View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

JIOS, Vol. 34, No. 2 (2010) Submitted 10/10; Accepted 12/10

UDC 004.032:004.652
Preliminary Communication Article

Determining semantic similarity of it systems based on the comparison of their graphical data models

Katarina Tomičić-Pupek *ktomicic@foi.hr*
University of Zagreb
Faculty of Organization and Informatics Varaždin

Abstract

Modelling is the basis for research and development of IT systems. Graphical models and graphic representations of models originally built in non-graphic languages and formalisms are often used. In modelling IT systems a need exists for comparing graphical models which can represent different variations of the same or similar modelled content or graphical models which, with certain revisions, could be applied in various domains. Graphical model in the latter case first needs to be translated into another form of predicate expressions or formal languages of modelled content representation. The lack of translation of the model for comparison is a time-consuming venture and may result in the loss of modelled relations due to differences in the "language" and representation symbols.

The goal of the paper is to explore and propose methods and procedures for determining similarities of IT systems based on the comparison of their graphical data models. The procedure of determining the similarities of graphical data models of the same type shall at the end of my research be based on semantic and structural similarity of models. In this article procedures for determining semantic similarity and their application is discussed together with examples and the procedure for determining structural similarity is proposed in roughly as it is in the finishing state of current research activities.

Keywords: semantic similarity, graphical data models, comparison

1. Introduction

Models are representations of selected relevant objects from the real world and of relations between the objects which are the subject of examination and interest. The models are built with the use of concepts which can describe the characteristics and/or relations between the concepts. Models based on graphical concepts are significantly developed and used in the IT. As all other models, graphical models are built for the purpose of researching, analyzing, simulating and representation and are developed in accordance with the rules of building models. They contain familiar, generally accepted or agreed graphical symbols for concepts, relations and characteristics of concepts.

A data model is a representation of relevant data objects of interest from the real world, defining the form, structure and content of the future IT system data base. Regardless of the building method, the model encompasses:

1. Set of concepts for modelling data structure (eg. attributes, entities, relations);
2. Set of concepts for modelling limitations and the preservation of consistency (eg. cardinality, copying, domain);
3. Set of concepts for managing data and changing the state.

A graphical data model is built with the application of the selected data modelling method. Basic and widely applicable data modelling methods are based on structural and object approach, such as Entity Relationship Attribute Modelling *(ERA models*), IDEF1X, relation model, UML diagrams (Object, Class, Package Diagrams).

A graphical data model with the previously described features can be compared to other graphical data models of the same type built for any other business domain from the real world and similarities between them can be investigated and determined. Depending on

model features and elements, semantic and structural similarities of models and their elements can be distinguished and methods, steps and comparison procedures defined.

Comparing graphical data models and drawing conclusions on their semantic and structural similarities may be useful and of interest for business or scientific reasons. Should the comparison of graphical data models of the same type reveal the existence of a set of semantically and structurally similar data models, this object set and all its sub-sets may be used as templates (or as patterns or even as metaobjects) in new business scenarios and IT data base development, thus shortening the time necessary for designing and developing the IT system. This reusability feature of data sets is of interest in business and scientific sense.

From the scientific point of view, the conclusion of analytical procedure must not necessarily be an expected conclusion. The analytical procedure itself and the manner of drawing the conclusion and the results of comparison may be of even larger significance. The procedure of determining semantic and structural similarity enables the verification of whether the same relations are valid between concepts of models of the same type built for various business domains, which can then be applicable to the analysis and comparison of other types of graphical models, such as metamodels, conceptual and reference data models.

2. Methods and procedures of comparing model objects according to semantic similarity

2.1. Overview of the results of the most resembling research on semantic similarity

Analyzing the literature in the field of exploring semantic similarity between graphical data models, two articles have been selected [1], [7], of which one most resembling was selected as the source reference. The article has been written by Song, Johannesson and Bubenko who searched for semantic similarity between data model objects [7]. They investigated semantic similarity of graphical data model objects in order to enable the integration of a number of different data shemas into a unique model (shema) which does not have any redundant objects. The authors use the term semantic similarity implying semantic resemblance of two data model objects via entity titles, entity attributes or data utilization context.

