

Data mining techniques for the study of online learning from an extended approach

Sánchez-Sordo José Manuel*

*Universidad Nacional Autónoma de México, FES Iztacala, Departamento de Psicología,
Área de procesos estadísticos.*

* Corresponding author:
Email: jose.sordo@ired.unam.mx

Received: 2019-07-01; Accepted: 2019-04-23

Abstract

In the latest years information technologies have impacted society changing the way human beings learn, and because of that it is necessary to study the intimate relationship between humans and their technological tools. On this path the extended mind thesis posits human cognition as a process that occurs in conjunction between biological and non-biological components, furthermore Connectivism is stated as a learning theory for the digital age. Based on such approaches this work presents a summary of a research whose objective was to know how people extend their cognitive processes with the aim of learning through the internet. Methodologically, an artificial intelligence algorithm for supervised learning (J48) was used to analyze the data of 336 participants with the aim of obtaining classification rules (patterns) of internet use. Finally, the results show that people who report visiting specialized websites, read electronic books and take into account the spelling of the resources they are looking at on the internet are the ones with optimal strategies for learning online.

Keywords

Cognition, connectivism, data mining, e-learning, extended mind, machine learning.

1. Introduction

In recent years, information technologies have massively impacted society, and the new generations for Marqués (2012), have been naturally assimilating this new culture attached to the digital. Derived from this, international organizations such as UNESCO (2006), have suggested development plans based on informatics focused on the educational, which according to the “knowledge society” aim for a sustainable development (UNESCO 2013). However, current technologies have not only transformed the education system or the economic model, since the use of technological tools or instruments far from being just a matter of the “digital age” or of the current times, is contemplated as foundational of the human genre, because the use and generation of technology allows the modification of nature by human beings, situation that brings with it profound implications on the plane of the evolutionary and particularly of the cognitive, insofar as human activity beyond the biological combines in its core the integration of external tools that allow it to continue transforming the world and transforming itself, since these, tools as Vygotski (1995) stated, are artificial organs and it is because of this that it is proposed that the intimate relationship between human beings and their technologies for learning purposes should be addressed nowadays from novel theoretical and methodological approaches from various disciplines such as psychology, philosophy and computer sciences.

In the same way, one of the main approaches or models that currently addresses the interaction of tools in a deeper sense is the extended mind thesis, which posits human cognition as a process that occurs in conjunction between biological and non-biological components as proposed by Clark and Chalmers (2011). This supposes the integration of tools as a part of the cognitive process, thus giving origin to the notion of extension of the mind—mind as understood from the behaviorist approach of Ryle (2009), that beyond how criticizable it is, his vision of the mind as something that connects with other notions prevents the quick assimilation of itself as an organ, either material or immaterial, as mentioned by Calvo in Clark et al. (2011)— so their thesis (Clark 2008; Clark et al. 2011),

states broadly that the brain, the objects and the world coordinate and extend as one thanks to cognitive action. In respect to this, the authors emphasize that said integration of the world's objects as a fundamental part of the cognitive process is achieved thanks to the functional parity that exists between the functions of both the organism and the object (Clark 2001), that is to say that the tools (whether digital devices or a pencil) are assimilated as part of cognition not because of their physical or material constitution but by the functions they perform.

Likewise, there are proposals similar to the Extended Mind, such as Hutchins's (1995), Distributed Cognition, which postulates that the mind is in the world, in contrast with the notion that the world is in the mind. Because for this approach cognition, as also proposed by Clark (2011), is distributed amongst the people, objects and tools that belong to certain contexts. By way of example, Hutchins (1995) mentions that the knowledge needed to operate a naval ship does not exist only inside a person's head, but that the process is distributed through objects, people and tools in the environment itself. Being then the objective of distributed cognition (2000), to know and describe the distributed units that coordinate to perform cognitive action, as well as the contextual framework in which the activity is performed.

Based on such postulates, this work presents a research whose goal was to know, with the aid of artificial intelligence algorithms, how people, from the perspective of Clark (2008) interact with information networks and the internet to extend their cognitive processes to generate learning in conjunction with software and electronic devices such as web browsers, search engines, computers and smartphones, which within this work's theoretical framework should be understood as cognitive extensions, because by using them we can overcome our natural limitations, thus harboring the cognitive process in a complex composed of biology and technology. For which the study of learning from an extended approach becomes important nowadays since it is necessary to know how learning occurs

in the digital age in conjunction with the tools, and not addressing it more only under the popular but somewhat obsolete concept of ICT for education, that far from reflecting about the psychological processes that are implied and modified with the use of technologies for learning, it seems in part to only open the market for a few applications and artifacts that are believed to place the classrooms and educational institutions at the “vanguard” as suggested by Marqués (2012).

