HAL
archives-ouvertes.fr

# Implicit knowledge extraction and structuration from electrical diagrams

Ikram Chraibi Kaadoud, Nicolas P. Rougier, Frédéric Alexandre

▶ **To cite this version:**

Ikram Chraibi Kaadoud, Nicolas P. Rougier, Frédéric Alexandre. Implicit knowledge extraction and structuration from electrical diagrams. 2016. hal-01525015v2

**HAL Id: hal-01525015**
**https://hal.inria.fr/hal-01525015v2**

Preprint submitted on 22 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial| 4.0 International License

# 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, Centre National de la Recherche Scientifique, Talence, France
(4) Institut des Maladies Neurodégénératives, UMR 5293, Centre National de la Recherche Scientifique, Université de 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 structured view, which is company dependent, of any electrical diagram for a given company.

**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 to be enforced as well during the actual physical construction of the circuit. In that regards, electrical workers are generally given a booklet describing the overall circuit even though this diagram has to be split into several subparts such that each subpart fits 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 existing electrical diagrams of the customer, to analyse them and adapt the database of their softwares. Furthemore, if customers want to find back their concepts and vocabulary into the new electrical diagrams once they are generated by Algo'Tech informatique's softwares, customers are also mostly interested to find their "footprints", their way of drawing and to dispose electrical components inside a circuit. The customers experts have indeed the ability to recognize the work of their company whenever they see it, and they are attached to continue to have it. Concretely, the difference between the schema of a company A and the schema of a company B, that both represent the same circuit, will lies mainly in the configuration of the electrical components in a sheet. To extract the footprints, Algo'Tech Informatique has to review the internal arrangement of electrical components into each page of each electrical diagram (on PDFs) and to compare the arrangements between them to find similarities. Thus, this work results from an extensive and non-automated collaboration between the software company and the customer. It is also time-consuming and informal (and not stricly replicable) in the sense that it results from non constrained interviews between humans. Hence, to get this company-specific synthetic view, 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, which is often the case with implicit knowledge. So, the work of expertise extraction is a difficult work since it needs to support experts in the exploration of their own habits and knowledge. It is also often the opportunity, during this process, to realise a review of the expertise by itself, by validating or not some actions experts used to do and to put in place new processes. Consequently, we aim for an automatic knowledge extraction process specific to a customer and based on its past projects (which are generally materialized as a restricted set of scanned PDF booklets). However, such documents mix texts and diagrams and this is what makes them difficult to analyze even though they also possess their own local structure that we can take advantage of as we will explain in this article. Our goal on the long term is to help Algo'Tech company in its work of customers expertise extraction and formalization inside its own softwares that will assist the designer in her or his work of designing new diagrams.

In this paper we propose two approaches for knowledge extraction and representation in order to describe in a more explicit way electrical diagrams with respect to the work and drawing habits of the owner company, but also with respect to the mental representation of the technical diagram designer. In this regards, concretely, we first suggest, in the following section a review of existing works in technical diagrams exploration. Then, in section 3, we describe the knowledge extraction process we developed to transform every raw data stuck

into the PDF booklets, into explicit, usable and relevant information. Then, in section 4, we detail a content approach related to the electrical components, that aims to put in light the footprints of customers using a data mining solution. We conclude by discussing our approaches and the next steps of our work.

## 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 like mechanical engineering, city maps, hand-drawn figures, geographical maps and electrical circuitry, is mostly focused on image analysis and recognition [1]. On the other hand, there is a great deal of works in the field of document analysis, including PDFs, in order to retrieve the text content from these documents, but again, that leaves aside the graphical content of these documents [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 field that didn't receive a lot of attention (compared to symbol recognition by computer vision for example) is the field of diagram interpretation. The identification of the structure in a diagram, the semantics of its constituents and their relationship, is almost always domain-specific. Recently, [3] focuses on the problem of diagram interpretation and reasoning, also defined as syntactic parsing (detection of the constituents and their syntaxic relationships in a diagram) and semantic interpretation (tasks of mapping constituents and their relationships to semantic entities or real-word concepts) in order to propose a global approach of the diagrams, regardless of the domain. However, the approach benefits and actually exploits a huge dataset. More precisely, it exploits a deep learning algorithm that learns on a basis of 5,000 diagrams, a size 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. This is the main reason that 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 capitalization of the knowledge expertise. Such knowledge is represented implicitly through several documents, and represents an intellectual capital resulting from the knowledge and the experience from the collaborators and/or the experts. 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]

