REPPE
Revista de Produtos Educacionais e Pesquisas em Ensino

Cluster analysis characterization of research trends connecting social media to learning in the United Kingdom

Caracterização por análise de clusters das tendências da pesquisa envolvendo mídias sociais e aprendizagem no Reino Unido

Ana Lucia PEREIRA[1]
Cristina COSTA[2]
Jose Tadeu LUNARDI[3]

Abstract
In this work we present a characterization of and discuss the research trends connecting social media to learning in the United Kingdom in the last six years. The data set for this research comprises articles published in educational journals indexed in the *Web of Science*® database. A cluster analysis was used to group similar articles in the data set. We characterized the main research trends by identifying the typical features of the articles within each of the main groups that emerged from this analysis.

Key words: Social Media; Learning; Research.

Resumo
Neste trabalho apresentamos uma caracterização e discussão das tendências da pesquisa envolvendo mídias sociais e aprendizagem no Reino Unido nos últimos seis anos. O conjunto de dados da pesquisa compreende artigos originais publicados em revistas de educação indexadas na base *Web of Science*®. Usamos análise de *clusters* para agrupar artigos similares no conjunto de dados. As principais tendências de pesquisa foram caracterizadas por meio da identificação das características típicas dos artigos em cada um dos grupos que emergiram dessa análise.

Palavras-chave: Mídias sociais; Aprendizagem; Pesquisa.

Introduction

The presence of social media in education has often been at the periphery of pedagogy (SELWYN; BULFIN, 2016) in that it is more easily used to demonstrate interest in innovation rather than intention of transforming learning and teaching

[1]University of Strathclyde, UK and State University of Ponta Grossa, Brazil.
[2] University of Strathclyde, UK.
[3] University of Glasgow, UK and State University of Ponta Grossa, Brazil.

practices in the context of the 21st century where social media plays a vital role in the production and dissemination of knowledge.

The inclusion of social media in education requires not only the reassessment of the role of teachers and learners, but also a reassessment of the role of technologies in education. In this regard, social media is more than just a tool for promoting learning and teaching practices through; it is also an enabler of participation and engagement in wider learning contexts. In other words, the effectiveness of social media is not solely reliant on the technological solutions it presents, but also, and above all, on the approaches individuals take to learning and teaching in the environments it facilitates.

Our main aim in this paper is to characterize the main themes the researchers in the UK are interested in when approaching social media and learning. In order to arrive to such a characterization, we try to identify the main research trends in the UK by analysing a data set containing articles published in educational journals that address the topics of "social media" and "learning" in the last six years. This is done by using a *cluster* analysis to group the articles into subgroups (clusters) of similar articles (see, for example, BATTAGLIA *et al.,* 2016; HUBERTY *et al.,* 2005). The main research trends are identified by the typical features of the articles falling into the main clusters that emerge from this analysis. After obtaining such a characterization by using cluster analysis, we discuss it by using the Figueiredo's learning contexts framework (FIGUEIREDO, 2016).

Theoretical background

We base our proposal on the pedagogical approaches proposed by Figueiredo (2010; 2016) who explores *a pedagogy of learning contexts*. A pedagogy of learning contexts brings together a body of theoretical knowledge applied to learning practices rather than the enhancement of teaching. Realising this difference is a crucial step to reflect on the role of social technologies in education and devise relevant pedagogical practices for the demands of the contemporary knowledge society. Moreover, a pedagogy of learning context is underpinned by principles and values of freedom and emancipation (FREIRE, 1970), democracy and experience (DEWEY, 1916), as well as forms of individual autonomy (ILLCH, 1971) and collective agency (WENGER, 1990). These conceptions of education were conceived prior to the advent of social media, but only now can be materialised through the affordances the web provides. These same principles can be easily associated with a set of emergent cultural practices that

are being mediated by social media, and which denote both a shift in practice and an approach of how individuals engage with both information and other learners.

The true contribution of social media to education is that they metaphorically free learners from Weber's pedagogical *Iron Cage,* i.e., a pedagogy of knowledge transmission and authority (FIGUEIREDO, 2016) and give way to a pedagogy of freedom where individual and collective agency becomes part and parcel of the learning and teaching strategies. Pedagogical approaches influenced by social media are but a reflection of the social and cultural practices such means of knowledge production inspire. As McLuhan (1964) reminds us media – as a tool and a means of communication – is shaped by our current needs as much as it shapes our practices. Digital social media has a similar effect in that it is shaping not only the way individuals live, but also work and learn (JISC, 2013).