1.2. Extensions and learning in the digital age:

As stated, in this approach human cognition is conceived as an extended process that does not occur entirely inside the skull (Clark et al. 2011), which implies the “fusion” of man with his tools in a psychological sense. The acceptance of such philosophical proposal suggests then a different way of approaching the study of learning and education, being in the sphere of the educational Connectivism (Siemens 2006), a proposal that is stated as a learning theory for the digital age, that is, a referential and explanatory framework of the learning process that occurs in formal or informal digital educational environments (Sánchez 2014), and not a psychological theory that attempts to explain human development or behavior in its entirety, but a theory that provides or tries to provide an answer as to how people learn and increase their current state of knowledge in a specific context: computer networks. This clarification is pertinent given the criticisms that Connectivism has received as a learning theory, such as Zapata-Ros’s (2015), which adhering to old theoretical conceptions refuses to accept Siemens’s (2004), proposal that “knowledge can reside in non human appliances”, approach that is in accord with what Clark (2011) proposes, and that reflects the underlying social context that gives life to these new theories: technology as an active agent in the learning process.

At a conceptual level Connectivism integrates principles of networks, complexity, self-organization and aspects related to the extended mind, since it emphasizes that a large part of the knowledge and learning that is generated occurs outside of people’s heads. For

Downes (2011), Connectivism is the thesis in which learning is distributed throughout a network of connections, and therefore learning consists of the skill to build and navigate those networks. Regarding this, Redecker (2009), states that such networks exist both externally and internally, in the external they are the structures that we create in order to be up to date and to continuously create and connect with new knowledge; and their nodes are the entities (people, online encyclopedias, websites, wikis, forums, applications, etc.) with which individuals connect to form a network. Internally, learning networks can be perceived as structures that exist in our minds in the connection and creation of comprehension patterns, given that as Siemens (2006) affirms we psychologically adapt our brain's connections to process the environment in which we move, given that the brain restructures its neural connections with the use of technology, which clearly implies a direct relation between the organization and functioning of our brain and what we can learn about the world.

For this approach then the value of the study of learning is to comprehend the capacity of individuals to generate connections and networks in informal educational environments on the internet that promote the specialized connection between sets of information that allow individuals to increase their knowledge, thus emphasizing the skill and autonomy of the individual to obtain and store information in structures external to him (mostly computer networks or digital devices), that from the approach proposed here perform as cognitive extensions.

Concerning the above, the study of strategies or ways in which people interact with technology to generate connections and networks for learning becomes necessary since it indicates the use and perception that people have of the learning that is generated in a distributed way on the web. On this path Garay et al. (2013), state as cognitive strategies related to learning on the web 2.0 the following: Search (of information), Compilation, Management, Reflection and Practices, which at the same time are associated with particular online tools like forums, blogs, wikis and social networks.

Although from Connectivism the term cognitive strategies is not used, Downes (2009) mentions that knowledge is no longer monopolized by official instances, but that it is diversified in positions, facts and opinions, being then the skill or strategy to differentiate between reliable and unreliable sources of knowledge something of great importance within connective learning.

2. Methodological perspective

Objective:

To know and describe with artificial intelligence algorithms, how people extend their cognitive processes when interacting with computer networks with the goal of generating knowledge from a connectivist logic.

Type of Study:

A non-experimental study with a quantitative approach using artificial intelligence techniques for data mining was carried out.

Sample:

For this work a non-probability sample of 336 volunteers was required. The majority (71%) were women from Mexico City belonging to the professional areas of Biological Sciences (50%) and Humanities (21%).