demonstrated the existence of expert-novice differences in mental representation using a series of drawing tasks. [4] confirmed that experts represent given information in a domain-specific manner, concerned with the deep semantic structure of that information, whereas novices, in their mental representations, focus upon superficial domain-general aspects. These results were confirmed by [5] who also asserted that for one given field, the novice experts have a rather verbatim-type conceptualization, which means detailed, analytical, controlled, low-level concepts, whereas senior experts have a conceptualization rather of gist-type, more fuzzy, conceptual, intuitive, using prototypes and high level concepts. This means that according to the level of expertise, the mental representation of a technical diagram is not the same. For example, if a novice sees geometric forms on a paper sheet, he will recognize shapes like "square" or "triangle", whereas an expert may describe them as series of pixels (or tiny dots) of varying shades arranged to form the image of a "square" or a "triangle". 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 First approach : Extraction and representation of raw data into concepts

The knowledge discovery of data (KDD) process was described by [6] as the process of discovering useful knowledge from data. Iterative and interactive, it involves many steps with many users decisions. In the present work, we chose to follow those steps with one specific goal : to put in light the implicit structure hidden in different kinds of electrical diagrams.

### 3.1 Raw data : the 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 DXF format ones, the only format with open access that can be handled by almost every Computer-aided design (CAD) system [7]. In this section, we describe how we process these kinds of files.

**Electrical diagrams on PDFs files** According to the complexity of the circuit represented in the diagram, it can take hundreds of pages to represent only one electrical diagram. Usually the document has a title, a list of pages, and thus, the circuit that goes through many pages, as in Figure 1. In each of them (named "folio") there are three important elements: First, the schema (the electrical diagram), second, the equipment and the electrical components (symbols of motors, power, security, etc.) and third, the text (folio's title, its

number (different from the page number), the equipment names, voltage indications and indications toward other folios). In the majority of the folios, there are indications to other folios: these pieces of information are important because they are the relations between different folios according to the wiring cables. For example, on the power equipment folio, there will be the number of the motor equipment folio and conversely. Hence, from the first folio containing the start of the schema and by following the folios indications on every folio, it is possible to extract the electrical diagram. In this paper, we chose to focus on the exploitation of the text extracted from an electrical diagram, because we want to explore the semantic value of text for knowledge representation and reasoning process. The PDF files are important : it is by using the folios indications, the folios titles and the folios number, that we realise the first approach described in the current section.

**Electrical diagrams on DXFs files** DXF, for Drawing Interchange Format, or Drawing Exchange Format, is a 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 and that is organized by sections (HEADER, BLOCKS, ENTITIES, etc.). A DXF file is thus an ASCII translation of a drawing file : by analysing the blocks and their attributes, it is possible to extract data and meta-data (coordinates, attributes, block name, etc.) related to electrical components. If the extracted text from a folio of a PDF allows us to get the name of an electrical component for example, the DXF processing allows us to get its x, y coordinates inside that folio. It is this information that allows us to process the content approach that we are going to describe in section 4. Hence, the alliance of PDF analysis and DXF analysis enables a quasi-complete analysis of electrical diagram and a more exhaustive interpretation without having to deal with computer vision or image recognition techniques and issues.