Figueiredo (2010) reflects on this aspect by reviewing the role of social media – and particularly the web – on individuals' knowledge practices. In this vein, Figueiredo maps out the differences of knowledge practices between a non-web generation (1.0) and the web generation (2.0). The differences between the two generations are not only expressed by their relationship with information and literacies, but also by their engagement with technology. The knowledge practices of the generation 2.0 is characteristically more embedded if not embodied in a mediated world where social media is not only a tool, but also an environment that supports, justifies and represents learners' learning practices.

Methodology

The basic units of our data set are research articles, which are grouped into subgroups of similar articles by using a cluster analysis in order to allow the identification of the main features of the articles falling into the main subgroups. These typical features will be interpreted as the main research trends we are trying to identify in the data set.

To obtain our data set we used the automated search tool contained in the internet site of the database *Web of Science* (Thomson Reuters)[4] to gather research publications involving the topics of "social media" and "learning". The initial search outcomes were further refined to only include articles published between 2011-2016

[4]www.webofscience.com

and which were written in English and focused on the UK, namely, England, Scotland, Wales and Northern Ireland. Finally, all the articles' titles, abstracts and keywords were carefully read. The articles which did not contain properly the concepts of "social media" or "learning" were discarded by the authors. The remaining ones constituted our data set.

We used a simplified version of the *cluster analysis* proposed by Battaglia *et al.,* (2016) in order to partition the obtained data set into subgroups containing similar articles. To do so, we first analysed all the articles' titles, abstracts and keywords to identify what the main themes were. We then devised a set of categories representing these main themes by using a *Discourse Analysis* strategy as proposed by Bardin (2011), which consists in three chronological stages: pre-analysis, exploration of the material and a qualitative data analysis with possibilities of inferences and interpretations. After the identification of a set of relevant categories, each article was given a binary code built as a sequence of digits 0's & 1's, the position of each digit in the code corresponding to a specific category, in a way that digit 1 means that the corresponding category was addressed in that article and digit 0 means that it was not. The notion of "similarity" between two articles was defined in terms of the similarity between the corresponding binary codes which, by its turn, was defined in terms of a "distance" between the two codes[5]. We consider that two codes are identical if the distance between them is zero. The greater the distance between the two corresponding codes, the more dissimilar the articles are. The central idea of the clustering procedure is that two elements in the same cluster are more similar between them them than two elements belonging to different clusters.

Translating the above technical concepts into the language of our investigation we say that two articles are more similar between them when they share a significant amount of addressed categories, and in this case they tend to fall into the same cluster. On the other hand, when two articles address very different subsets of categories then they should be considered as dissimilar and, therefore, they tend to fall into different clusters.

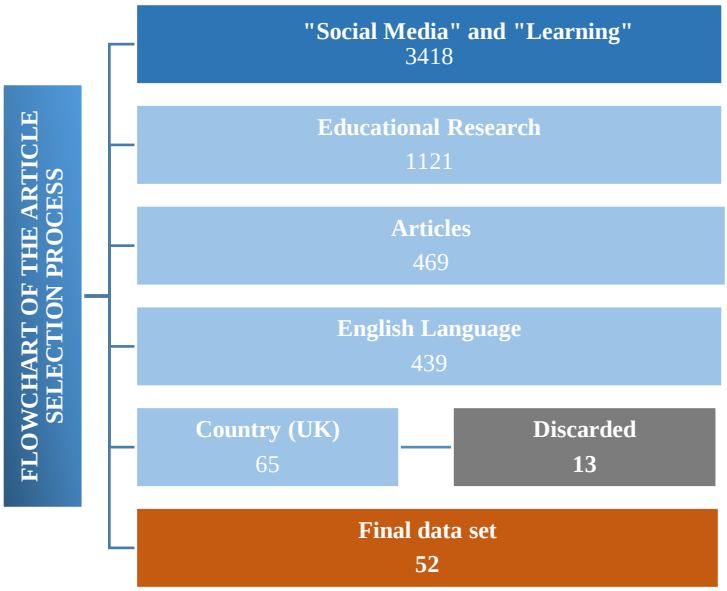
Among the several strategies to partition a data set into clusters, we chose a hierarchical clustering procedure, in which each hierarchical level can be defined by setting a maximum distance in a way that elements (or clusters in a lower level) are

[5]We used the "Hamming distance", that simply measures the number of digits which differ in the two codes.

joined by their next neighbour in a hierarchical sequence until the maximum clustering distance is reached. We defined the distance between two clusters as being the average distance between their elements. As we consider larger distances more clusters/elements join with their nearest neighbours, forming larger clusters. In the extreme cases we have all the elements isolated (zero maximum distance) or a single cluster containing all the elements of the data set (by choosing a sufficiently large maximum distance). Between these two extreme cases we have a defined number of cluster at each hierarchical level. We used an *ad hoc* criterion to select a suitable hierarchical level in such a way that the number of resulting clusters were not so small in a way that no differences could be observed, nor was it so large that no similarities could be noted. For the clustering procedure we used the built-in clustering commands in the software *Mathematica*© (Wolfram Research).

