Implicit knowledge extraction and structuration from electrical diagrams

Ikram Chraibi Kaadoud, Nicolas P. Rougier, Frédéric Alexandre

Implicit knowledge extraction and structuration from electrical diagrams

Ikram Chraibi Kaadoud[1,2,3,4], Nicolas Rougier[2,3,4], and Frederic Alexandre[2,3,4]

(1) Algo'Tech Informatique, Technopole Izarbel, Bidart (2) INRIA Bordeaux Sud-Ouest, Talence, France (3) LaBRI, UMR 5800, CNRS, Talence, France (4) IMN, UMR 5293, CNRS, U Bordeaux, Bordeaux, France
{ikram.chraibi-kaadoud,nicolas.rougier,frederic.alexandre}@inria.fr

Abstract. The electrical domain, either domestic or industrial, benefits from a huge set of well-defined norms at both the national and international levels. However and surprisingly enough, there is no such norm regarding the actual conception and structuration of electrical diagrams, even though the basic symbols and notations remain the same. Each company is actually free to design such diagram relative to its own experience, expertise and know-how. The difficulty is that such diagrams, which are most of the time materialized as a PDF booklet, do not reflect this implicit knowledge. In this paper, we introduce our work on the extraction and the structuration of such knowledge using ad-hoc graph and text analysis as well as clustering techniques. Starting from a set of raw documents, we propose an end-to-end solution that offers a company dependent structured view, of any electrical diagram.

Keywords: Knowledge extraction, knowledge representation, hierarchical clustering, knowledge structuration, electrical diagrams

1 Introduction

The electrical domain, either domestic or industrial, benefits from a huge set of well defined norms at both the national and international levels. These norms have to be enforced when designing a new electrical diagram and during the actual physical construction of the circuit. In that regards, electrical workers have generally to deal with a booklet describing the overall circuit splited into several subparts, each one fitting on a regular A4 sheet paper. This arbitrary and constrained segmentation of the whole diagram implies a graph structure that does not follow the logical structure of the underlying object. It is indeed quite similar to the decomposition of an image into pixels sharing no similarities with the inner structure of the image. The problem we are interested in this paper is to find original methods for discovering the implicit structure of electrical diagrams using the PDF description. In this context, we are working with a software company, Algo'Tech Informatique, that provides a set of specialized tools for the design of electrical diagrams. These tools are able to provide a synthetic view of any electrical diagram and to produce the corresponding booklet. To do so,

whenever Algo'Tech Informatique has a new customer, it gets its existing electrical diagrams to analyse them and then adapt the database and the softwares according them. Hence, Algo'Tech informatique experts face two challenges : the first one is to understand the reasoning and planning process of electrical diagrams by their customers in order to formalize it. The second challenge is to make the customer's experts explain their work by supporting them during the exploration of their own habits and knowledge. Besides, since each customer designs freely diagrams according its own experience, expertise and know-how, diagrams do not reflect the implicit knowledge. And yet this knowledge is important, since it holds the footprint of the customers. These ones are indeed attached to their way of drawing and disposing electrical components inside a circuit (they are able to recognize the work of their company whenever they see it) and they are attached to continue to have it once they used Algo'Tech Informatique softwares. So, to extract the "footprints", Algo'Tech Informatique has to review the internal arrangement of electrical components into each page of each electrical diagram and to compare the arrangements between them to find similarities. This work results from an extensive and non-automated collaboration between the software company and the customer. It is also time-consuming, informal (and not stricly replicable) and complex (documents mix texts and diagrams) in the sense that it results from non constrained interviews between humans. Consequently, we aim for an automatic knowledge extraction process specific to a customer and based on its past projects (generally materialized as a restricted set of scanned PDF booklets).

2 Review of existing work

2.1 Technical documents analysis

The capitalization of the knowledge of technical documents is a real problem in the industry. From one hand, the field of analysis of technical documents, which has been investigated for several years in many kinds of domains is mostly focused on image analysis and recognition [1]. On the other hand, there is a great deal of work in the field of document analysis, including PDFs, in order to retrieve the text content, but again, that leaves aside the graphical content [2]. However in the domain of technical diagrams, both of these elements, text content and image content, are important for the understanding and the interpretation of documents. Another neglected field is the diagram interpretation one. The identification of the structure in a diagram, the semantics of its constituents and their relationship, is almost always domain-specific which make global approach difficult. But, recently, [3] focuses on the problem of diagram interpretation and reasoning by exploiting a deep learning algorithm that learns on a basis of 5,000 diagrams, one order of magnitude bigger than our 160 PDF booklets that we are currently analysing (plus the 500 ones that we would like to test). These reasons makes us to look for an innovative solution to analyse small volume of electrical diagrams, without using standard methods.

