ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2008, Vol. 37, No. 1

FORMAL CONCEPT ANALYSIS FOR BUSINESS INFORMATION SYSTEMS

Algirdas Laukaitis, Olegas Vasilecas

Vilnius Gediminas Technical University
Sauletekio al. 11, LT-10223 Vilnius-40, Lithuania

Darius Plikynas

Vilnius Management School
J. Basanaviciaus 29A, LT-03109 Vilnius, Lithuania

Abstract. In this paper, we present a methodology based on formal concept analysis (FCA) for the information system model verification. We show that FCA can be useful for understanding conceptual model topology and it can be used to improve the structure of the conceptual model. IBM's Information Framework (IFW) Financial Services Data Model (FSDM) has been used for the present research. By using FSDM, we demonstrate that the IS model can be recreated with formal concept analysis.

Keywords: Information systems engineering, formal concept analysis.

1. Introduction[1]

Software engineers and business analysts spend hours in defining information systems requirements and finding common ground of understanding. Quite often one finds that there were several meetings just for understanding and defining only one IS concept. Then, automatic verification of the information system model and integration of such technology into information systems engineering and modeling is an important factor in meeting challenges created by computing. The development of the information systems in such complex domains like finance or insurance services requires flexible and bottom-up integrated representations. Enterprise architectural frameworks like Zachman's Information Systems Architecture (ISA) [12] or ARIS [10] meet genuine challenges when applied to the whole enterprise up to the detail representations level. The ripple effect caused by changes in some cell of the framework matrix produces overwhelming communication and management cost.

In this paper, the formal concept analysis [2] is suggested to reinterpret IS model and to verify the comprehensibility and soundness of the information system model. All presented ideas and methodological inference have been tested with the IBM Information FrameWork (IFW) [6], which is a comprehensive set of banking specific business models from IBM corporation. For our research we have chosen the set of models under the name *Banking Data Warehouse*.

The rest of the paper is organised as follows. First, we present the general framework of automated model generation system from the IS documentation and engineers utterance. Automatically generated hierarchical concept lattice can be used by the modeler as the tool for finalizing IS model and avoiding errors in the model. Next, we present IBM's IFW solution and the models from it which we used in our experiments. In that section we present formal concept analysis as the formal technique to analyze IS model on the *object:attribute* sets.

2. General framework of the solution

Conceptual models offer an abstracted view on certain characteristics of the domain under consideration. They are used for different purposes, such as a communication instrument between users and developers, for managing and understanding the complexity within the application domain, etc. The presence

[1]The work was supported by project Reg. No. BPD2004-ERPF-3.1.7-06-06/0014.

of tools and methodology that supports integration of the requirements textual documents and communication utterance into knowledge bases is crucial for the successful IS architectural framework development.

Even more, we can say that the essence of modeling is the ability of modeler to classify the textual information and then to represent it by some formal modeling language. Then, artificial intelligence technologies, that will attempt to automate IS modeling, must follow that cognition process of human modeler. In this paper we suggest the use of self-organizing maps to classify IS documentation and IS utterance on a supervised and an unsupervised basis. The self-organizing maps have been extensively studied in the field of textual analysis. Such projects like WEBSOM [7], [8] have shown that the self-organizing map algorithm can organize very large text collections and that SOM is suitable for visualization and intuitive exploration of the documents collection. The experiments with the Reuters corpus (a popular benchmark for text classification) have been investigated in the paper [5] and there were presented an evidence that SOM can outperform other alternatives.

Nevertheless, in the field of information systems modeling the connectionist paradigm has been met with some scepticism. The reason is that IS architects and modelers want to give the credibility on how clusters received from documents processing are related and explain semantic meaning of the underlying documents topology. To overcome this problem, we suggest that the formal concept analysis (FCA) [2] can give more on that account by formally analyzing the set of objects and their attributes. On the other hand, when directly applied to the big data set of textual information, formal concept analysis gives little meaning with the presentation of overwhelming lattice. Those arguments motivate integration of the formal concept analysis and other text clustering techniques. In that sense our work bears some resemblance with the work of Hotho et.al. [4]. They used BiSec-kk-Means algorithm for text clustering and then formal concept analysis was applied to explain relationships between clusters. The authors of that paper have shown the usability of such approach in explaining the relationships between clusters of the Reuters-21578 text collection.