2.1. Procedure

Participants were asked to fill the online questionnaire titled “Strategies for the selection and use of information for online learning” developed for this study (content validity index of .95), and which is composed of **60 items** (attributes) that address the possible strategies involved in the *online learning* process. Based on the bibliography (Garay, Lujan, and

Etxebarria 2013; Head and Eisenberg 2010; Hernández 2010) it was decided that the dimensions addressed by the questionnaire were the following:

1. **Sociodemographic data:** Gathers information about the participant, like age and place of residence.
2. **Connectivity:** Aims to know aspects related to the users' internet access.
3. **Type of sources consulted:** Evaluates the number and format of the online resources that the participant consults (what do participants consult on the web?).
4. **Search and access to information:** Provides a list of options to know how the participants search for and access information on the internet (How do they find the information?).
5. **Validation of the information:** Gathers information about what criteria are used by the participants to determine the consulted information as true or reliable.
6. **Storage and recovery:** Inquires about what the participants do to store the resources they consult on the web in order to be able to access them in the future.
7. **Self-evaluation:** Allows the participant to rate their strategies of internet use in relation to the generation of knowledge.

2.2. Data Analysis

For the data analysis, the KDD (Knowledge Discovery in Databases) model was used, it is a multistep process for the discovery of knowledge in large data collections. Nigro, Xodo, Corti and Terren (2004), mentioned that the KDD process is iterative by nature, and depends on interaction for a dynamic decision making. One of the more complete definitions of KDD is proposed by Fayyad (1997), who assured:

Knowledge Discovery in Databases are rapidly evolving areas of research that are at the intersection of several disciplines, including artificial intelligence, statistics, machine learning, pattern recognition and visualization for extracting knowledge automatically from databases. (p.2)

To perform the KDD the following steps are required:

1. **Data selection.** In this stage data sources and the type of information to be used are determined.
2. **Pre-processing.** This stage consists of the preparation and cleaning of data extracted from the various data sources in a manageable form, necessary for the following phases.
3. **Transformation.** Consists of the preliminary treatment of data, transformation and generation of new variables from already existing ones with an appropriate data structure.
4. **Data Mining.** It is properly the modeling phase, in which intelligent methods are applied with the objective of extracting patterns previously unknown, valid, new, potentially useful and comprehensible that were contained or “hidden” within the data.
5. **Interpretation and Evaluation.** The truly interesting patterns obtained are identified based on certain measures and the obtained results are evaluated and interpreted.

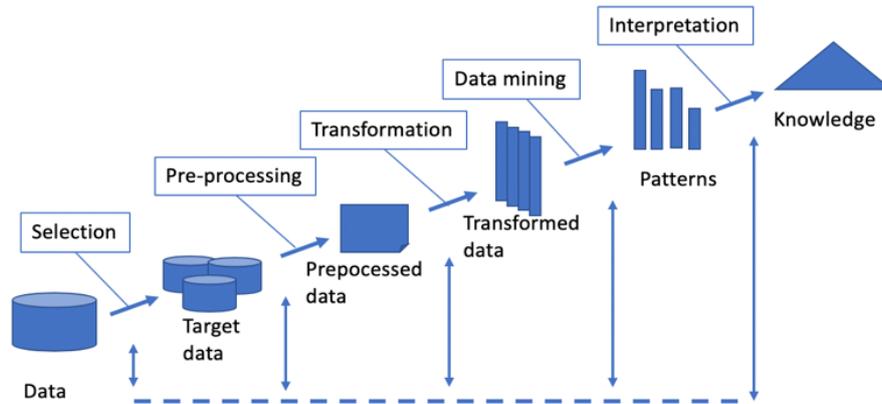


Figure 1. KDD steps outline.

2.2.1 Description of the performed analysis and processes:

Regarding the objective set for this research, and using the previously described KDD model, quantitative analyses were carried out through data mining with the J48 supervised learning algorithm and descriptive statistics in order to be able to interpret the gathered information as it is explained:

- The data derived from the application of the online questionnaire “Strategies for the selection and use of information for online learning” to the 336 participants was analyzed with the goal of finding classification rules (patterns) in the data for the subsequent classification of the subjects according to their internet use strategies (good, regular and poor) using the J48 algorithm. Subsequently descriptive data were obtained according to the participants’ perception of their internet use.