For the analysis of similarity according to the object title, the authors differentiate between 3 levels of similarity: 1) the objects have the same title; 2) the titles of objects are synonyms; 3) and the title of one object is the abbreviation of the title of the other object. For the second level of similarity the authors suggest additional clarification of the synonym by introducing semantic dictionary for the interpretation of synonym similarity of objects. Two levels of synonym similarity of a pair of objects have been defined: strong synonyms (or simply synonyms) and weak synonyms (or similar objects). The authors realize the incompleteness of semantic dictionary implementation elaboration and state the need for further research of the topic and the elaboration of object similarity levels in the semantic dictionary. It is easy to agree to such claims by the authors, as it should be researched if context similarity in the object utilization exists, regardless of how similar the titles of data model objects are.

For similarity analysis according to object attributes, the authors define 4 levels of similarity of data model objects:

1. Weak semantic similarity: two objects have weak semantic similarity if their attributes are partially overlapping;
2. Compatible semantic similarity: two objects have compatible semantic similarity if their key attributes are overlapping;
3. Equivalence semantic similarity: two objects have equivalence semantic similarity if their key attributes are identical;
4. Mergable semantic similarity: two objects have mergeable semantic similarity if all their attributes are identical.

For the analysis of similarity according to the data utilization context, the authors define the following context structure: let Ent be an entity type, and let the set of relationship types associated to Ent be {Rel*1*,Rel*2*,…Rel*n*} which are denoted to {Ent*1*,Ent*2*,…,Ent*n*} respectively, then the context of the entity type Ent is Cntx(Ent)={(Rel*i*, Ent*i*) i $\leq$i$\leq$n}.

By that definition entity context is a set of relationships which associate Ent to other entities. The contextual similarity is determined by comparing context of Ent*1* i Ent*2* based on comparing relationship sets Rel*1 i* Rel*2* to other entities Ent*1′* i Ent*2′*, respectively. Furthermore, objects are compared and their relations recognize 3 levels of similarity: 1) weak contextual connection, describing objects with weak semantic similarity and similar relations, i.e. if Cntx1=(Rel*1*, Ent*1′*)⊆context(Ent*1*) and Cntx2=(Rel*2*,Ent*2′*)⊆ context(Ent*2*), and Rel*1 i* Rel*2* are similar; 2) compatible contextual connection with compatible similarity of entities and relations, i.e. if Cntx1=(Rel*1*, Ent*1′*)⊆context(Ent*1*) and Cntx2=(Rel*2*,Ent*2′*)⊆ context(Ent*2*), Rel*1 i* Rel*2* are compatible relations to compatible related Ent*1′* i Ent*2′* ; and 3) equivalence contextual similarity, i.e. if Cntx1=(Rel*1*,Ent*1′*)⊆context(Ent*1*) *and* Cntx2=(Rel*2*,Ent*2′*)⊆context(Ent*2*), Rel*1 i* Rel*2* are equivalent relations to equivalently related Ent*1′* i Ent*2′*. Contextual similarity is the focus of further research of article authors whereby the authors plan to add degree to the context. The degree would implicate the number of similar subcontexts between entities.

2.2. The Proposal of procedures for the determination of semantic similarity between data models

As the development of procedures for comparison of graphical data models based on semantic and structural similarity is the goal of my research, the mentioned similarity levels of data model objects given by Song, Johannesson and Bubenko are identified as only partially applicable. Therefore there is a need for a proposal of a set of procedures for the determination of semantic similarity which take former research in the field in account, but also offer specific methods for the comparison of semantic similarity. My proposal of procedures is a research output on this subject and is described in Table 1.

The first procedure is intended for finding pairs of data objects in two data models which are semantically similar at the level of object titles. This procedure is based on the semantic dictionary concept suggested by Song, Johannesson and Bubenko. In my opinion the semantic dictionary should be extended by a descriptive title clarification allowing recognition of data object title analogies in various business domains. This extension allows data objects from the first model, that have no homonyms, abbreviations or synonyms by their title in the other model to be taken into consideration when comparing data objects because of their title analogies.

The second procedure is also based on the suggested comparison method from Song, Johannesson and Bubenko, and is used to demonstrate that the previous extension (the descriptive clarification in the first procedure) should be applied to the comparison of the attributes as well. This will allow a comparison of data object attributes which have completely different names or meanings in various business domains but which may be similar in some other context of data utilization.

