ISSN 1870-4069

Text Mining for Domain Structure Analysis in a Training System for Electrical Procedures

Yasmín Hernández[1], Guillermo Santamaría-Bonfil[1,2], Víctor Pecero[1]

[1] Instituto Nacional de Electricidad y Energías Limpias,
Gerencia de Tecnologías de la Información, Cuernavaca,
Mexico

[2] CONACYT-INEEL, Mexico

{yasmin.hernandez, guillermo.santamaria, 57679vpr}@ineel.mx

Abstract. Learning environments themselves constitute a source of knowledge to understand and improve educational settings. We have developed some training systems where instructional content consists of separated electrical procedures stored in text documents. Since, the adequate characterization of the domain is a key factor in learning, we conducted a text mining study to understand relationships between topics, tasks and procedures in the training systems. In turn, this knowledge can help to provide students with an adaptive training, to help instructors to plan courses and to improve the training systems. We rely on a machine learning approach applying bag of words model and a clustering algorithm. Results on mining partial instructional content are encouraging since they show some useful relationships. Here, general proposal and preliminary results are presented.

Keywords: Bag of words, clustering algorithms, educational data mining, text mining, training system.

1 Introduction

The advancement of information technologies and the growing usage of educational environments, such as e-learning systems, intelligent learning environments, and massive online open courses, among others, have produced an amazing volume of data. The Educational Data Mining (EDM) is focused in processing such data to provide information about aspects and elements included in the learning process, for instance, student-system interaction, performance of the learning environments, and learning process itself. This emerging field exploits statistical, machine-learning and data mining algorithms [1], and it is defined as an emerging discipline, concerned with development of methods for exploring the unique types of data that come from educational settings, and using those methods to understand students, and the settings in which they learn [2].

On the other hand, the electrical field requires efficient training in order to minimize training time, costs, and equipment damage, and most important, to prevent accidents that could injure electricians. However, training on electrical procedures faces some problematic situations, such as limited opportunity to practice in an actual installation, since they are operating on a regular basis. To cope with this situation, we have developed several training systems based on virtual reality to support traditional training and also to allow distance training. Trainees attend classroom courses and have field practice and they complement their learning and practice aided by the training systems. This blended training is guided by instructors [3].

The systems provide training on a number of electrical procedures which are composed by a number of steps, and in turn each step is composed by a number of sub-steps. Among other functionalities, the electrical procedures are explained by means of text describing the whole procedure in virtual environments.

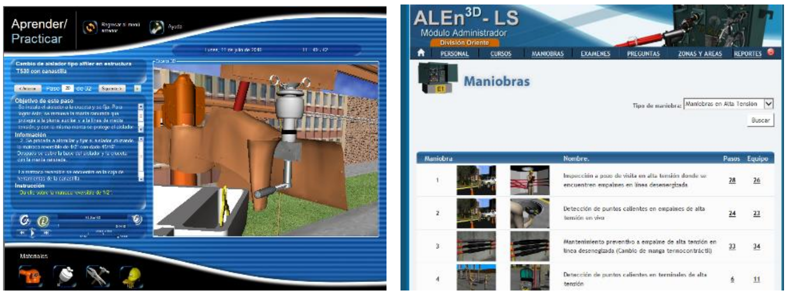
The successful of the learning environments depends to a large degree in the adequate characterization of the domain, which allows us to understand the relationships between topics, tasks and procedures. In turn, this knowledge can help to provide students with an adaptive training, to help instructors to plan courses and to improve the training systems.

We conducted a study to understand the similarity among the different electrical procedures. We want to analyze the domain structure by applying EDM techniques to lessons in the training systems. We rely on the Bag of Words model [4] which consists in a statistical processing of the text and a machine learning algorithm, we used *K-means* clustering algorithm [5].

Here we describe our general proposal and present preliminary results. The rest of the paper is organized as follows: Section 2 presents the electrical training systems. Section 3 describes our proposal to analyze structure domain. Section 4 presents the results of the initial evaluation of the student model. Finally, conclusions and future work are presented in section 5.