The algorithm used to perform the data analysis is described below:

Decision trees (J48): it is a free version of the commercial C4.5 algorithm developed by Quinlan (Sancho 2018) and has the objective of decreasing the entropy of the data by using information gain. This is to find the attribute that better divides or “arranges” the data according to the categories in which they are to be classified using the function:

$$E(S) = \sum_{i=1}^n p_i \log_2(p_i)$$

Therefore, in order to obtain the attribute that generates more homogeneous branches within the decision tree Sancho (2018), indicates the following:

1. Total entropy is calculated.
2. The data set is divided according to the different attributes.
3. The entropy of each branch is calculated and then they are added proportionally to calculate total entropy:

$$E(T, X) = \sum_{c \in X} p(c) E(S_c)$$

4. This result is subtracted from the original entropy, obtaining as a result the information gain (entropy decrease) using this attribute.

$$\text{Gain}(T, X) = E(T) - E(T, X)$$

5. The attribute with the highest gain (of information) is selected as a decision node, it is to say, the attribute through which the data classification will begin. Finally, this process is repeated with all the nodes that do not perform as outputs until null entropy levels are obtained (leaves).

3. Results and findings

A. Classification rules (patterns) of internet use for online learning:

Next, the results obtained from the application of the J48 algorithm for the discovery of internet use classification rules (patterns) are shown *according to what was reported* by the 336 participants in the “Strategies for the selection and use of information for online learning” questionnaire, it should be mentioned that the Correlation Attribute Eval (Weka 3.8.3) filter was applied in order to estimate correlations between questionnaire attributes and in this way select the most pertinent to analyze with the algorithm.

The model generated by the algorithm to classify the data according to the participants’ internet use strategies showed very high levels of area under the ROC curve (**ROC**), which means that the patterns shown below are accurate indicators to know and predict the participants’ behavior on the internet for education or learning purposes.

Table 1. Statistical details of the J48 classifier tree “Internet use strategies for learning”

Kappa:	0.842
Correctly classified instances	91.6667%
ROC area:	.966
Number of leaves	35

Shown on the following image is the decision tree (pruned) product of the model generated with the J48 algorithm, shown in said tree are the nodes (attributes) that better arranged the distribution of all the data for their classification:

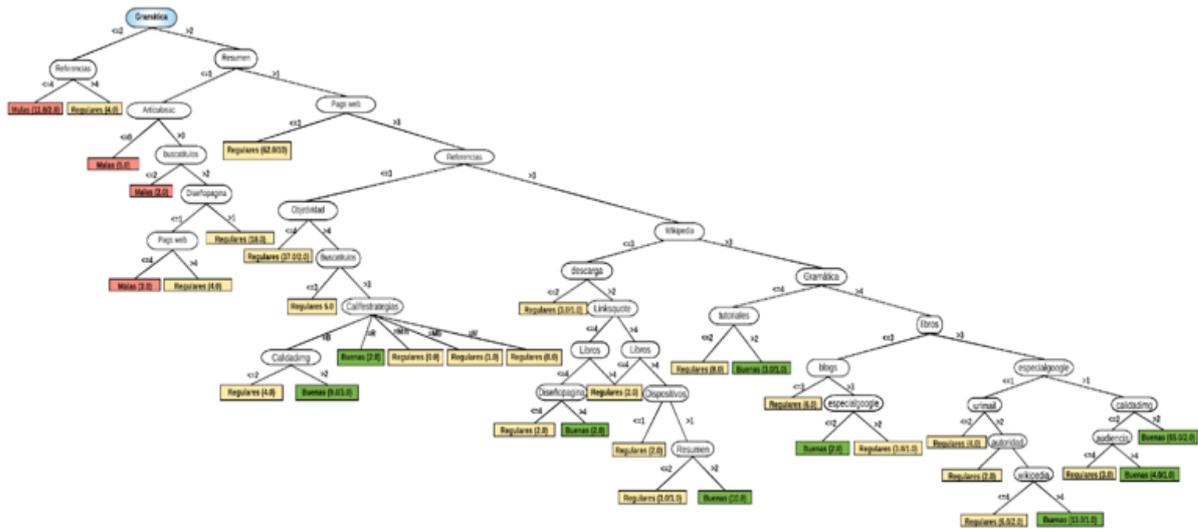


Diagram 1. Decision tree model obtained for the “Strategies for the selection and use of information for online learning” questionnaire data (n=336). **Objective:** to find patterns of internet use for learning and with it classify the participants according to their strategies in “**Poor, Regular, and Good**”, and infer possible patterns of cognitive extensions (see appendix for better image quality).

As it can be observed in the diagram above, 35 patterns to classify the subjects (n=336) were found according to the internet use they reported and the resources they access, store and recover to generate learning. As it can be observed, the tree includes rectangles (leaves) colored green, yellow and red, each corresponding to a classification respectively; **green** for **Good (strategies)**, **yellow** for **Regular (strategies)** and **red** for **Poor (strategies)**.