The third procedure is an original proposal of a procedure for determination of semantic similarity between object of two (or more) data models by spreading the comparison to the level of comparing processes that use the data objects from data models as data flows. The comparison of processes and their data flows should be based on comparing process models representing typical business scenarios. The analysis of processes and the data the processes are receiving or sending enables drawing analogies based on the premise that similar processes are using similar data in a similar way.

The last procedure is also an original proposal of a procedure for determination of semantic similarity between object of two (or more) data models by widening the comparison to the level of comparing the data object dynamics represented as automata using transient state diagrams.

Procedure	Description	Executor	Result
Determination of semantic similarity according to the object title	Develop a semantic dictionary containing one column for each data model taken into the comparison process (e.g. if two data model are compared, the semantic dictionary will consist of two columns). In the first column object title of the first model is added followed by a descriptive title clarification allowing recognition of data object title analogies in various business domains. Object title of the second model is added in the second column if it is a homonym, abbreviation or synonym to the title in the first column, or if its descriptive title clarification corresponds to the descriptive title clarification of the object in the first column.	IT system designer	Pairs of objects semantically similar in object titles
Determination of semantic similarity according to object attributes	Comparing object attributes of two models – if they match or are identical, the pair objects will be considered similar in key attributes. The purpose of the step is to determine which data model objects may be used for analyzing the model structure using a graph theory based algorithm.	IT system designer	Pairs of objects semantically similar in object attributes
Determination of semantic similarity of data utilization	Analyzing typical business scenarios of data utilization and concluding by analogy on the possible pairs of similar objects. The premise is that similar processes use similar data in a similar way.	IT system designer in coordination with business experts	Pairs of objects semantically similar in utilization context
Determination of semantic similarity of the objects dynamics	With the transition state diagram research the life cycle of data objects. During life cycle of a data object the initial state of life cycle of other data objects may be activated – such cases need to be researched further.	Business analysts and IT system designer	Pairs of objects with similar semantic relations to other objects

Table 1: Procedures of determination of semantic similarity

The comparison procedures of semantic similarities of graphical models are described in detail together with examples as follows.

2.2.1. Determining semantic similarity according to the object title

Develop a semantic dictionary containing object titles and their descriptive clarification of one model in the first column. Object title of the second model is added in the second column if it is a homonym, abbreviation, a synonym or a analogue data object title to the title in the first column. The resulting object pairs in the same column shall be considered similar. Table 2 shows a segment of the semantic dictionary row illustrating the structure of the semantic dictionary.

First model object	*Semantically similar second model object*
Schedule The schedule dictates the allocation of resources and defines the procedures of generating services for the end buyer – student.	**Production order** Production order dictates the allocation of resources and defines the procedures of generating services for the end buyer.
Student Special class of business partner obtaining their status by enrolment in a university. By enrolling in a university, students "order" education services from the university. Students are given an invoice for the education service. The invoice can be paid by the student or another entity (so called free education, when the education service is paid by the competent ministry).	**Business partner** Special class of business partner obtaining their status by signing a cooperation contract, ordering a product or service or in any other way of contracting the rights and obligations of business partners.

Table 2: Example of the semantic dictionary

The effectiveness of this procedure depends on the ability to recognize data object title analogies in various business domains and on competencies of experts involved in the development of the semantic dictionary.

2.2.2. Determining semantic similarity according to object attributes

Determining semantic similarity according to data object attributes may be implemented on all object attributes. Comparison criteria may be the identification or description feature, the attribute title (synonyms and homonyms in title), attribute domain, syllable length, optionality of data input, referential integrity and other restricting or descriptive features. Figure 1 shows the example of pairing key attributes of two entity object groups. Group 1 is made up of entities Person-enrolment-Program of Study, and group 2 is made up of entities Business partner-contract-form of cooperation.