2 Training Systems

We have developed several systems for training on different electrical topics [3, 6-9]. Fig. 1 shows some screenshots of these systems. They include a number of electrical procedures. Each procedure is composed by a different number of steps, and in turn each step is composed by a different number of sub-steps. Electrical procedures are taught by means of text explanations. In Table 1, a step and three sub-steps are shown (fragment), in each sub-step there are a description and an instruction; the instruction is an action to be executed by the trainee. As the system is in Spanish, texts have been translated to English. In addition, every procedure refers to handbooks and technical reports. The systems keep track of trainees' progress, however we are still working on integrating some intelligence, so such a way they are able to keep a trainee model of trainees and to respond in consequence [10].



Maniobra	Nombre.	Pasos	Equipo
1	Inspección a pozo de visita en alta tensión donde se encuentren empalmes en línea desenergizada	28	26
2	Detección de puntos calientes en empalmes de alta tensión en vivo	24	23
3	Mantenimiento preventivo a empalme de alta tensión en línea desenegizada (Cambio de manga termocontráctil)	33	34
4	Detección de puntos calientes en terminales de alta tensión	6	11

Fig. 1. Training systems for different electrical domains. Left: medium tension power lines maintenance, right: underground lines.

Table 1. Example of a step of the electrical procedure "Structure conversion from TS30 to VS30 using a platform". Only four sub-steps are shown.

Procedure: Structure conversion from TS30 to VS30 using a platform

Step 35: Place the last insulator, remove the medium voltage line from the temporary conductor holder and rotate the platform

	Sub-step	Instruction
1	Place the new pin insulator and screw it to the crossarm	Select the 13PC pin insulator from menu of materials
2	The lineman places the pin insulator in the crossarm. The isolator is previously climbed up using the errand bucket	Click on the 13PC pin insulator
3	Proceed to screw and fix the insulator using the 1/2" reversible ratchet with a 15/16" socket. Then the insulator base and the crossarm are covered back with the rubber blanket	Click on the 1/2" reversible ratchet

The electrical knowledge is represented as a tree structure (Fig. 2). However, this representation does not include any relationship like dependence or hierarchy between topics, material, and tools. For example, if trainee correctly executes the first electrical procedure, we do not have any information about other electrical procedures. For these reason, we want to apply EDM techniques to the text describing the electrical procedures.

3 Model for Domain Structure Analysis

Data mining or Knowledge Discovery in Databases is the field of discovering novel and potentially useful information from large amounts of data [11]. Educational data mining methods are often different from standard data mining methods, due to the need to explicitly account for the multi-level hierarchy and non-independence in educational

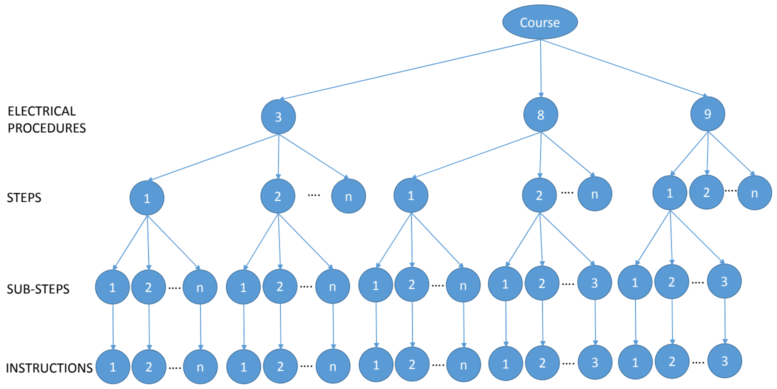


Fig. 2. Knowledge structure for a training course consisting of three electrical procedures composed by steps, sub-steps and instructions.

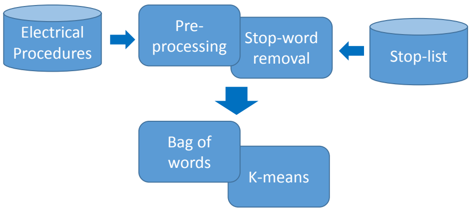


Fig. 3. Proposal for text mining model relying on machine-learning approach using the Bag of Words technique for text-mining.

data [12]. Educational data mining techniques play an important role in augmenting and improving learning environments. Extensive EDM work has been conducted for a wide number of applications: communicating to stakeholders, maintaining and improving courses, generating recommendation, predicting student grades and learning outcomes, student modeling, domain structure analysis [13].