Table 2. Patterns and classifications of the use of internet for learning

Classification:	Poor	Regular	Good
Number of Patterns:	4 patterns (classification rules)	22 patterns classification rules)	9 patterns classification rules)
Number of instances (persons):	21 (19 correct, two wrong)	205 (187 correct, 18 wrong)	110 (102 correct, eight wrong)

As it can be observed on table 2, the largest number of patterns obtained was for the regular strategies category (22), which means that the majority of the participants have regular strategies of internet use for academic purposes.

A. Patterns with more classified instances:

Poor: 1 (13 persons), 2 (5 persons).

Regular: 1 (62 persons), 2 (37 persons).

Good: 1 (65 persons), 2 (13 persons).

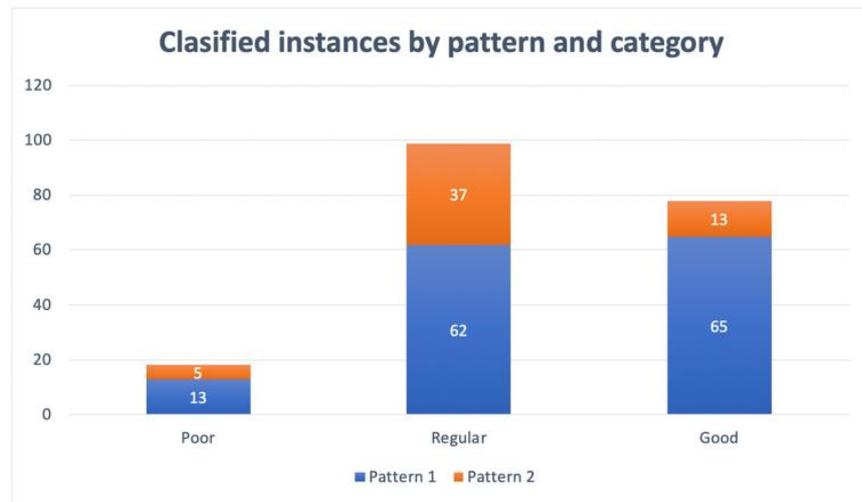


Figure 1. Shown by the graph are the two most representative patterns of each category according to the number of classified instances in their interior.

Then, of the 35 patterns obtained with the algorithm (diagram 1) individual patterns are shown and interpreted for each category that contains the most instances and that better classifies and predicts the participants' actions. With them we can infer how people extend their cognitive processes on the internet, since these individual patterns show the combinations of attributes that the majority of the participants perform in relation to what tools they use and how they do it to connect (Sánchez 2019), and in this way generate knowledge and online learning.

Pattern no.1 of classification rules for good use of the internet)

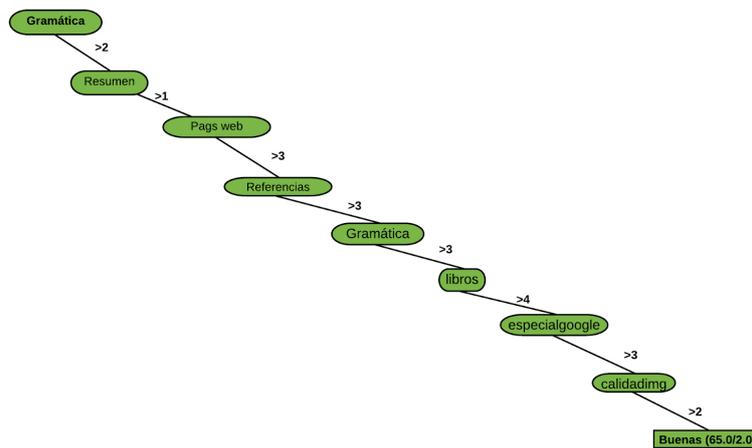


Diagram 2. Main pattern of good strategies for the use of internet for learning.

Interpretation:

Based on this pattern (no.1) we can induce that **most** of the participants (65/110) with **good strategies** of internet use for learning employ digital tools in the following manner:

They are people who almost always take into account the quality of the spelling and writing of the resources they consult online to consider them valid; on occasion they make summaries or take notes of the information they consider relevant, they consult plenty of specialized websites and almost always look up the bibliographical references that are cited within the sources they consult. They also use Wikipedia and check many (4-5) books or book chapters in electronic formats and use many of Google’s specialized search options and almost always take into account the quality of the images and graphics that are included on the online documents they go over.

