Automatic Detection of Semantic Clusters in Glossaries

Marcela Ridao
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

Jorge Horacio Doorn
Universidad Nacional del Oeste, Argentina & Universidad Nacional de Tres de Febrero, Argentina

INTRODUCTION

The most important principle of the Requirements Engineering is to perform all of its activities within the framework of the clients and users culture. This is the major difference between Requirements Engineering and the system analysis, since the latter models the context in which the future software system will be inserted, using computer science techniques, models and resources (Pressman, 2005; Sommerville, 2011). To plan the functionalities that a future software system will provide, in terms of the business process peculiarities, is not an easy task for the requirements engineer. As a first step, he or she must deeply understand the peculiarities of the context in which the future software system will run. There have been proposed several approaches to elicit and model such knowledge (Castro, Kolp, & Mylopoulos, 2002; Van Lamsweerde, 2001). Creating a glossary (Weidenhaupt, Pohl, Jarke, & Haumer, 1998) of the vocabulary used in Universe of Discourse (UofD), at the beginning of the project, is one of them. The UofD is the context in which the software is developed. It includes sources of information and people involved with the software: users, software engineers, domain experts, etc.

The research, whose main results are presented in this chapter, was done in a process, which precisely begins with developing of the Language Extended Lexicon (LEL), which is by itself a sort of enriched glossary (Leite & Franco, 1993). The fact that language carries knowledge, cultural information and reflects the substantial and particular ways of thinking of people (Crozect & Lidedicont 2000; Nettle & Romaine, 2000; Salim, 2017) has allowed, along many previous studies, to accelerate and to systematize the comprehension of the context for which the system will be developed (Leite, Hadad, Doorn & Kaplan, 2000). Besides the LEL, all the remainder models needed to establish the software system requirements, are also created using natural language; this is done to improve the communication among all stakeholders.

Lately, a deep analysis of the LEL has proven that the way in which it was used, did not take full advantage of its potential. There is some sort of hidden information, not easily perceived. In the chapter “Displaying Hidden Information in Glossaries”, included in the fourth edition of the Encyclopedia of Information Science and Technology, a strategy to visualize the grouping of terms in the LEL was presented (Ridao & Doorn, 2018). Experiments performed on real world cases have shown that the clusters obtained using syntactic resources of the LEL model, coincide with information nuclei of the application domain. Such strategy was based on graphs constructed using hypertext links embedded in the LEL model. Unfortunately, graphs of a relevant quantity of LELs do not present clearly distinguishable clusters, even when groups of related terms exist, and they may be found by human observers, after a careful scrutiny. This is the weakness of relying only on the detection of groups of terms by means

DOI: 10.4018/978-1-7998-3473-1.ch052

of the visual perception of clusters in the graph. Coping with this weakness was the motivation of the research project whose results are reported here.

Facing Structurally Complex Disciplines By Automatic Graph Construction

Disciplines dedicated to the understanding of phenomena whose dominant aspect is their structural complexity, instead of the essential complexity of their components, have lately acquired an increasing importance (Barabasi, 2002; Dorogovtsev & Mendes, 2003). The classic visual representation of these networks is done by means of graphs. What actually happens is that when the number of nodes surpasses a moderated limit, those graphs are no longer useful to observe the relevant characteristics of the network. This limitation may be tempered by using automatic graph construction. The automatic graph construction procures to reach an optimal nodes placement. The meaning of such optimal nodes placement may be rather different, depending on the aspects of the network that are to be emphasized. Some examples are the visualization of groupings or clusters, the emphasis on edges that disconnect the graph or the identification of nodes with few connections, among others. There are many examples, where the detection of groupings provides a notorious contribution to the comprehension of the phenomenon being studied (Mo, Cao, & Wang, 2012; Sarmah, Kalita, & Bhattacharyya, 2011; Zimmermann, Ntoutsi, Siddiqui, Spiliopoulou, & Kriegel, 2012). Most evident examples of this kind of problems are the organizational networks, social networks, bibliographic references networks and interest groups among many others. Configurations of nodes and connections also occur in a great diversity of other situations. They may represent physical networks, such as electrical circuits, roadways, or organic molecules. They are also used when representing less tangible interactions as might occur in ecosystems, sociological relationships, databases, or in the control flow in a computer program (Gross & Yellen, 2003, 2006; Kamada & Kawai, 1989).

Structural Approaches in Requirement Engineering

Several activities in different Requirements Engineering processes, take advantage of the use of graphs. For example, the requirements prioritization method proposed by Duan, Laurent, Cleland-Huang and Kwiatkowski (2009), analyzes documents of the Requirements Engineering process by means of detecting clusters of requirements. In Duan proposal, the requirements are included in clusters by means of an iterative algorithm. Those clusters are object of negotiations among the stakeholders to assign priorities. Requirements inherit the priority assigned to the cluster to which they belong.

In the specific case of the process in which this project was rooted, graph depicting may be used almost everywhere. This is because natural language models allow drawing many links from one model to another; but these links very soon will become almost useless even in medium sized projects. In other words, automatic graph construction may be useful along the whole process. However, the best results have been obtained from the graphs created from the LEL. The LEL model records the vocabulary used in the application domain. It describes the words and phrases used by clients and users with a meaning specific to the application domain.

Surveying the LEL from a structural point of view, it may be immediately seen that it can be depicted using a graph, where the vertices symbolize the LEL terms and the edges represent the hypertext links among them. From this point of view, the LEL may be visualized as a sort of linguistic network with a complex structure; in such a way that besides the explicit information stored in each symbol, there is implicit information stored in the relationships among symbols. To build and to analyze the graph of