### 3.2 Knowledge extraction and representation process using graphs

We propose in this subsection different ways to represent an electrical diagram. If we consider a list of folios describing a wiring diagram, it indicates a linear approach of the diagram (one folio by one folio). But in reality, the electrical diagram is more complex than this : there are many cable pathways. So the review and analysis of previous electrical diagrams of one customer, which is mainly manual, can take a lot of time and may generate misinterpretations. We present below the approach that we chose to enable a quick browsing of the electrical diagrams in order to get the concepts of the customer's job, its ways of drawing but also a quick detection of particular cases or mistakes. Among all the texts extracted from electrical diagrams, we focused on the folios title, the folios number and the folios indications (Figure 1, label A and B). In the field of electrical diagrams, there are global standards for symbols that represent components and inside each company, internal standards about the naming. So technical diagram designers have to follow explicit conventions to name folios. We thus chose to analyse these elements as a starting point of our KDD process.

**Transform a diagram into a graph** While reading an electrical diagram, the expert has to follow the folio's indication (Figure 1, label C) in order to navigate between folios. In that regard, the first challenge is to sail from an electrical circuit portion to another in order to build a first, global and abstract representation of the electrical diagram. Since PDF booklets have different size, this work can be very complex. Our first work was to transform a linear booklet of many pages, into one single representation holding in one sheet. Mimicking the expert's reasoning, we used the following process to transform any electrical diagram into a graph : for each folio of each diagram, we use the folio's number as the id of a node and the related folio's title as the label of that node. Then, we use the folio's indication to establish links between nodes. We obtain thus a graph representing the electrical diagram with a first level of abstraction.

**Get homogenized graphs** After the previous step applied on several booklets, we obtain a set of graphs, each one representing one diagram, with its own titles. Then, we worked on the titles in order to replace the nodes labels by a limited set of words in order to get homogenized graphs from the vocabulary point of view. For that, we analysed the titles to put in light the common words. So, we extract all the words in the titles all booklets combined and compute an intersection between them. In this way, we obtain a set of words that are common to all booklets. We will refer to these words as the concepts of the customer : they are the concepts that are recurrent and present in all the projects of a given customer. These concepts are thus essential to the business logic of the customer. A collaboration with experts at this step is necessary to control the concepts, and validate them. Once it is done, these concepts replace the original nodes labels (the folio's titles). Nodes having the same label get merged (but by preserving the links) in order to have, at the end, one occurence of each label in each graph. In this way, all the links are preserved, but the number of nodes have been reduced, and labels of the nodes have been standardized. So the complexity of the graphs has been reduced. We thus obtain, a first local representation of the knowledge present in each electrical diagram through homogenized graphs.

**Getting the global graph of customer's concepts** The main goal for experts while working and going into the analysis of each diagram is to understand and extract the customer business expertise : Concepts, rules, standards, etc. Once we get the set of standardized graphs, one step remains : making a single view of the customer business logic. Since our graphs have the same labels, we can merge them into one global graph of concepts by aggregating links or nodes gradually. The global graph of concepts contains, thus, at the end of the process, all the concepts that the customer have already used in his previous electrical diagrams, and all the possible relations between these concepts. The relations between nodes represent by their existence, a habit of work of the client : indeed, if there is a link between concepts "Equipment" and "Power", it does not only mean that it is possible from a functionnal point of view, but also that the customer, in his job, at least once put a folio's indication (a link) between a folio that belongs to the power concept and a folio that belongs to the equipment

concept. At this point, the graph is a representation, at an abstract level, of the knowledge of the customer.

In this first part, we proposed a process to go from an electrical diagram, raw data, to an abstract view of the customer's knowledge that put in light concepts and relations between them. This approach allows also to facilitate the groupement of the folios of the same nature since that 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 second approach.

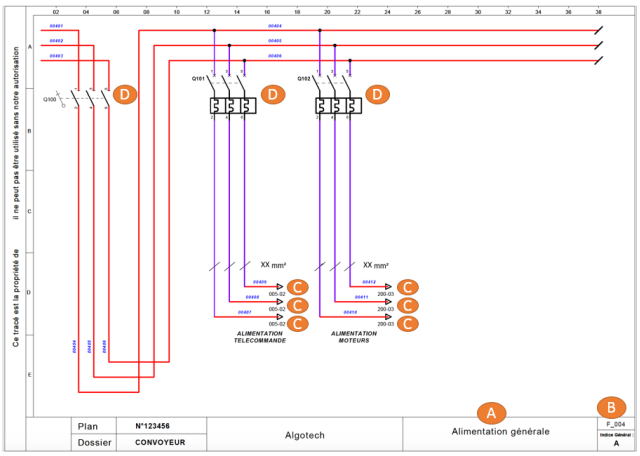


**Fig. 1.** The organisation and content of a folio : A- the folio's title, B- the folio's number, C- the folio's indications, D- the electrical components