With this information it can be said that this particular combination of tools and activities is how the participants extend their cognitive processes in conjunction with tools to generate connections on the internet in an optimal way with the purpose of learning.

Pattern no.2 of classification rules for regular use of the internet)

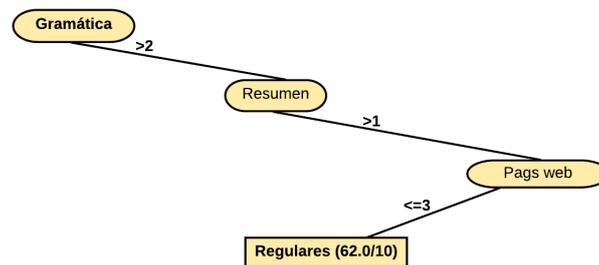


Diagram 3. Main pattern of regular strategies for the use of internet for learning.

Pattern (no.2). In this pattern the participants (62/187) with **regular strategies** of internet use for learning employ digital tools in the following way:

Sometimes they consider writing and grammar as something important to select internet resources, also on occasion they make summaries of the relevant information, but they never or almost never visit specialized websites.

Based on this interpretation it can be said that the pattern predicts that this combination of tools and activities is how most of the participants extend their cognitive processes in conjunction with online learning tools in a way that can be considered as regular, because they take into account spelling and make summaries or notes, but they do not consult specialized websites.

Pattern no.3 of classification rules for poor use of the internet)

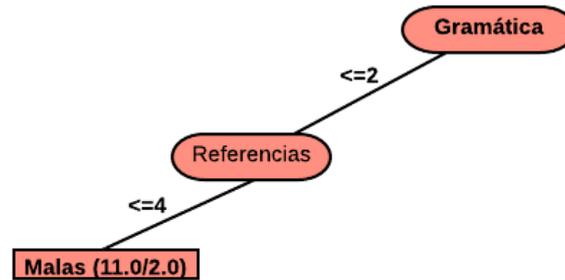


Diagram 4. Main pattern of poor strategies for the use of internet for learning.

Pattern (no.3). In this pattern the participants (11/19) with **poor strategies** of internet use for learning employ digital tools on the following way:

They never or almost never take into account the quality of the writing or spelling of the online resources they consult, as well as they never or almost never look up the bibliographical references that are cited and also visit none or very few specialized websites.

This way we can predict that people that do not take into account grammar nor cited references in online sources are the ones that make a deficient use of online tools (extensions) to connect with new learning.

B. Descriptive analysis:

Shown in this section are a couple of graphics with results obtained from the analysis of the frequencies of the participants' opinions in relation to how technology influences some of their cognitive processes.

The first aspect addressed was: “By being able to store the information you consult online with academic or learning purposes inside virtual folders, your ‘internal’ memory processes were amplified and/or modified.”

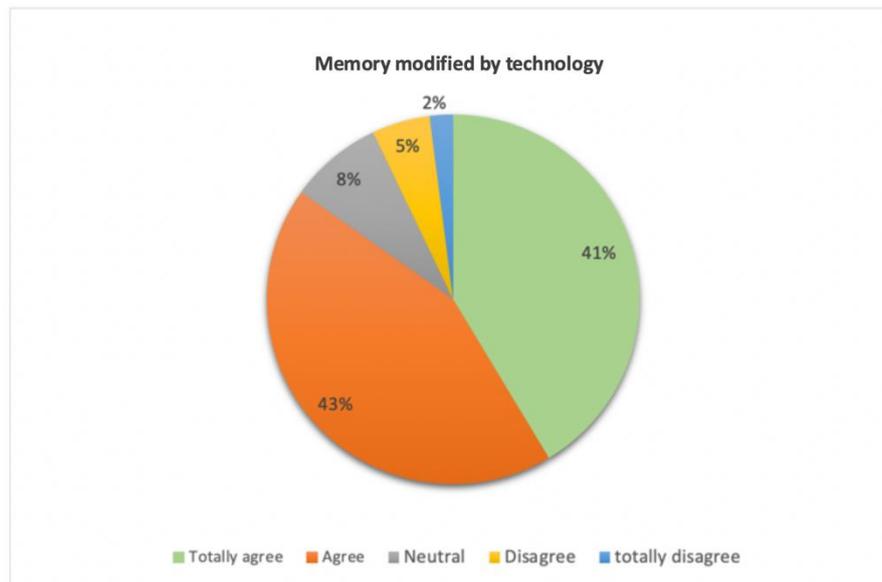


Figure 2. Memory amplified or modified by technology.