2.2 Expertise extraction

Another problem that industry faces is the expertise capitalization. Defined as an intellectual capital resulting from the knowledge and the experience from the collaborators and/or the experts, the expertise is represented implicitly through several documents. Using many different methods, companies try to collect and disseminate such expertise to all the employees, so that they can take advantage from it. But even then, the challenge is still complex because of the experts mental representation of their work. Indeed [4] confirmed the existence of difference between senior experts and beginner experts mental representation. The first ones have a conceptualization rather of gist-type, more fuzzy, conceptual, intuitive, using prototypes and high level concepts. In the opposite, beginner experts have a rather verbatim-type conceptualization, which means detailed, analytical, controlled, using low-level concepts.

There is thus different ways to describe diagrams : through the explicit structured data and through the implicit structure or mental representation of the experts. In the following section, we describe the knowledge extraction of raw data using the titles and text into electrical diagrams (Figure 1, label A, B, C). From this point, we will refer to Algo'Tech Informatique experts as experts, and we will precise in other cases.

3 Raw data : Electrical diagrams in PDFs and DXFs files

Companies that draw electrical diagrams have to deal mostly with two type of files : the PDF ones, the most commonly used type, and the Autodesk's Drawing Exchange Format (DXF). PDFs mix texts and diagrams and possess their own local structure : Usually the PDF document has a title, a list of pages, and thus, the circuit that goes through many pages, as in Figure 1. In each folio, there are four important elements: its title, its number (different from the page number), the electrical components (and voltage indications) and its indication to other folios. These last ones are the link that make the rebuild of the whole diagram possible. A DXF file, in the other hand, is a Computer-aided design (CAD) data file that enables data interoperability between many CAD softwares. It is considered an open access format because it is basicaly an ASCII file that can be read by text editors. Considered as an ASCII translation of a drawing file, a DXF file, allow, by analysing the blocks and their attributes, to extract data and meta-data related to electrical components symbols, in particular coordinates, attributes and block name. It permits thus to avoid computer vision or image recognition on the PDF files. Hence, since PDF booklets give a graphical view of electrical diagrams using symbols and texts, and DXF files contain more technical data about the layout of the diagram into pages, the alliance of their analysis enables a quasi-complete analysis of electrical diagram and a more exhaustive interpretation.

4 1^{st} approach : From raw data to concepts

We present in this subsection a first approach that allow us to confirm hypothesise regarding our data before to go further : for a given customer, just by studying his electrical diagrams (PDF mainly), it is possible to put in light a structure and dominant concepts that has a business meaning for him. We wanted to get a global view of each electrical diagram to get a global idea about the struture. For that, we transformed each diagram into graphs by mimicking the expert's reasoning. We used the following process : for each folio of each diagram, we use the folio's number (Figure 1, label B) as the id of a node and the related folio's title (Figure 1, label A) as the label of that node. Then, we use the folio's indication (Figure 1, label C) to establish links between nodes. We thus manage to transform a linear booklet of many pages, into one single representation holding in one sheet. This first step make us realize redundancy in folios names, so we extracted all the titles all booklets combined, to analyze them and compute an intersection between them. We obtain thus a set of common words to all booklets that we will refer to as the concepts of the customer from this point. These words were used as a replacement for the nodes labels, which results in an homogenized set of graphs. Finally, all the simplified graphs were merged into one global graph of concepts by aggregating links and nodes gradually. We thus, obtain a single view of the customer business logic grouping all the concepts that the customer have already used in his previous electrical diagrams, and all the possible relations between these concepts. The final graph is, hence, a representation of the knowledge of the customer. Through this preliminary work, we manage to do 2 things : First, we managed to group folios according their content. Indeed, by doing the reverse path (from the abstract view to the raw data), it is possible to group and analyse folios according to the concept they belong to. This capacity to group folios according to the concept will be used for the next section. Second, we structured raw data into a view of the customer's knowledge that put in light concepts and relations between them. By doing this, we confirm the existence of redundancy in the work or electrical diagrams designers for a given designer, at least at the global level. In the next section, we chose to focus on the layout of electrical components (Figure 1, label D) into folios, in order to confirm the existence of redundancy, and by extend a footprint in the drawing.

5 2^{nd} approach : Footprint extraction using datamining

This second approach is based on a second hypothesis : each electrical diagram is an instanciation of the customer knowledge and by studying it, it is possible to extract it. We thus based our work on the work of [5] and more recently [6], that both defined the KDD process as 3 steps : a pre-processing step, a data mining step and a post-processing step.

Preprocessing Two treatments are done : the creation of data sets and their transformation. Considered data are sequences : a set of electrical components

extracted from each folio and ordered according their growing x, y coordinates. We created two data sets : the first one group sequences according the concept they belong to, whereas the second gather together all sequences independently of the concepts they belong to. Finally, each sequence is translated into a binary vector of 25 units (since 25 families of electrical components exists according experts), each indicating the presence or absence of a family into a folio.