## 4 Second approach : Extraction of the footprint with hierarchical clustering

### 4.1 A content approach through the electrical components

The second approach used for the analysis of the electrical diagram is a content approach through the electrical components (Figure 1, label D). DXFs files give us information about electrical components as a symbol drawn into a folio. One important information is the x,y coordinates of the components. So for each folio of each electrical diagram, we extract a set of electrical components in order to obtain as many sets as folios, all booklets combined. Since it is the internal arrangement of components inside a folio that interrest us, we chose to order the components inside each set, according to their growing x, y coordinates. We call an order set a sequence : it is the formalisation of the internal arrangments into a folio. Each sequence describing one folio. For example, in figure 1, the sequence is Q100, Q101, Q102. It is an important notion for our work since it is inside a

sequence that can be found a footprint of a customer. Studying electrical components, made us also exchange with experts about their meaning and functions. Together, we listed 25 families (categories) of electrical components, each one composed by sub-families. These families are quite common in the global field of electrical diagrams: as they have been define once, they can be used for any diagrams in the electrical field. We use the sequences and the list of categories for the building of our data sets and the data mining process described below.

### 4.2 The data mining process : Hierarchical clustering

In his description, [6] puts in light one element regarding the KDD process and the data mining : the latter one is defined as a step into the former one. Data mining is not a KDD process by itself, but the application of a particular algorithm in order to extract patterns from data. Indeed [8] defined the KDD process as 3 steps : a pre-processing step, a data mining step and a post-processing step. Here is, below, our work according to these three steps.

**Preprocessing** Two treatments are done on the sequences : the creation of data sets and their transformation.
First, considering all electrical diagrams, we group together the sequences that belong to the same concept (common word) : it is the first set of data, composed by subsets, one for each concept. Second, considering again all electrical diagrams, we group together all the sequences, independently of the concept they belong to. We thus create the second set of data. Finally, each sequence is translated into a binary vector of 25 units (since we listed 25 families of electrical components), each indicating the presence or absence of an electrical category into a folio.

**Data mining** Once the data prepared, we chose to apply a hierarchical clustering method on the two sets in order, from one hand, to study, observe 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 [9], who use the hierarchical cluster analysis to rebuild the grammar (set of rules and patterns) from sequences. The hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It needs a distance metric (like 'euclidean', 'hamming' or 'jaccard') and a linkage method that compute the distance between two clusters. It starts with n individual samples considered as singleton clusters and at each iteration, it merged the two clusters with the smallest distance until only one remains. To choose the best combination "distance metric-linkage method", it is advised to compute the Cophenetic correlation coefficient. The more its value is close to 1, the more the clustering preserves the original distances. The result can be graphically represented as a dendogram (the standard representation) or as a cluster-tree (a graph) with folios as leafs and nodes as clusters. For further technical details, please refer to [10].
In our work, on each set of data, we use the Jaccard distance and average linkage method as parameters to our hierarchical clustering, since it was the combination that shows the better Cophenetic correlation coefficient. For every sets and subsets, we thus obtain a dendogram. The graphical result of this analysis on the

second set of data (sequences from 3 electrical diagrams) is presented in Figure 2.

**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 allowed to compare the clusters obtained with the second data set, with clusters obtained with the first data set : Exception apart (folios which weren't plot at the expected location), we found in the cluster-tree of the second data set, all the cluster-trees of the first data sets. We thus realized that this approach allowed us to put in light particular cases or mistakes in an electrical diagram, for example, when a folio (its sequence) was ploted at an expected location. On top of that, the fact that we obtained this results shows that there is indeed recurrence and regularities into the way of working of a customer : it confirms the existence of a footprint.