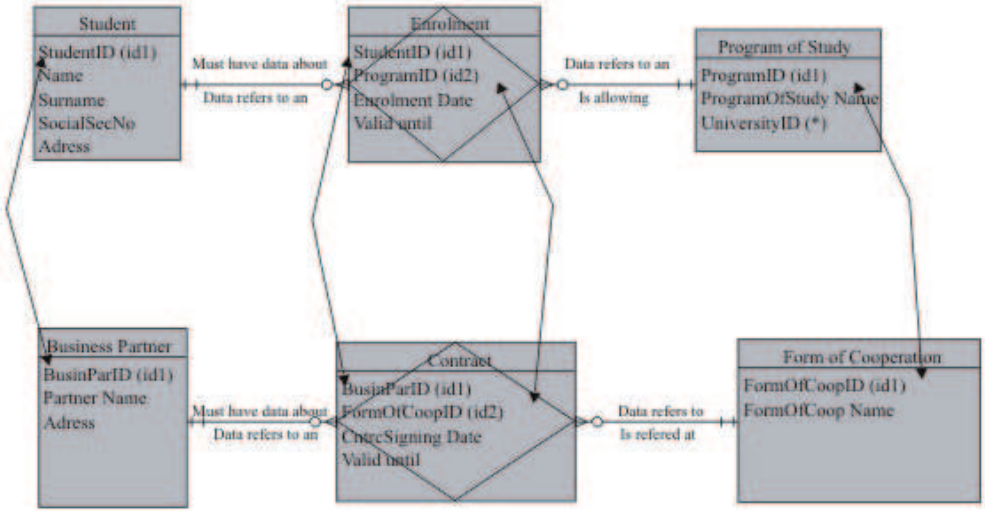

Figure 1: Determining semantic similarity according to attributes

Results of comparison and pairing of data model objects according to semantic similarity based on the similarity of attributes, the following conclusions may be drawn:

1. Object pairs of different models in which weak semantic similarity is recognized may exist. These need not be taken in account in analyzing the structure of the model.

2. Object pairs of different models in which compatible semantic similarity is recognized may exist, such as Student-Business partner, enrolment-contract.
3. Object pairs of different models in which equivalent semantic similarity is recognized may exist.
4. Object pairs of different models in which mergeable semantic similarity is recognized may exist.

Graphical data model objects with compatible, equivalent and mergeable level of semantic similarity should be used for research and determination of structural similarity of models. The effectiveness of this procedure depends on the competencies of experts involved in the analysis of similarity of data object attributes and the ability to recognize and pair objects from various business domains.

2.2.3. Determining semantic similarity of data utilization

The third procedure of determining similarities of IT system data models is related to data utilization by business processes in various business contexts. Business context is determined by messages exchanged between the system (or objects in the system) and the systems and objects within its environment. If the process is considered as a system, then the business context is determined by data contents exchanged between processes. Determining semantic similarity of data utilization is based on an analysis of data flows within an static models that show the processes, data and data flows that describe how processes use data. The assumption that allows the determination of semantic similarity of these relations is as follows: similar processes are using similar data in a similar way. This assumption requires the definition of the similarity of the process, data, and use of data. The similarity of the data can be examined by comparison according to data object names and attributes (chapter 2.1. And 2.2.).

Semantically similar processes can be processes from different business domains or industries (or other elements that describe an action, such as activities, operations, tasks, etc.) that have a similar role in overall business workflow of the business they belong to. The significance of business processes can be estimated by analyzing the level of process complexity [4] as individual, functional i.e. vertical, horizontal i.e. cross-functional process or by process classification based on Porters value chain [6]. Analysis of processes which constitute Porter value chain provides the assessment of process's significance as a supporting set of activities or as those directly involved in generating profits, and fulfilling the mission of the organization (also called primary activities). If the two processes from to different business domains (or models) belong to a same group of supporting activities or to a same of primary activities, then we can say that the processes have similar significance in overall business workflow and that these processes can be considered as semantically similar.

Data utilization is influenced by the character of the process, or in other words, the meaning of the process derives from the sense and role in the process of transforming the input data content in the output. Way to compare and determine the similarities of the data utilization depends on the method of modelling the static connections between model objects. I give two examples:

- Analysis of data flow model elements is performed by comparing the flows between processes and data stores. Processes and data stores are origin and destination nodes of data flows. Data flows are channels for exchanging contents stored in the data store. If similar processes (from two different models) are associated with similar data store via similarly directed data flows, then the data utilization is also similar.
- Analysis of data utilization in the UML activity diagram is performed by comparing the activities, objects and flows that connect the activities and objects. If an object flow starts in an activity and ends in an object that indicates that the activity can create or update the object. If an object flow starts at the object and ends in an activity, then that means that the activity is using an object. If similar activities (from two different models) are associated with similar data store via similarly directed data flows, then the data utilization is also similar.