Overall, 84% of the participants agree or totally agree that their “internal” memory processes have been modified by the use of current information technologies. Situation accordant with the ideas proposed by Sánchez et al (2019), about the extension of the mind, that state that cognitive processes, in this case memory, are shared or distributed between neural/biological processes and electronic devices.

Likewise, the participants were asked if: “When using computer networks or digital devices to learn or generate knowledge, they consider that their learning is occurring not only ‘inside’ their heads but occurring in conjunction with the devices or networks”.

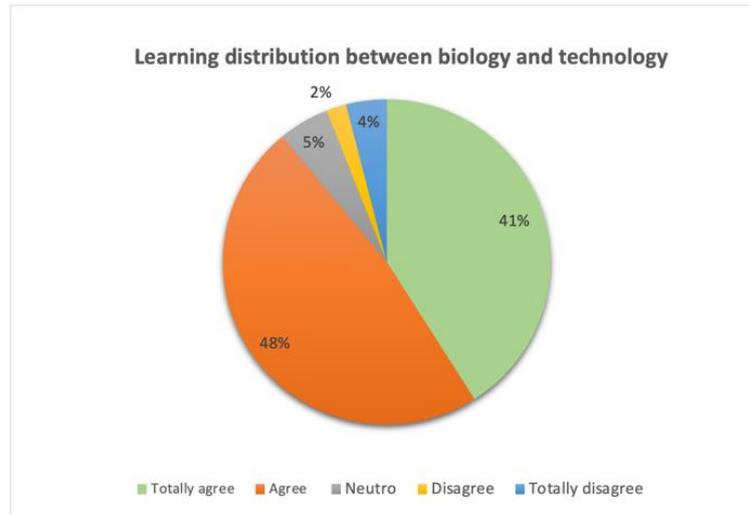


Figure 3. Learning distributed between “inside the head” and technological devices.

We can observe that most of the participants consider that their learning processes occur not only inside their heads, but that they are distributed with devices and computer networks, situation that in some way reflects what is mentioned by Siemens (2004) and Clark (2008), about learning and cognition that occurs “outside” the heads of individuals.

4. Conclusions

The findings of this research show in some way how cognition is shared or extended when we use digital tools, based on the fact that we psychologically adapt our neural connections with the use of technology (Siemenes 2006), which clearly implies a direct relationship between the organization and functioning of our brain and the tools we use (Maravita and Iriki 2004). With this we show that Connectivism is a theoretical framework that shares postulates with the philosophical thesis of the extended mind proposed by Clark & Chalmers (2011), so addressing cognition from these approaches together gives a current and complex view of the phenomenon of learning in digital educational environments.

In relation to this, it is considered that the results of this research provide data of interest to people involved in non-formal and online education, because interesting facts were found about which tools favor online learning, being some of the most favorable abilities or strategies: to check the spelling of online resources, to make summaries of reliable information, as well as the use of some specialized search engines. Also, other relevant result is that most of the participants agree that ICT have modified their “internal” cognitive processes such as memory.

Also, the KDD model allowed us to discover from the data relevant information about how the participants of this study use certain technological tools on the Internet, allowing us to then generate usability profiles that allowed to classify the subjects as having optimal, regular and poor use of the internet with learning purposes, contributing thus with relevant data for learning and educational psychology in the digital age.

In addition, the KDD model and the use of artificial intelligence algorithms such as the ones used in this research proved to be effective tools for the discovery of new knowledge within psychological research, because its use approaches psychology to different methodological paradigms that allow the analysis of human behavior in an innovative way (Fayyad 1997), allowing also the analysis of large amounts of data.

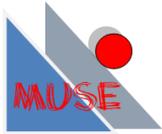
Finally, we can mention that this work’s contribution is to address education from psychology and philosophy with artificial intelligence in a digital environment, areas of knowledge that are giving a fundamental shift to education nowadays. Basically, the proposal of this work is based on an evaluation of the learning process using artificial intelligence techniques for the classification of patterns in the use of information and tools on the internet, thus giving basic clues to support the concept of extended mind, which should be addressed in future educational research starting from these bases and having other considerations as there are other types of learning and ways in which people interact with tools (Apud 2014).