Our approach differs in two important respects. First, our goal is not text clustering. Our goal is automated generation of the ontology from textual documents if there is no knowledge base produced by human experts. In case when the knowledge base has already been developed, we seek for a method that formally measures the comprehensibility of the knowledge base topology and in case of new documents and concepts automatically integrates them into the knowledge base.

The overall process of automatically clustering concepts descriptions and then deriving concept hierarchies from self-organizing map is presented in Figure 1. First, the corpus is created form the knowledge base concept descriptions. In the figure it is named the domain descriptions. Then, vector space of the corpus is created using natural language processing framework, domain ontology and WordNet ontology [9]. The self-organizing network is built and used for cluster analysis. Next, with conceptual context and formal concept lattice improvements are made in the understanding of clusters relationships. All process is interactive with the analyst, who checks it under manually created conceptual model.

Some philosophical arguments for the use of the connectionist paradigm in the context of information systems development can be found in the paper of Honkela [3]. By the reference to the works of Von Foerster, he supposed that most information systems are developed as "trivial machines" to be predictable and controllable. The requirements for more flexibility and adaptability for the information systems can lead to the use of connectionist paradigm. All results in this paper can be interpreted as an implementation of these philosophical statements.

3. Business knowledge bases and formal concept analysis

Conceptual centric modeling can be an effective tool for driving ambiguity and vagueness out of information systems. But data centric enterprise wide models are rarely built and few organizations even tried to surround their information systems and business activities with such models. The problem with data centric enterprise wide models is that they are difficult to understand. Their abstract and generic concepts are unfamiliar to both business users and information systems professionals, and remote from their local organizational contexts [1]. The natural language processing and understanding techniques can be used to solve mentioned problems. But before applying the NLP techniques for the IS engineering, we must to have some formal method to deal with the sets of $\{$*classes, object and attributes*$\}$ which are products from systems of natural language processing. In this section we will introduce the formal concept analysis as the method for automatically building hierarchical structure of concepts (or classes) from the $\{$*object:attribure*$\}$ set.

In left side of the Figure 2 we can see the small part of the IBM IFW financial services data model (FSDM) [6], which is a domain specific model based

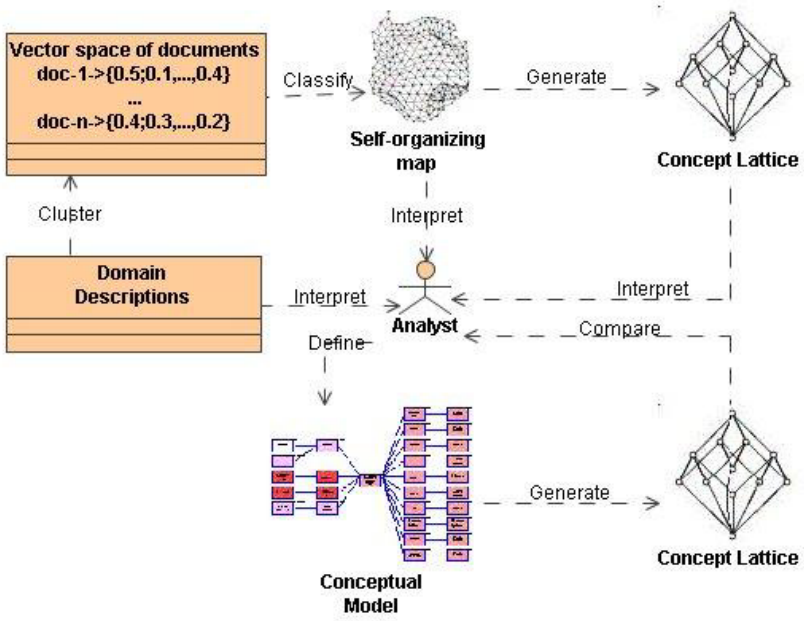

Figure1. Process of integration: Conceptual modeling, clusters detection and interpretation by use of formal concept analysis

on the ideas from the experts in the IBM financial service solutions center. The IBM financial services data model is shown to consist of a high level strategic classification of domain classes integrated with particular business solutions (e.g. Credit Risk Analysis) and logical and physical data entity-relationship (ER) models.