Based on EDM techniques, we built a student model over the performance of trainees in the training systems [10]. Now, we want to analyze and improve the instructional material by discovering similarities and relationships between topics, procedures, tools, etc. We have relied on machine-learning approach using the well-known Bag of Words technique for text-mining.

3.1 Bag of Words Representation

In the *bag of words* (BOW) model, a text (sentence or document) is represented as a feature vector. This model omits grammar and word order, and it is interested in the number of words occurrences within the text [4].

Our initial proposal consists in preprocessing the text to convert it into a bag of words representation and apply *k-means* clustering algorithm. The pre-processing includes to

Table 2. Sample of 4 procedures titles in English and Spanish.

Procedure	Procedure title	
1	Rescue of injured lineman	Rescate de liniero accidentado
2	Installation of stirrups in structure TS30 with platform	Instalación de estribos en estructura TS30 con plataforma
3	Change of pin insulator in the PS30 structure with platform	Cambio de aislador tipo alfiler en estructura PS30 con plataforma
4	Change of fuse cutout (FCO) in structure TS30/RD3	Cambio de cortacircuito fusible (CCF) en estructura TS30/RD3

remove stop-words since they do not provide us with relevant information. This proposal is presented in Fig. 3.

As we mentioned, each training system includes a different number of electrical procedures and every procedure is different from the others as they contain different numbers of steps and sub-steps. As a first attempt, we decided to apply EDM only to the 43 titles of the electrical procedures in only one training system. In Table 2, we show an excerpt of the source text.

To obtain the BOW representation, for each procedure the text was converted in lists of words. For example, for the second procedure we have the list: [“*Instalación*”, “*de*”, “*estribos*”, “*en*”, “*estructura*”, “*TS30*”, “*con*”, “*plataforma*”]. Then, the stop words were removed in each vector. The stop list consists of the words: [*a*, *con*, *de*, *en*, *un*, *una*, *por*, *y*]. For the same example, we obtained the vector: [“*Instalación*”, “*estribos*”, “*estructura*”, “*TS30*”, “*plataforma*”].

After that, a list with all the different word in all the lists is constructed. This list is composed by 67 words and it is the bag of words representation.

After transforming the text into a bag of words, various measures can be calculated to characterize the text. We use term frequency which is the most common type of characteristics, or features calculated from BOW model. Term frequency refers to the number of times a term appears in the text. For our case, electrical procedures titles, we constructed 43 lists to record the term frequencies of all the distinct words.

These 43 lists of 46 elements, can be seen as a matrix of 46x43, an extract is shown in Table 3. The columns represent the words and the rows represent the procedures.

3.2 Clustering

We obtained 67 different words in the 43 electrical procedures, this is a grouping problem. However, as it has been argued, unlike in classification problems, in data grouping or cluster analysis we are not interested in modeling relationships between set of multivariate data and a certain set of outcomes, but we intent to discover and model the groups in which data are often clustered, according to some similarity measure. It is required unsupervised learning techniques to approach this problem, in turn this problem more challenging as the input space dimensionality increases and, as a result, data become more sparsely represented [14].

Tab ency
mea

Ele Pro e Id	S30
1	0
2	0
3	1
4	0

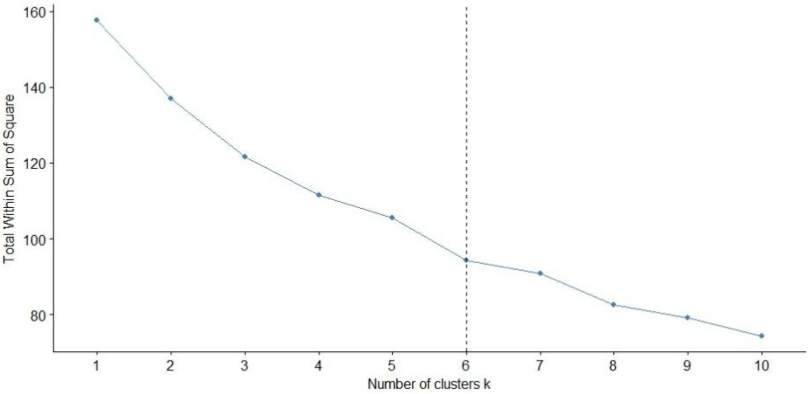


Fig. 4. Result elbow model which does not shows a clear elbow.