The semantic similarity of data utilization can be verified in drawing up and analyzing typical business scenarios for data utilization and in drawing conclusions on possible pairs of

similar objects on the basis of analogies. Business scenarios can be modelled in graphical process models or process relations and matrixes. Figure 2 shows a simple example of comparing data utilization presented in process models.

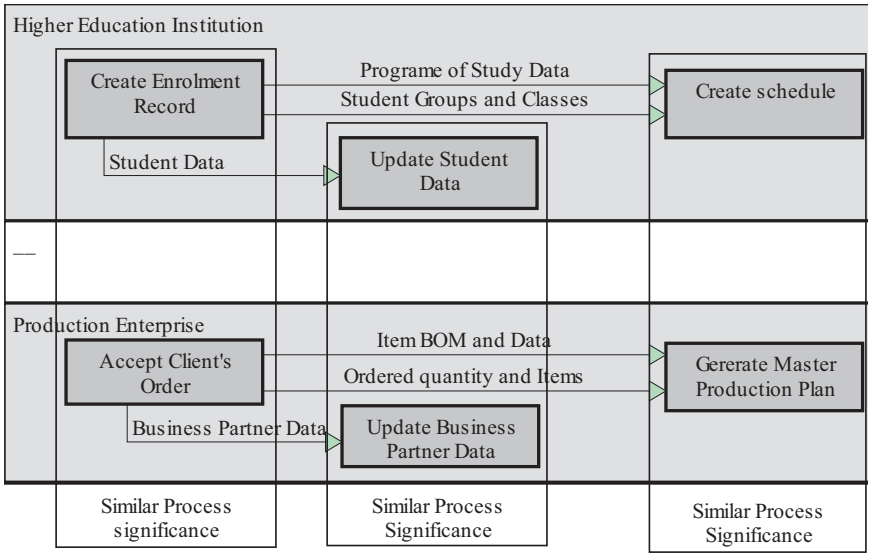

Figure 2: Example of comparing business scenarios and determining the similarities of processes and their data contents

The effectiveness of this procedure depends largely on the competencies of experts involved in the analysis and modelling of scenarios, understanding business processes and the ability to recognize concept analogies in various business domains.

2.2.4. Determining semantic similarities of object dynamics

The fourth procedure of determining similarities of IT system data models is related to data object dynamics. Determining semantic similarities of data object dynamics is based on an analysis of the life cycle of objects. Each data object used in business has its life cycle which is combined of a number of states that the object goes through. The first state is the design state or the data input state, while the deletion or permanent disabling of accessing data is the last state. Between these two end states, a number of states exist where data are used for various purposes. The instance of object is found in one and only one state in every moment. The assumption that allows the determination of semantic similarity of object dynamics is as follows: If the two diagrams contain analogue or states of the same type, or analogue or events of the same type which cause the transition from one state to another, it can be concluded that there exists a certain semantic similarity between them.

The stages in which an object can be found may be represented in automaton or *transient state diagram.* This diagram represents the manner of changing the object by calling application procedures for data processing. The states are represented by circles (S1 to Sn). The states represented in double circles are the so called final states. Transitions from one state to another are caused by certain events. The diagram shows events in the form of arrows (d1 to dm). If we apply the same automaton to another data model object of another business domain (eg. enrolment list of students), we can investigate and determine if it shows the same states and transitions, and the relation through events towards other objects, the life cycles of which are related to the enrolment list in the same way as the Production order, business partner and items with the client's Order.