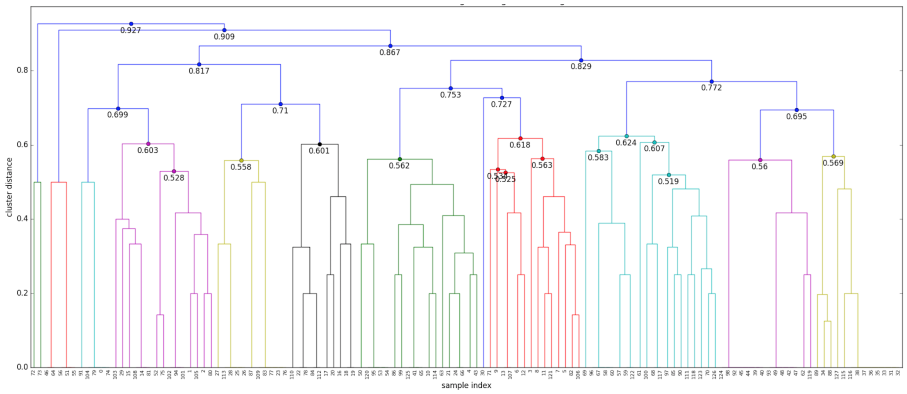


**Fig. 2.** A dendogram, a graphical representation of the hierarchical clustering of the sequences of all electrical diagrams combined : numbers on the x axis identify the sequences, whereas the numbers in the dendogram identify the distance.

## 5 Discussion

In this paper, we exposed two approaches we developped for knowledge extraction and representation in the field of electrical diagrams. We first described the approach that create a global view of the customer business logic using graphs, on the basis text extracted from electrical diagrams. By doing this, from one hand, we showed that the exploitation of text from electrical diagrams allowed us to unravel the concepts that are important for the customer. From the other hand, we provided a graph representation that is accessible for both the experts and for the novices such as external collaborators. The second approach aims to spell implicit knowledge out from the internal arrangments of all the folios electrical components. Using clustering analysis, we managed to detect and extract patterns from sets of data that represent customers footprints. 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, we showed through our work, that each electrical diagram is an instanciation of the customer knowledge from which it is possible 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 with the hypothesis that such sequences held more information about the habits of work for a given customer. Such sequences represents a small part of the global knowledge as well as the drawing habit (the arrangements of components). To explore this, we have used a neural network approach in order to discover the "grammar" rules governing the sequence arrangement : the Elman model [9]. Our first preliminary results indicates it is possible to learn and to predict as long the sequence length is not too big. Tests and analysis are still on going and alternatives, like other recurrent neural networks, are also considered.

On the medium term, as well as improving the expertise analysis and by extend the tools that assists the designer in her or his work, we aim at showing that there is an implicit dimension in the planing in every electrical diagrams as it was shown in [11].

## References

1. Antoine, D., Collin, S., Tombre, K.: Analysis of technical documents: The REDRAW system. In: Structured document image analysis, pp. 385–402. 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, pp. 1007–1014 (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. Lowe, R.K.: Constructing a mental representation from an abstract technical diagram. Learning and Instruction 3(3), 157–179 (1993)
5. Aimé, X., Charlet, J.: IC: Ingénierie des Connaissances ou Ingénierie du Conformisme? In: IC-24èmes Journées francophones d'Ingénierie des Connaissances (2013)
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996)
7. Weber, N., Henrich, A.: Retrieval of technical drawings in DXF format-concepts and problems. In: LWA. pp. 213–220. Citeseer (2007)
8. 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)
9. Servan-Schreiber, D., Cleeremans, A., McClelland, J.L.: Encoding sequential structure in simple recurrent networks. Tech. rep., DTIC Document (1989)
10. Hees, J.: SciPy Hierarchical Clustering and Dendrogram Tutorial (Aug 2015), https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/
11. Cleeremans, A., McClelland, J.L.: Learning the structure of event sequences. Journal of Experimental Psychology: General 120(3), 235 (1991)