The model is divided into a number of levels with a different degree of abstraction: the 'A' level with nine data concepts that define the scope of the enterprise model (involved party, products, arrangement, event, location, resource items, condition, classification, business), the 'B' level with business concepts hierarchies (more than 3000 concepts), the 'A/B' level with business solutions (integrates business solutions with more than 6000 concepts) and 'C' level - entity relationship ER diagram with about 6000 entities, relationships and attributes. The concept lattice of shown model extract has been produced by formal concept analysis with Galicia software [11] and is shown in the right-hand side of Figure 2. As we can see it is consistent with the original model. It replicates underlying structure of conceptual model originally produced by a human expert team and, in addition, suggests one formal concept that aggregates *Arrangement* and *Resource Item*: the two top concepts from the original model. Next, we will give a small introduction to the Formal Concept Analysis (FCA).

3.1. Lattice of the IS conceptual model

Formal Concept Analysis is used to represent underling data in the hierarchical form of the concepts. The most adapted form in the FCA analysis for the data representation is the concept lattice (CL). Due to its comprehensive form in visualising underlaying hierarchical structure of the data and rigorous mathematical formalism FCA grown up to mature theory for data analysis from its introduction in the 1980s [2]. FCA successfully has been used in many applicable areas, but our interest in this paper is the ability to use it in the area of the information systems modeling. In defining the concepts and attributes FCA takes similarities with the database theory and object orientated system design. Due to this fact the FCA has been often applied for class diagram design in information systems [2]. Next, we introduce the definitions

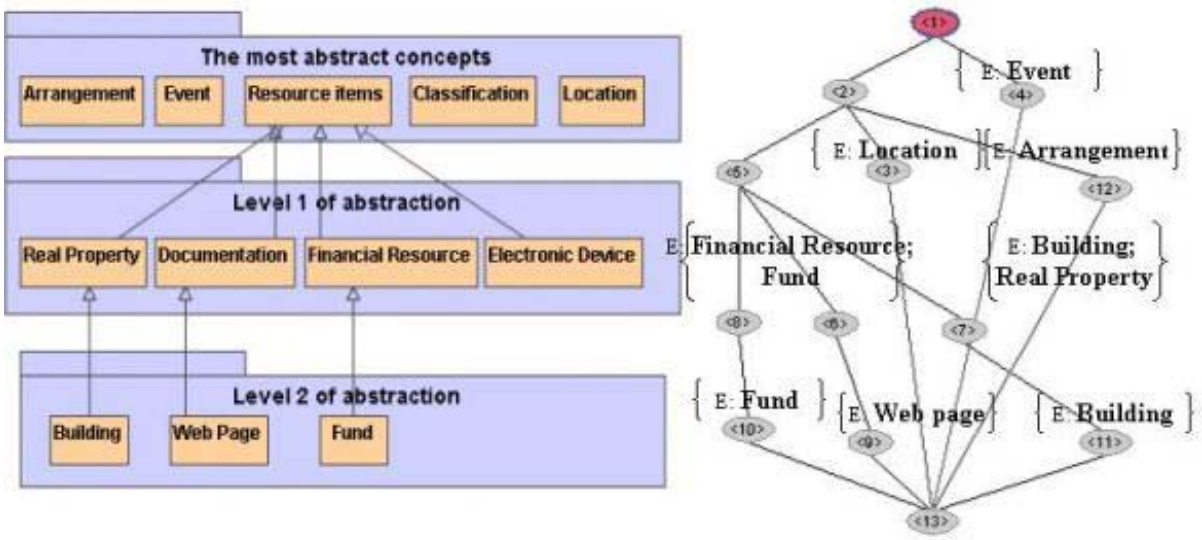

Figure 2. Left side:A small extract from the financial services conceptual model.Rightside: Concept lattice from this conceptual model.(We see that FCA depicts th estructure from the conceptual model

of the FCA and then present a more gentle example in explaining rigorous mathematical definitions from the area of financial services domain.

Let G be a set of objects that we able to identify in some domain (e.g. loan, mortgage, leasing, person, client etc.). Let M be the set of attributes. We identify the index I as a binary relationship between two sets G and M i.e. $I \subseteq G \times M$. A triple $\mathbb{K} := (G, M, I)$ is called a formal context. For $A \subseteq G$ we define

$$A^{'} := \{m \in M | (g, m) \in I \ \texttt{for all} \ g \in G\}$$

and dually, for $B \subseteq M$,

$$B^{'} := \{g \in G | (g, m) \in I \ \texttt{for all} \ m \in B\}.$$

A formal concept of a formal context (G, M, I) is defined as a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A^{'} = B$ and $B^{'} = A$. The sets A and B are called extend and intend of the formal concept (A, B) . The set of all formal concepts $\mathfrak{B}(\mathbb{K})$ of a context (G, M, I) together with the partial order $(A_1, B_1) \leqslant (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2$ is called the concept lattice of context (G, M, I) .