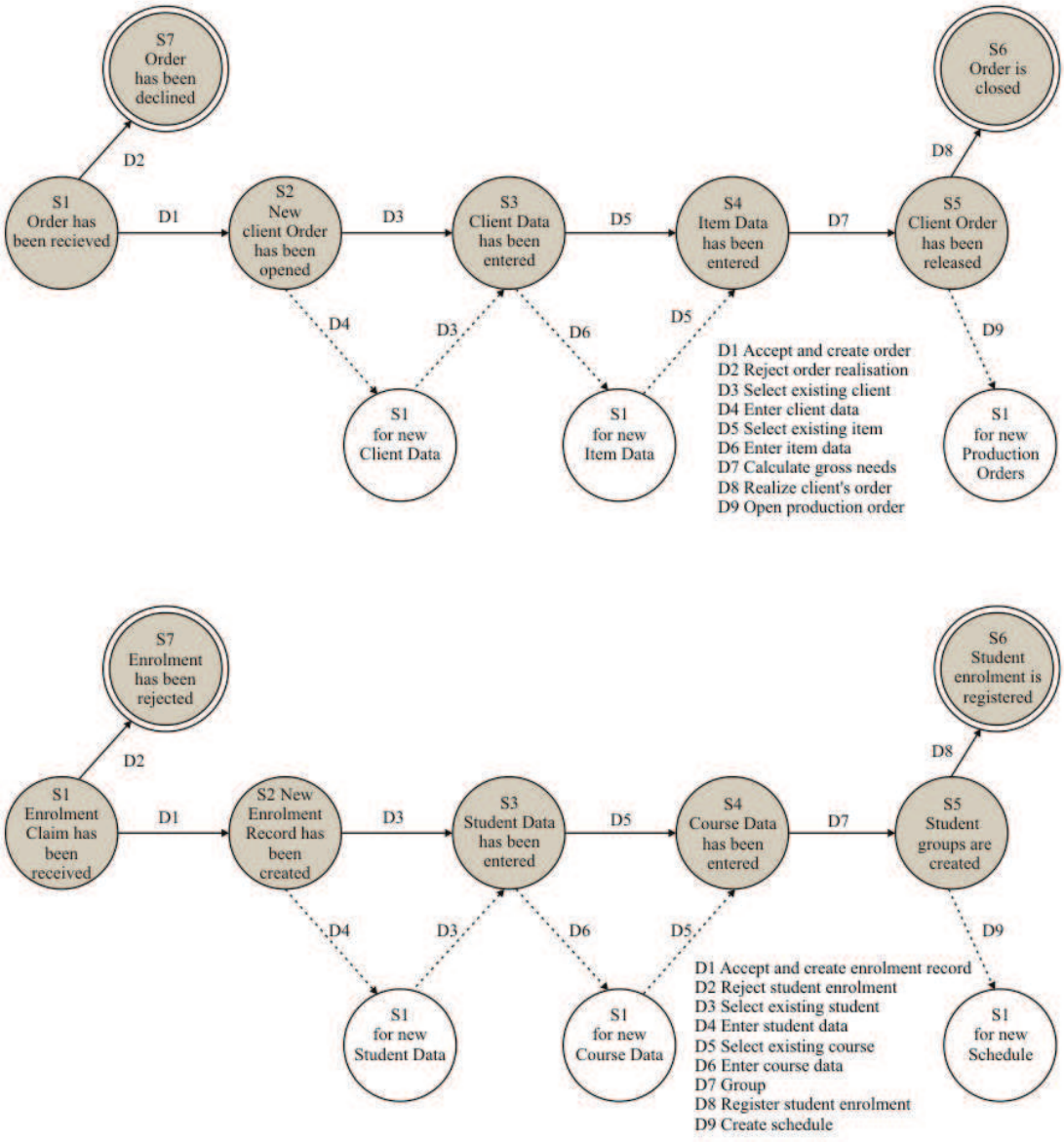

Figure 3: Example of determining logical connections between objects using transient state diagrams

The effectiveness of this procedure depends largely on the competencies of experts involved in the analysis and modelling of events and states, understanding IT application functionality and the ability to recognize concept analogies in various business domains.

3. Comparison of model structure using graph theory-based algorithm

Data models built for two different business domains from the real world can be compared, researched and their structural similarity determined only if the mentioned models have been built using the same modelling method and if they use the same concept sets for modelling data structure, limitations and the preservation of consistency and data management. If this prerequisite of the same type of models and concept has not been met, the models need to be transformed into the required form. The ERA model of hospital service provision and the UML Object Diagram of a registration office therefore cannot be compared.

The second minimal prerequisite for comparison is the analysis of whether the models contain analogue or terms of the same type which would represent reference points for

analysis and comparison of models. If this prerequisite has not been met, there is no point in comparing models, as relations between concepts describe incomparable relations of the first and the second model.

