Methodology of learning curve analysis for development of incoming material clustering neural network

Boris Onykiy[*], Evheniy Tretyakov[†], Larisa Pronicheva[‡], Ilya Galin[§], Kristina Ionkina[**], Andrey Cherkasskiy[††]
National Research Nuclear University MEPhI, Moscow, Russian Federation
bnonykij@mephi.ru, evheniytretyakov@yandex.ru, lvpronicheva@mephi.ru, ilia.galin@gmail.com, ionkinakristina@gmail.com, aicherkasskij@mephi.ru

Abstract
This paper describes the methodology of learning curve analysis for development of incoming material clustering neural network. This methodology helps to understand deeply the learning curve adequate level and to bring learning curve structure to the relevant one of the thematic scope of incoming materials. The methodology is based on visual analysis and comprises the building of directed graphs in order to identify data templates. As the battlefield for material clustering the Nuclear Infrastructure Development Section (NIDS) of the International Atom Energy Agency (IAEA) is selected as the support from NIDS' experts had been available during the research. Some of the challenges the NIDS faces are data aggregation for Country Nuclear Infrastructure Profiles (CNIP) and data assessment after Nuclear Infrastructure Review Missions (INIR).

Keywords: neural network, material clustering, learning curve, visual analysis

1 Introduction

The Nuclear Infrastructure Development Section (NIDS) works with Member States to improve: understanding of the requirements and obligations essential to implementing nuclear power programmes; and abilities to develop the necessary infrastructure for introducing nuclear power (Onykiy, et al., 2016).

[*] Development of methodology concept
[†] Development of supporting software
[‡] Terminology extraction for the NIDS ontology
[§] Development of the NIDS ontology structure
[**] Development of directed-graph of the NIDS ontology
[††] Analysis of the working documents

Member States that are new for nuclear technologies are called newcomers and if newcomers are intended to get nuclear technologies for peaceful purposes, they have to meet special conditions on every phase and relevant issues in order to develop infrastructure for nuclear power plant. These phases and issues are described in the "Milestones in the development of a national infrastructure for nuclear power" (NG-G 3.1), so called the "Milestone document". The "Milestone document" is also used by newcomers to assess their own development status, and to prioritize their activities towards the development of nuclear infrastructure for nuclear power plant.

Also, the NIDS provides newcomers with special types of Integrated Nuclear Infrastructure Review (INIR) missions to assess infrastructure development and to provide newcomers with guidelines, recommendations and relevant documentation. The assessment of nuclear infrastructure development is performed in compliance with "Evaluation of the Status of National Nuclear Infrastructure Development" (NGT-3.2 Rev.1).

Besides, the experts of NIDS aggregate information from different sources to the Country Nuclear Infrastructure Profile database (CNIP) to be well informed and to keep tracking of the nuclear infrastructure development, thereby experts meet the challenge of structuring large amounts of information.

The development of incoming material clustering neural network is evaluated as one of the solutions towards structuring large amounts of information. Before starting building neural network the development of the ontology was performed in order to prepare learning curve. This paper describes the methodology of learning curve analysis for development of incoming material clustering neural network.

2 Methodology

The methodology enable researcher to identify the completeness of learning curve and completeness of terminology sources. The process of gathering and analyzing data for learning curve comprises several stages:

- identification of the NIDS' working documents, the terminology from the documents are to be included into learning curve.
- ontology structure development, the structure should reflect the structure of the database where incoming materials will be stored that means the database structure has to explain well thematic areas of incoming materials.
- the NIDS' working documents distribution, the working documents have to be distributed among main classes of the developed ontology structure.
- the terminology extraction, the specific terminology from identified working documents has to be extracted and distributed among relevant ontology structure classes due to complete ontology.
- weights calculation, weights reveal the level of relevance of the term to the thematic areas.
- building of ontology chart, ontology chart reveals relations between terminology and ontology classes.
- analyzation of ontology chart, the completeness of ontology and the completeness of working materials.

The IAEA publications were selected as the main source for ontology terminology. These publications are working documents for NIDS and they are used as assessment materials during INIR missions or materials that are used as guidelines for newcomers. The list of working documents is further:

- "Milestones in the development of a national infrastructure for nuclear power" (NG-G 3.1).
- "Managing Human Resources in the Field of Nuclear Energy" (NG-G 2.1).
- "Workforce Planning for New Nuclear Power Programmes" (NG-T 3.10).

- "Responsibilities and Capabilities of a Nuclear Energy Programme Implementing Organization" (NG-T-3.6).
- "Managing Siting Activities for Nuclear Power Plants" (NG-T-3.7).
- "Establishing the Safety Infrastructure for a Nuclear Power Programme" (SSG-16).