For the more informal introduction to the area of the formal concept analysis we can return to Figure 2. The conceptual model extract from the figure has 12 objects and 137 attributes (the whole model has more than 1000 objects and more than 4000 attributes). FCA algorithm *Incremental Lattice Builder* generated 11 formal concepts. In the lattice diagram, the name of an object g is attached to the circle and represents the smallest concept with g in its extent. The name of an attribute m is always attached to the circle representing the largest concept with m in its intent. In the lattice diagram, an object g has an attribute m if and only if there is an ascending path from the circle labeled by g to the circle labeled by m. The extent of the formal concept includes all objects whose labels are below in the hierarchy, and the intent includes all attributes attached to the concepts above. For example, the concept 7 has {*Building; Real Property*} as extend (the label *E:* in the diagram), and {*Postal Address; Environmental Problem Type;Owner;... etc.*} as intent (due to the huge number of attributes they are not shown in the figure).

4. Conclusion

Conceptual models and other forms of knowledge bases can be viewed as the products emerged from human natural language processing. The self-organization is the key property of humans mental activity and the present research investigated what self-organization properties can be found in the knowledge bases. We have shown that with the self-organizing map and formal concept analysis we can indicate inadequateness of the concept descriptions and improve the process of knowledge base development. Presented methodology can serve as the tool for maintaining and improving enterprise-wide knowl-

edge bases. Additionally, we provided evidence that high quality documentation can be reused as the separate module in the IS natural language interfaces.

References

[1] **P. Darke, G. Shanks.** Understanding Corporate Data Models. *Information and Management* 35, 1999, 19-30

[2] B. Ganter, R. Wille. Formal Concept Analysis: Mathematical Foundations. *Springer, Berlin-Heidelberg*, 1999.

[3] **T. Honkela.** Von Foerster meets Kohonen – Approaches to Artificial Intelligence. *Cognitive Science and Information Systems Development. Kybernetes*, 34, 1/2, 2005, 40-53.

[4] **A. Hotho, S. Staab, G. Stumme.** Explaining text clustering results using semantic structures. *In Principles of Data Mining and Knowledge Discovery, 7th European Conference, PKDD* 2003, *Croatia. LNCS. Springer*, 2003 22-26.

[5] **C. Hung, S. Wermter, P. Smith.** Hybrid Neural Document Clustering Using Guided Self-organisation and WordNet. *Issue of IEEE Intelligent Systems*, 2004, 68-77.

[6] IBM. IBM Banking Data Warehouse General Information Manual. *Available from on the IBM corporate site http://www.ibm.com*.

[7] **S. Kaski, T. Honkela, K. Lagus, T. Kohonen.** WEBSOM self-organizing maps of document collections. *Neurocomputing*, 21, 1998, 101-117.

[8] **K. Lagus, T. Honkela, S. Kaski, T. Kohonen.** WEBSOM for textual datamining. *Articial Intelligence Review,* 13 (5/6), 1999, 345-364.

[9] Miller, G.A.: WordNet: A Dictionary Browser, Proc. 1st Int'l Conf. Information in Data, (1985) 25-28.

[10] A.-W. Scheer, M. Nüttgens. ARIS Architecture and Reference Models for Business Process Management. *In: W.v.d. Aalst, J. Desel, A. Oberweis (ed.): Business Process Management – Models, Techniques, and Empirical Studies. Berlin et al*., 2000, 376-389.

[11] **P. Valtchev, D. Grosser, C. Roume, H.M. Rouane.** GALICIA: an open platform for lattices. *In A. de Moor B. Ganter, editor, Using Conceptual Structures: Contributions to* 11*th Intl. Conference on Conceptual Structures*, 2003, 241-254.

[12] **J.A. Zachman.** A Framework for Information Systems Architecture. *IBM Systems Journal*, 26, *No*.3, (1987, 276-292.

Received May 2007.

DOI: 10.5755/j01.itc.37.1.11923