There are several formal methods which could be considered as suitable for graphical data models formalization. Vatanawood and Rivepiboon [8] and Yugopuspito and Araki [9] explored the application of Z notation in data modeling, M. Keet [2] conducted a comparison of conceptual data modeling languages based on Description Logic Languages, Mammar, and Laleau [3] proposed modeling of relational concepts shown in UML diagrams with the B-method, Kim Pilho mentions the possibility of data model presentation using graphs in the dissertation "E-model: event-based graph data model theory and Implementation" [5] and claims that "graphs", (are a type of data model) "graph data models are applied in areas in which information about data interconnectivity or topology is more important, or as important, as the data itself".

In this case, graph theory is selected as a formal mathematical method for graphical data model formalization and comparison in favor of other considered methods for following reasons:

- In graph theory there is simplicity of graph concepts which can be used to represent data objects from graphical data models as a new structure.
- In graph theory there is a limited set of applicable operations (e.g. matrix permutation, elimination of matrix rows and columns) over a small set of basic elements (nodes and arces).
- In order to draw a graph and to compare graphs using a limited set of operations and algorithms a certain independence of applications is an advantage to some formal languages.
- Intuitive language based on common known graphical symbols is necessary because at some point business experts with no or little IT skills could be involved in the comparison process.
- The adequacy of the use of graph theory for data model representation has been advocated in present literature.
- Data models are usually already graphical representations and the intention of expressing them in predicate expressions or formal language may be a time-consuming venture.
- The complexity of formal methods requires involvement of competent experts to avoid losses of modelled relations due to differences in the "language" and representation symbols.

If a data model can be presented by a "graph" then mathematical methods based on graph theory can be applied for the comparison of models. The structure, limitations and operations on data model concepts may be recognized in the structure, limitations and operations on graph nodes and edges. A graph consists of nodes (entities in a data model) and arces (relations between entities). A certain arc weight defines the content of incidence matrix and could express cardinality, referential integrity, and other relation quantities. If the comparison of graphs via reduction of incidence matrixes reveals a sub graph with at least two nodes then a certain structural similarity of graphs elements exists, and then it can be expected that data models from which the graphs originated also have structurally similar elements. For the reduction of incidence matrixes in order to find similar sub graphs an appropriate algorithm needs to be developed. This is the subject of further research to be conducted in this sense.

4. Conclusion

Development of procedures for comparison of graphical data models on the basis of semantic and structural similarity may be useful and of interest for business or scientific reasons as mentioned earlier. In the article some procedures for comparing data models and drawing conclusions on their semantic similarities are proposed. These procedures are meant to be

applicable for determination of semantic similarity according to the object title and according to object attributes, determination of semantic similarity of data utilization and determination of semantic similarity of objects dynamics. More research is needed in the area of determination of structural similarities of data models, whereby in this article the application of graph theory is implied as an appropriate method for comparison of graphical data models.

References

[1] Kashyap, V.; Sheth A. Semantic and Schematic Similarities between Database Objects A Context-based approach. *The VLDB Journal — The International Journal on Very Large Data Bases,* Vol.5 No.4, p.276-304, 1996

[2] Keet, C. M. A Formal Comparison of Conceptual Data Modeling Languages. http://www.meteck.org/files/EMMSAD08CMcompCMK.pdf (downloaded January, 25th 2010)

[3] Mammar, A.; Laleau, R. From a B formal specification to an executable code: application to the relational database domain. *Information and Software Technology*, Vol. 48, Issue 4, p.253-279, 2006

[4] Laguna, M.; Marklund, J. *Business Process Modeling, Simulation, and Design*. Pearson Prentice Hall, New Jersey, 2005

[5] Pilho, K. E-model: event-based graph data model theory and Implementation, http://smartech.gatech.edu/handle/1853/29608?show=full (downloaded October, 20th 2010)

[6] Porter, M. *Competitive Advantage Creating and sustaining*. The Free Press, New York, 1985

[7] Song, W.W.; Johannesson, P.; Bubenko Jr. J.A. Semantic Similarity Relations in Shema Integration, *Lecture Notes in Computer Science*, Vol.645/1992, p.97-120, 1992

[8] Vatanawood, W.; Rivepiboon, W. Formal Specification Synthesis for Relational Database Model. *International journal of intelligent systems*, Vol.19, p.159-175, 2004

[9] Yugopuspito, P.; Araki, K. Transformational Object-Relational Database Model in FormalMethods. *Transactions of Information Processing Society of Japan*, Vol.42, No.Sig.5, p.71-80, 2001