The list of these documents was provided by NIDS' experts. The access for their support was available during the research. Also these materials are used in the development of "Competency framework" – the database comprises key activities to be implemented towards development of nuclear infrastructure. The relevant issues for these documents are also provided by NIDS' experts.

The structure for ontology was developed on the basis of "Milestone Approach" that is described in "Milestone document". The "Milestone Approach" identifies phases that newcomers have to reach due to complete nuclear infrastructure development and issues that describes thematic areas of main obstacles during the development of nuclear infrastructure. Totally three phases and nineteen issues identified in the "Milestone Approach" the chart of nodes and edges are presented in the Figure1.

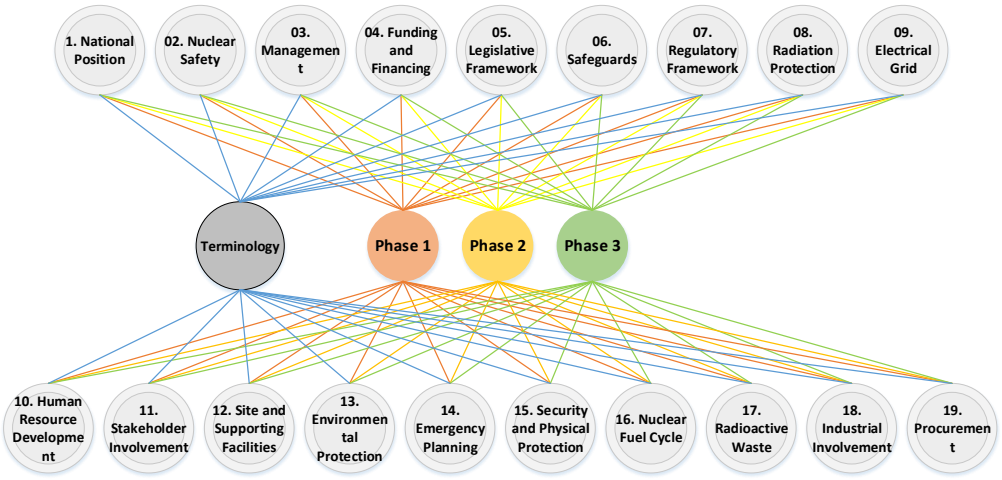

Figure 1: The ontology structure

The extraction of special terminology from working documents was performed manually by young specialists in this field, such practice minimizes the probability of wrong data extraction. One thousand three hundred terms were extracted that were divided into fifty-seven vocabularies – nineteen vocabularies per phase.

The weights were calculated after the terminology extraction. The weights represent level of relevance of the term to the specific issue and issue to the specific phase, mind that every issue is different in every phase (Ananieva, Artamonov, Galin, Tretyakov, & Kshnyakov, 2015). Weights for terminology are calculated as the frequency of term usage in every document relevant to every issue and weights for issues are calculated as sum of relevant terms' weights. The weight of term to issue is

$$W = \sum_{text}^{texts} F(term, text)$$

Formula 2: The weight of singular term relevant to an issue

calculated with the help of Formula 1. The weights of edges between terminology and issues were calculated with help of scripts written in Python 3.5.

The provision for building directed graph are nodes, edges and weights of edges (Artamonov, et al., 2014). Nodes are recognized as classes of the developed ontology: first class – phases, second class – issues and third class – terminology. Edges represent the linkage between classes, in our directed graph they are: terminology to issues and issues to phases.

The graph presented in Figure 2 was built using the Gephi software and algorithm of Yifan Hu (Hu, 2005).

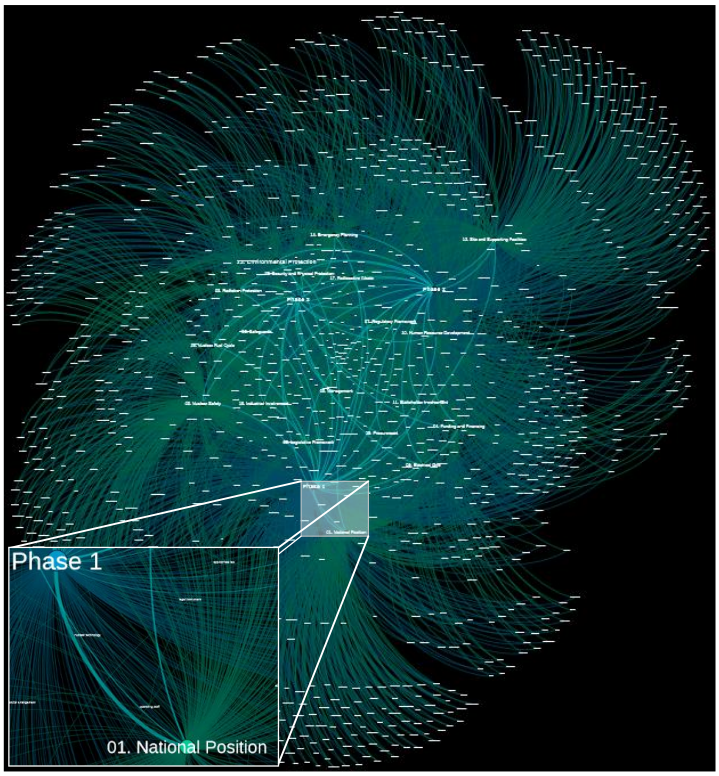

Figure 2: The NIDS ontology graph

The graph shows well the more and the less frequent terminology. The most frequent terms are located in the center and the least ones are on the borders.

Terminology that are located on the borders get high biases in neural network as they are relevant to specific issues. Analyze of weights of the most frequent terms that are located in the center shows the adequate level of learning curve, i.e. the term NEPIO has large relevance to first phase, less to the second phase and the least to the third phase as well as on practice.

3 Results and Discussion

In this subsection we are presenting the experimental results obtained during the development of learning curve for incoming material clustering neural network. The learning curve is represented as ontology with three main classes – phases, nineteen subclasses – issues and terms that take the role of characteristics in the neural network. Totally six specialized documents (listed above) were processed to extract more than 1300 terms. The sample of final data is presented in Table 1.

Term	Issue	Weight
site	Phase 2 Site and Supporting Facilities	147
safety	Phase 1 Nuclear Safety	123
nuclear	Phase 2 Human Resource Development	85

nuclear	Phase 1 National Position	70
regulator	Phase 2 Regulatory Framework	67
nuclear	Phase 1 Nuclear Safety	64
organization	Phase 1 Nuclear Safety	58
nuclear	Phase 3 Human Resource Development	57
safety	Phase 3 Nuclear Safety	55
NEPIO	Phase 1 National Position	51
programme	Phase 2 Human Resource Development	51
training	Phase 2 Human Resource Development	49
criteria	Phase 2 Site and Supporting Facilities	47
nuclear	Phase 2 Radioactive Waste	46
nuclear	Phase 1 Human Resource Development	45
regulatory body	Phase 2 Regulatory Framework	45

Table 1: The sample of data for weights of terms to issues

The analysis of the graph showed the terminology that is critical for incoming materials neural network. This terminology is located on the borders of the graph and points well on relevant issues and phases. The terminology that is in the center of graph is more common but their weights describes the frequency of terms usage that is helpful to identify the adequate level of learning curve.

However, some experts highlighted the fact of incompleteness of ontology that can lead to poor accuracy of the neural network. Probably, the incompleteness of working documents could lead to the incompleteness of the ontology. Nevertheless, this issue is not related to the methodology moreover with the help of methodology this issues became obvious.

Conclusion

The analysis of learning curve is an important process during the whole building of neural network. It helps developers to understand why a neural network works right or wrong. Very important is to get the support from experts in relevant field as they can point out issues in it. The methodology can be applied for various collections of document analysis in organizations in tasks related to identification of hidden relations between working and incoming douments

In these circumstances, it is highly recommended to use visual analysis for learning curve analysis. The methodology presented in this paper proved its utility with learning curve analysis as it delivered a good picture to identify critical terminology, the incompleteness of working documents and adequate level of learning curve.

Acknowledgements

This work was supported by the Competitiveness Program of NRNU “MEPhI”.

References

Ananieva, A. G., Artamonov, A. A., Galin, I. U., Tretyakov, E. S., & Kshnyakov, D. O. (2015, November). Algoritmizatiom of search operations in multiagent information-analytical systems. *Journal of Theoretical and Applied Information Technology, 81*(1), pp. 11-17.

Artamonov, A., Leonov, D., Nikolaev, V., Onykiy, B., Pronicheva, L., Sokolina, K., & Ushmarov, I. (2014). Visualization of semantic relations in multi-agent systems. *Scientific Visualization, 6*(3), 68-76.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American, 304*(5), pp. 35-43.

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal, 10*(1), 37-71.

Onykiy, B., Suslina, A., Ionkina, K., Ananieva, A., Pronicheva, L., Artamonov, A., & Tretyakov, E. (2016). Agent Technologies for Polythematic Organizations Information-Analytical Support. *Procedia Computer Science, 88*, pp. 336-340.