Data mining Once the data prepared, we applied a Hierarchical clustering (HC) method on the two sets in order, from one hand, to study and compare the clusters that appear, and from the other hand, to put in light the footprint into the sequences. We get inspired by the work of [7], who use the HC analysis to rebuild the grammar (set of rules and patterns) from sequences. For further technical details on this method, please refer to [8]. On each set of data, the algorithm uses the Jaccard distance and average linkage method as parameters, since it was the combination that shows the better Cophenetic correlation coefficient. For every sets and subsets, it results in a dendogram (the standard representation in HC) and a cluster-tree (a graph).

Post-Processing For every sets of vectors, each cluster obtained is described as a binary vector corresponding to the union of the binary vectors of the previous clusters. This work of identification of clusters makes possible the comparison between the clusters obtained with both data sets : Exception apart (folios which weren't plot at the expected location), all the cluster-trees of the first data set were found in the cluster-tree of the second data set. Thus, besides this approach allows to put in light particular cases or mistakes in an electrical diagram, it also shows that there is indeed recurrence and regularities into a cutomer's working way : it then confirms the existence of a footprint.

6 Discussion

In this paper, we exposed two approaches we developed for knowledge extraction and representation in the field of electrical diagrams : the first one creates a global view of the customer business logic using graphs, on the basis of extracted text. The second approach spells implicit knowledge out from the internal arrangments of all the folios electrical components, using clustering analysis. It detects and extracts patterns, that represent customers footprints, from data sets. These two approaches complement each other in order to give a view of the electrical diagrams that assist, improve and accelerate experts analysis. In a more global dimension, this work shows that each electrical diagram is an instantiation of the customer knowledge from which it is possible to reconstruct the expertise knowledge as well as habits of work. The study of such habits is still an ongoing work. In order to have the full picture, and to provide more assistance to the experts, we are currently studying the electrical components sequences according the wiring cables with the hypothesis that such sequences held more information about the working rules and habits for a given customer. To explore this, we propose a neural network approach in order to discover the "grammar" rules governing the sequence arrangement : the Elman model [7]. Our preliminary

results indicate that it is possible to learn and to predict as long as the sequence length is not too large. Tests and analysis are still on going and alternatives, like other recurrent neural networks, are also considered. On the medium term, by improving the expertise analysis and by extending the tools that assist the designer in her or his work, we aim at showing that there is an implicit dimension in the planning of every electrical diagrams as it was shown in [9].

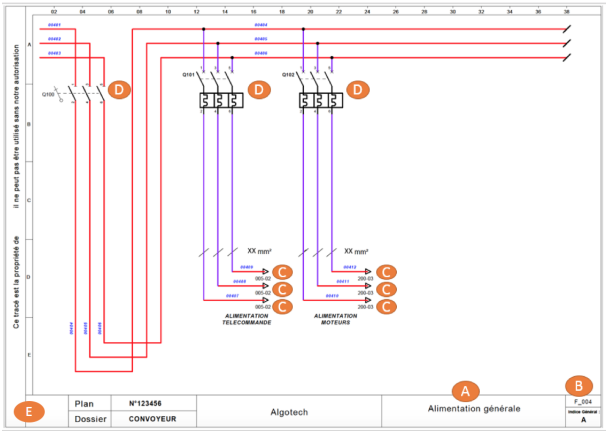

Fig. 1. The organisation and content of a folio : A- the folio's title, B- the folio's number, C- the folio's indications, D- the electrical components, E- the caption block, a frame surrounding the diagram. Sequences of electrical components according the wiring cables in this folio are : Q100, Q101 and Q100, Q102.

References

1. Antoine, D., Collin, S., Tombre, K.: Analysis of technical documents: The REDRAW system. In: Structured document image analysis. Springer (1992)
2. Futrelle, R.P., Shao, M., Cieslik, C., Grimes, A.E.: Extraction, layout analysis and classification of diagrams in PDF documents. In: ICDAR. vol. 3 (2003)
3. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. arXiv preprint arXiv:1603.07396 (2016)
4. Aimé, X., Charlet, J.: IC: Ingénierie des Connaissances ou Ingénierie du Conformisme? In: 24èmes Journées francophones d'Ingénierie des Connaissances (2013)
5. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996)
6. Ramos, S., Figueiredo, V., Rodrigues, F., Pinheiro, R., Vale, Z.: Knowledge extraction from medium voltage load diagrams to support the definition of electrical tariffs. Engineering Intelligent Systems for Electrical Engineering and Communications 15(3), 143–149 (2007)
7. Servan-Schreiber, D., Cleeremans, A., McClelland, J.L.: Encoding sequential structure in simple recurrent networks. Tech. rep., DTIC Document (1989)
8. Hees, J.: SciPy Hierarchical Clustering and Dendrogram Tutorial (Aug 2015)
9. Cleeremans, A., McClelland, J.L.: Learning the structure of event sequences. Journal of Experimental Psychology: General 120(3), 235 (1991)