5. References

- Apud, I. (2014). ¿La mente se extiende a través de los artefactos? Algunas cuestiones sobre el concepto de cognición distribuida aplicado a la interacción mente-tecnología. *Revista de Filosofía*. 39 (1), 137-161.
- Castañeda, L., & Adell, J. (2013). *Entornos personales de aprendizaje: claves para el ecosistema educativo en red*. Alcoy: Marfil.
- Clark A. (2001) *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford: Oxford University Press
- Clark, A. (2008). *Supersizing the mind: Embodiment, action and cognition extension*. Oxford: Oxford University Press.
- Clark, A., & Chalmers, D. (2011). *La mente extendida*. Oviedo: KRK Ediciones.
- Downes, S. (2009). *The New Nature of Knowledge*. Online [Available] <https://www.downes.ca/cgi-bin/page.cgi?post=53404>
- Downes, S. (2011). *Connectivism and Connective Knowledge* [Online] Available: https://www.huffingtonpost.com/stephen-downes/connectivism-and-connecti_b_804653.htm
- Fayyad, U. (1997). *Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases*. SSDBM '97 Proceedings of the Ninth International Conference on Scientific and Statistical Database Management , 2-11.
- Garay, U., Lujan, C., & Etxebarria, A. (2013). El empleo de herramientas de la Web 2.0 para el desarrollo de estrategias cognitivas: un estudio comparativo. *Porta Linguarum*. 20, 169-186.
- Head, A., & Eisenberg, M. (2010). *How College Students Evaluate and Use Information in the Digital Age* [Project information literacy progress report]. Washington, D.C.: University of Washington.

- Hernández, S. (2010). Uso de estrategias de aprendizaje con apoyo en Internet considerando el género y el nivel educativo. VIII Congreso Iberoamericano de Ciencia Tecnología y Género. Curitiba.
- Hutchins E. (1995), Cognition in the wild, Bradford Book.
- Hutchins, E. (2000) Distributed cognition, Online [Available]: <http://comphacker.org/pdfs/631/DistributedCognition.pdf>
- Maravita, A., & Iriki, A. (2004). Tools for the body (schema). TRENDS in Cognitive Sciences. 8 (2), 79-86.
- Márqués P., (2012). Impacto de las TIC En La Educación: Funciones Y Limitaciones, Revista 3CTIC, 1, (3).
- Nigro, Xodo, Corti and Terren, (2004), Online [Available]: <http://sedici.unlp.edu.ar/handle/10915/21220>
- Redecker, C. (2009). Review of Learning 2.0 Practices: Study on the Impact of Web 2.0 Innovations on Education and Training in Europe. Bruselas: Joint Research Centre.
- Ryle G., (2009). The concept of mind, Roulledge
- Sánchez-Sordo, J (2019). Redes y cognición; abordando la mente extendida en ambientes conectivistas de aprendizaje, Revista Digital Internacional de Psicología y Ciencia Social, to be published.
- Sánchez-Sordo, J. (2014). Conectivismo y ecologías para la educación a distancia en la web 2.0. Revista Mexicana de Bachillerato a Distancia, 6(12), 11.
- Sancho, F. (2018). Aprendizaje Inductivo: Árboles de Decisión. Obtenido de Universidad de Sevilla: <http://www.cs.us.es/~fsancho/?e=104>
- Siemens, G. (2004). Conectivismo: Una teoría de aprendizaje para la era digital. Madrid: Ediciones Nodos Ele.
- Siemens, G. (2006a). Conociendo el conocimiento. Madrid: Ediciones Nodos Ele.

- UNESCO, (2006). Hacia las sociedades del conocimiento: informe mundial de la UNESCO. [Available]: <https://unesdoc.unesco.org/ark:/48223/pf0000141908> .
- UNESCO, (2013). Education for Sustainable Development Goals: learning objectives. [Available] <https://unesdoc.unesco.org/ark:/48223/pf0000247444>
- Vygotsky, L. (1995). Historia del desarrollo de las funciones psíquicas superiores. Madrid: Editorial Visor.
- Zapata-Ros, M. (2015). Teorías y modelos sobre el aprendizaje en entornos conectados y ubicuos. Bases para un nuevo modelo teórico a partir de una visión crítica del “conectivismo”. *Revistas VSAL*. 16 (1), 1-49.



Appendix:

