texas, USA 2005.
13 Morik, K., and Scholz, M. The MiningMart approach to knowledge discovery in databases. In N. Zhong and J. Liu (Eds.), Intelligent Technologies for Information Analysis, 47-65, Springer Verlag, 2004.
14 Cannatro, M., and Comito, C. A data mining ontology for grid programming. In Proceedings of 1st International Workshop on Semantics in Peer-to-Peer and Grid Computing, in conjunction with WWW2003, 113-134, 2003.
15 Hilario, M., Nguyen, P., Do, H., Woznica, A., and Kalousis, A. Ontology-based meta-mining of knowledge discovery workflows. In N. Jankowski, W. Duchs, and K. Grabczewski (Eds.), Studies in Computational Intelligence, Vol. 358/2011, Meta-Learning in Computational Intelligence, 273-315, Springer Verlag, 2011.
16 Kietz, J-U., Serban, F., and Bernstein, A. eProPlan: A tool to model automatic generation of data mining workflows. In P. Brazdil, A. Bernstein, and J-U. Kietz (Eds.), Proceedings of the 3rd Planning to Learn Workshop (WS9) at ECAI 2010, 2010.

17 Behja, H., Zemmouri, E., and Marzak, A. Viewpoint-based annotations for knowledge discovery in databases. In Proceedings of IEEE International Conference on Machine and Web Intelligence ICMWI, 299-302, Algiers, Algeria, 2010.
18 Anton, A. I. Goal based requirements analysis. In Proceedings of the 2nd International Conference on Requirements Engineering. ICRE'1996, Colorado Springs, Colorado, pp. 136-144, 15-18 April 1996.
19 Anton, A. I. Goal Identification and Refinement in the Specification of Information Systems, Ph.D. Thesis, Georgia Institute of Technology, June 1997.
20. Mylopoulos, J., Chung, L. and Yu, E., (1999). From Object-Oriented to Goal-Oriented Requirements Analysis. Communications of the ACM, Vol. 42 No. 1, pp. 31-37, January 1999.
21. Nuseibeh B., and Easterbrook S. Requirements engineering: a roadmap. In Proceedings of the Conference on The Future of Software Engineering ICSE '00, 35-46, ACM, New York, NY, USA, 2000.
22. Giorgini P., Rizzi S., and Garzetti M. Goal-oriented requirement analysis for data warehouse design. In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP (DOLAP '05), 47-56, ACM, New York, NY, USA, 2005.
23. Khouri, S., Bellatreche, L., and Marcel, P. Towards a Method for Persisting Requirements and Conceptual Models in Data Warehousing Context. 27ième Journées Bases de Données Avancées BDA2011, Rabat, Morocco, October 2011.
24. Kumar M., Gosain A., and Singh Y. Stakeholders Driven Requirements Engineering Approach for Data Warehouse Development. Journal of Information Processing Systems, Vol. 6(3), 385-402, 2010.
25. Glinz, M. On Non-Functional Requirements. 15th IEEE International Requirements Engineering Conference RE 2007, 21-26, October 2007.
26. R. Jarvis, G. McArthur, J. Mylopoulos, P. Rodriguez-Gianolli, and S. Zhou, “Semantic Models for Knowledge Management,” In Proceedings of the Second International Conference on Web Information Systems Engineering WISE'01, Vol. 1, pp. 8-16, IEEE Computer Society, Washington, DC, USA, 2001.
27. Giorgini P., Nicchiarelli E., Mylopoulous J., and Sebastiani R. Formal Reasoning Techniques for Goal Models. Journal on Data Semantics, LNCS, Vol. 2800/2003, pp. 1-20, Springer, 2004.
28. Aybuke A., and Claes W. Engineering and Managing Software Requirements. Springer-Verlag Berlin, 2005.
29. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. CRISP-DM 1.0 Step-by-step data mining guide. (Tech. Rep.) CRISM-DM consortium, 1999. Retrieved from, http://www.crispdm.org
30. Rolland, C., and Salinesi, C. Modeling Goals and Reasoning with Them. In Aybüke Aurum and Claes Wohlin (Eds.). Engineering and Managing Software Requirements (EMSR), 189-217, Springer-Verlag 2005.
31. FaCT++, “FaCT++ OWL-DL reasoner,” http://owl.man.ac.uk/factplusplus/
32. Pellet: OWL 2 Reasoner for Java, http://clarkparsia.com/pellet/
33. Apache Jena: a Java framework for building Semantic Web applications, http://jena.apache.org/