We used *k-means* clustering algorithm which allows to group a set of objects in such a way that objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. *K-means* clustering aims to divide *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [5].

Another situation is concerned with how many clusters we need for an adequate grouping. We use the Elbow method [15], which is a graphical device to select the *optimal* number of clusters. To determine the number of clusters, an iterative clustering procedure where *k* is increased by *1* is performed. For each iteration, the within-cluster sum of square is calculated. This is the sum of square distances between each element of the cluster and its centroid.

4 Results

After applying the machine learning approach for text mining, we have the following results. Base on the the elbow model, we determined that 6 clusters are required (Fig. 4) by the vertical dashed line. This is arbitrary since there is no a clear elbow in the graphics, and more clusters could be used. The six clusters are described in Table 4.

In Fig. 5 we present a snake plot of the 6 clusters centroids for the electrical procedures characterized as a bag of words. Centroids ranges from *C1* to *C6*. *X-axis* corresponds to the 67 dimensions of the bag of words, and *Y-axis* corresponds to frequency values. As can be observed, there are a few words that are very distinctive of their corresponding clusters, for instance *installation* for *C1* or *AD30* for *C5*. On the other hand, there is a large overlap on the centroids values particularly for the words *structure*, *change*, and *bucket truck* which appear in most of the clusters. Nevertheless, it can be appreciated that by including less frequent words clusters become more distinguishable.

Table 4. Resultant clusters from the *k-means* clustering algorithm.

Cluster	Size	Description
1	7	This cluster is characterized by the word *installation*. It contains several types of installation procedures. This feature is highly distinctive for this cluster since no other cluster contains it.
2	3	This cluster is characterized by words such as *structure*, *TS30*, *isolator*, and *electric pole*. Thus, it contains the procedures related to the replacement of isolators or electric poles for TS30 structures.
3	15	This is the largest cluster and is characterized by words such as *structure*, *change*, and *bucket truck*. It also includes several types of different structures defined by composite words such as *VS30/RD3*. Consequently, this cluster is related to procedures of change of several types of structures using a bucket truck.
4	7	This cluster is characterized by words *conversion*, *structure*, *TS30*. It contains all tasks which are related to the conversion of structure TS30 to other types of structures.
5	4	This cluster is characterized by words *structure*, *conversion*, *bucket truck*, *AD30*. Thus, it contains tasks that are related to structure conversions, from and to, AD30, using a bucket truck.
6	7	This is the most impure cluster: there is a sub set (4 tasks) which is related to changes of damaged structures using two platforms; however, it also contains tasks with very infrequent words like *injured*, *lineman*, *branch* or *fault*.

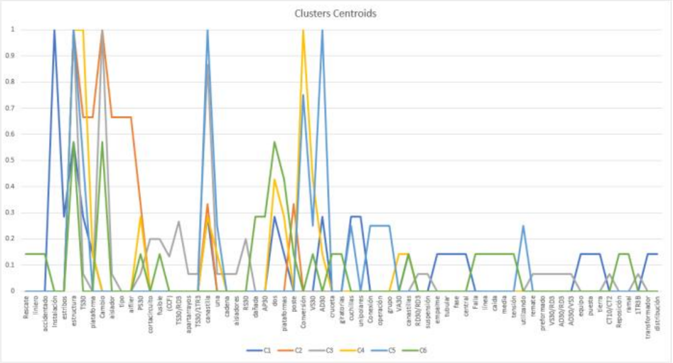


Fig. 5. Snake plot of the centroids for 6 clusters of procedures characterized as a bag of words where X-axis corresponds to the 67 dimensions of BOW, and Y-axis corresponds to frequency values.

Although this clustering separates electrical procedures into *reasonable* clusters, we can also observe that procedures descriptions provide a very limited characterization of the content. This is evident in the large overlap of the clusters in centroid's plot (Fig. 5), and the common words among the clusters (i.e. *structure*, *bucket truck*, *change*). Furthermore, cluster 6 might be divided into two distinct clusters, one containing changes of damaged structures with two platforms, and the other containing unrelated tasks.

5 Conclusions and Future Work

In this paper, we presented a model to mine instructional content in electrical training systems and preliminary results. We need to do more experimentation before to mine the entire electrical procedure knowledge.

There is a large pool of possible improvements for this work, for example: a) the usage of term-frequency instead of the inverse term-frequency will surely reveal more interesting relations between tasks, even by using BOW, b) stemming and replacing BOW by n-grams will reduce the feature space and reduce redundancy (words *platform* and *platforms* are considered different for our current setup), c) the automation of the stop-word removal procedure will improve time and reduce the appearance of irrelevant words in the feature space, and d) instead of using the elbow rule to select the proper number of clusters, more objective functions such as the Davies-Bouldin index or the Mean Silhouette Coefficient might be used.

Once we have more results, we can analyze the domain structure, and propose applications to several aspects of intelligent learning environments.

References

1. Romero, C., Ventura, S.: Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33:125–146 (2007)
2. Baker, R.S.J.d.: Mining Data for Student Models. In: Nkambou, R. Mizoguchi, R, Bourdeau, J. (Eds.), Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence, Vol. 308, Springer-Verlag, Berlin Heidelberg New York, pp. 323–337 (2010)
3. Hernández, Y., Pérez, M., Zatarain-Cabada, R., Barrón-Estrada, L., Alor-Hernández, G.: Designing Empathetic Animated Agents for a B-Learning Training Environment within the Electrical Domain. Educational Technology & Society, 19 (2):116–131 (2016)
4. Kwartler, T.: Text Mining in Practice with R. John Wiley & Sons, Ltd, Chichester, UK (2017)
5. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:881–892 (2002)
6. Ayala, A., Galvan, I., Arroyo, G., Muñoz, J, Rodríguez, E.: Virtual reality training system for maintenance and operation of high voltage overhead power lines. Virtual Reality, 20(1):27–40 (2016)
7. Hernández, Y., Pérez, M.: Virtual reality systems for training improvement in electrical distribution substations. In: Proceedings of 16th ICALT, pp. 75–76 (2016)
8. Hernández, Y., Pérez, M.: A B-Learning Model for Training within Electrical Tests Domain. Research in Computing Science, 87(1):44–52 (2014)

9. Galvan, I., Ayala, A., Rodriguez, E., Arroyo, G.: Virtual reality training system for the maintenance of underground lines in power distribution system. In: Proceedings of 3th International Conference on Innovative Computing Technology (2013)
10. Hernández, Y., Cervantes, M., Pérez, M., Mejía, M.: Data-driven Construction of a Student Model using Bayesian networks in an Electrical Domain. In: Proceedings of 15th MICAI 2016, LNCS, Springer, pp. 481–490 (2017)
11. Witten, I. H., Frank, E.: Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA (1999)
12. Sosnovsky, S., Brusilovsky, P.: Evaluation of Topic-based Adaptation and Student Modeling in QuizGuide. User Modeling and User-Adapted Interaction 1(4):371–424 (2015)
13. Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.): Handbook of Educational Data Mining. CRC Press, Boca Raton, FL (2010)
14. Vellido, A., Castro, F., Nebot, A.: Clustering Educational Data. In: Romero, C., et al. (Eds.) Handbook of Educational Data Mining. CRC Press, Boca Raton, FL, pp. 1–8 (2010)
15. Ketchen, D. J. Jr, Shook, C. L.: The application of cluster analysis in Strategic Management Research: An analysis and critique. Strategic Management Journal 17(6):441–458 (1996)