We handled the outcomes of the quantitative cluster analysis in the following way: The largest clusters obtained in the settled hierarchical level were interpreted as the most representative ones and, therefore, we restricted our analysis to them. We identified in each of these most representative clusters the “typical” categories which were addressed by the articles falling into the cluster. The typical categories within a given cluster were identified as those presenting a frequency of occurrence within the cluster greater than or equal to 50%. In this way, we were able to associate to each cluster a “typical article”, that is, a fictitious article associated with the cluster’s typical categories. The typical articles associated with the most representative clusters thus represent the main trends in the research topics we aimed to explore in this paper.

Fig. 1 – Sample selection procedure performed in the *Web of Science* database. The numbers in each block denote the number of outcomes in each step. The first step (block in dark blue) corresponds to the search by “Topic” in the automated search tool; the blocks in light blue correspond to further refinements of these results. The block in grey regards the number of papers discarded by the authors (BAINBRIDGE *et al.,* 2014, adapted*)*.



From: Authors.

Results

The number of articles resulting at each step of the search and filtering procedure performed in the *Web of Science* database (on 26th November 2016) is summarized in the Figure 1. The final data set contained 52 articles whose titles, abstracts and keywords were carefully analysed to find out the relevant categories. By using the discourse analysis by Bardin (2011) we obtained a set of 10 relevant categories, labelled as A, B, C, , J, as shown in Table 1. After that, a binary code containing 10 digits was associated to each article in the data set. The first digit in the code was associated with the category A, the second one to the category B, and so on.

Table 1 – The set of 10 categories identified in the sample of articles.

Label	Category
A	Socialization
B	Formal Educational Environments
C	Professional Development
D	Informal Educational Environments
E	Ethical Issues and Risks in the Virtual World
F	Community of Practice
G	Mobile learning

H	Online learning
I	Social Networks
J	Engagement

From: Authors.

As an illustration, the binary code associated to a specific article of our data set is shown in Table 2; that code shows that the given article address only the categories labelled as A, C, F, H and J, and does not address any other category of the set of categories given in Table 1.

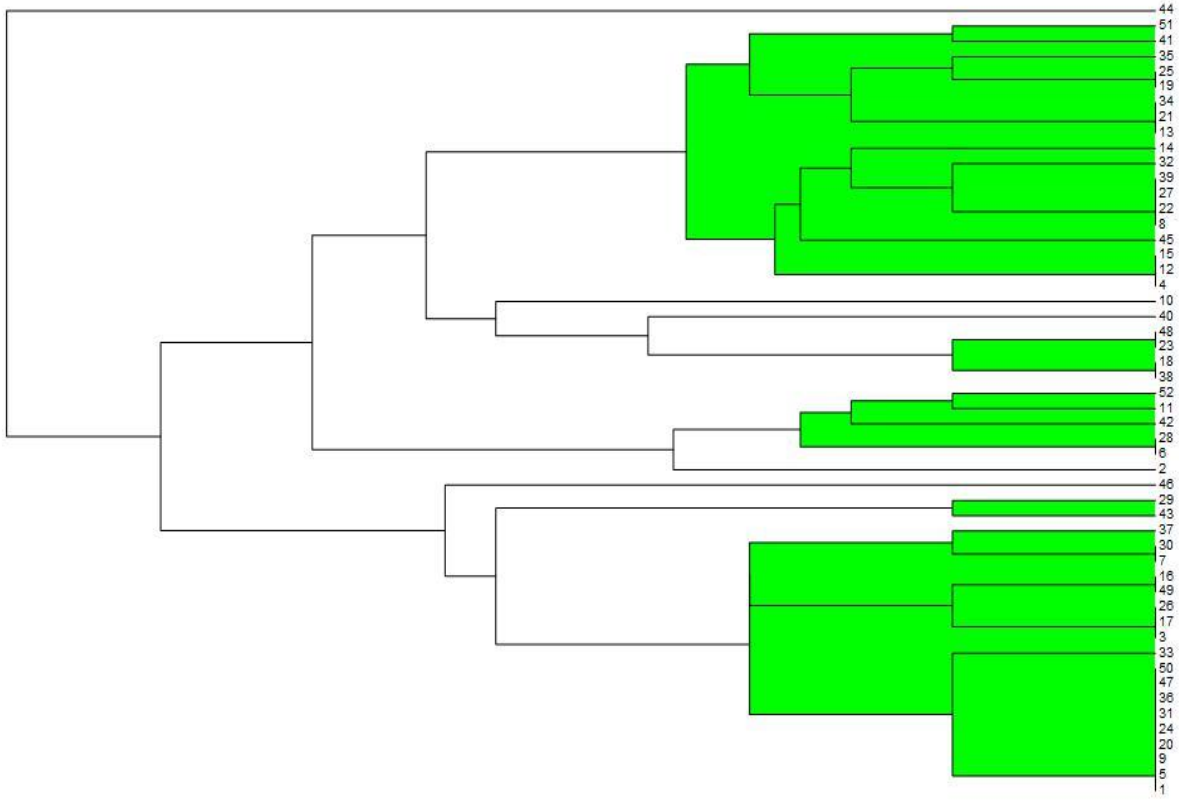
Table 2 – Binary code associated to the article labeled as A29 in the data set.

Categories	A	B	C	D	E	F	G	H	I	J
Code	1	0	1	0	0	1	0	1	0	1

From: Authors.

The hierarchical clustering procedure was done by using the built-in commands in the software *Mathematica©.* The result is shown in the form of a dendrogram in Figure 2. Each article in the data set is represented as a numeric label at the right side of this figure. By setting the maximum distance in 2.35 we obtained the 10 clusters highlighted in green. By labelling these clusters from the bottom to the top as C_1, C_2, etc, we observe the proeminence of two large clusters (C_1 and C_9, having 18 elements each), as well as three smaller clusters (C_5, C_6 and C_2, containing respectively 5, 4 and 2 elements each). The remaining 5 clusters contain a single element each. As the "typical article" of a given cluster we considered a virtual article whose binary code was formed by associating to each position in the code the most frequent digit (0 or 1) observed for that position, taking into account only the articles that fell into that cluster. From this analysis the clusters C_1 and C_9 emerged as the most representative ones, since they are the largest ones and contain the majority of the articles in the data set. Their "typical articles" representatives are shown in Table 3.

Fig. 2 – Dendrogram corresponding to the hierarchical clustering. The horizontal lines are proportional to the (average) Hamming distance between each pair of clusters/elements. The label identifying each article are in the right side of the figure. Highlighted in green are the 10 clusters formed at the hierarchical level corresponding to setting a maximum distance of 2.35 (at this level we note that are some clusters consisting of a single element).



From: Authors.

Here we also propose an alternative way to associate a "representative element" to a cluster, by means of a *fictitious* code where, instead of 0 and 1, we associate the *percentage* of occurrence of the digit 1 in the corresponding code position. This gives a better idea of how the addressing of each category is distributed among the articles within the given cluster (see Table 4). From Table 3 we observe that a typical article in cluster C_1 would address only the category A and would not address any other. From the percentages of Table 4 we observe that all of these articles (100%) indeed address category A and no one addresses categories B, C, D and F (0%). However, some of the articles in this cluster addresses also other categories, being categories H and I the next ones addressed predominantly (with frequencies 27.8%); other categories (G and J) are also addressed with lesser percentages. An article in cluster C_9, by its turn, typically addresses category C but doen't address any other (see Table 3). Table 4, however, shows that all the articles in this cluster address category C (100%) and no one addresses categories B and E (0%); when addressing the other categories, they concern predominantly category I (27.8%) and, successively F and G, H, A and J with lesser percentages. The same readings can be repeated for any of the other (less representative) clusters.

Table 3 – Typical representatives codes of clusters C_1 and C_9.

CLUSTER	CATEGORIES									
	A	B	C	D	E	F	G	H	I	J
C_1	1	0	0	0	0	0	0	0	0	0
C_9	0	0	1	0	0	0	0	0	0	0

From: Authors.

Discussion

From the dendrogram of Figure 2 we can see that at the highlighted level in the hierarchical clustering procedure the sample of articles considered in this study group into 10 clusters. Each cluster contains articles that tend to be similar among them in the sense that they address similar categories in their titles, abstracts or keywords. On the other hand, in this sense an article in a given cluster tends to be significantly dissimilar with another article in a different cluster. The two bigger ones are Clusters C_1 and C_9, containing 18 elements each; they are the most representative ones and we will interpret the typical features of the articles fallling within them as giving a characterization of the main trends of the research themes in our data set. Tables 3 and 4 show the main features of the articles belonging to each of these two most representative groups.

Translating the quantitative results, our investigation identified two major trends regarding the publication of research on “social media” and “learning” - as indexed in *Web of Science* database in the field of educational research - over the last 6 years in the UK. Articles in cluster C_1 cover approximately 35% of the research papers from the sample. Papers in this cluster have a strong similarity among themselves given that all of them address the role of “socialization” that social media have in education. This feature is reinforced by the simultanous addressing of the categories “online learning” and “social networks” in a significant amount of papers in this cluster, which is visible from Table 4. This research trend seems to evidenciate the importance that Figueiredo (2006; 2016) attributes to social media in the building of social relationships (in both their positive and negative aspects) and evidenciates, in a lesser extent, the importance of social media in their role as a facilitator to the teaching and learning process. The other important trend is identified through cluster C_9, that reveals a strong research interest in the subject of “professional development”, which is also reinforced

by the simultaneous addressing in some papers of the categories “social networks”, “comunity of practice” and “mobile learning”. This research trend is in line with McLuhan’s (1964) concept of media in what regards the impact of social media on the way people live and work, and again, in a less extent, seems to reinforce the importance that Figueiredo (2006; 2016) attributes to them as a facilitator to the acquisition of knowledge.

Table 4 – Fictitious codes representing clusters C_1 and C_9. In each position corresponding to categories in the codes are the frequencies of occurrence of the digit 1 as observed in the articles within the cluster.

CLUSTER	CATEGORIES									
	A	B	C	D	E	F	G	H	I	J
C_1	100	0	0	0	0	0	5.6	27.8	27.8	5.6
C_9	5.6	0	100	5.6	0	16.7	16.7	11.1	27.8	5.6

From: Authors.

This paper also aimed to illustrate the power that cluster analysis might have in helping to identify patterns of similarity (or dissimilarity) between objects traditionally analysed using qualitative methods. Even if in this paper we dealt with a not so large data set, and used just one cluster strategy (the hierarchical clustering), the most evident strenght of the cluster analysis is that it can deal with a very large number of objects and is able to identify subgroups of similarities which might be hardly identified by using only qualitative tools on a large number of data. The analysis could be enriched by using other cluster strategy complementarily, such as *k*-means clustering (BATTAGLIA *et al.,* 2016; HUBERTY *et al.,* 2005).

It is worth to mention that cluster analysis is a quantitative exploratory tool, and obviously does not substitute any of the qualitative tools in the human and social sciences literature. As any other quantitative tool in human or social sciences, its outcomes depend strongly on the way a numeric codification is attributed to qualitative data (as well as they depend of the clustering strategy adopted, the choice of the distance measure, etc.). Such outcomes may eventually not make sense to the researcher and he/she should rethink the cluster strategy in order to get meaningful results from it; when the researcher is successful in obtaining meaningful results from this exploratory analysis he/she can use it as a complementary tool to compare, deepen and enrich the qualitative discussion.

Acknowledgement. ALP thanks the partial support from CAPES, Brazil (Grant n. 99999.000403/2016-04).

Referências

BAINBRIDGE, R.; TSEY, K.; McCALMAN, J.; TOWLE, S. The quantity, quality and characteristics of Aboriginal and Torres Strait Islander Australian mentoring literature: a systematic review. **BMC Public Health**, v. 14, n. 1, 2014.

BARDIN, L. **Análise de conteúdo**. São Paulo: Edições 70, 2011.

BATTAGLIA, O. R.; DI PAOLA, B.; FAZIO, C. A New Approach to Investigate Students' Behavior by Using Cluster Analysis as an Unsupervised Methodology in the Field of Education. **Applied Mathematics**, v. 7, 2016.

DEWEY, J. **Democracy and Education**: An Introduction to the Philosophy of Education, 1916.

FREIRE, P. **Pedagogy of the oppressed**. Continuum International Publishing Group, 1970.

FIGUEIREDO, A. D. **A Geração 2.0 e os Novos Saberes, Seminário 'Papel dos Media' das Jornadas "Cá Fora Também se Aprende**, Conselho Nacional de Educação, 2010.

_____. A Pedagogia dos Contextos de Aprendizagem. **Revista e-Curriculum**, v. 14, n. 3, 2016.

HUBERTY, C. J.; MICHAEL JORDAN, E.; BRANDT, W. C. Cluster Analysis in higher education research. In: Smart, J. C. (Org). **Higher Education**: handbook of theory and research. Vol. XX**.** Springer, 2005.

ILLICH, I. **Deschooling society**. Harpercollins, 1971.

McLUHAN, M. **Understanding Media**: The Extensions of Man. London: Routledge, 1964.

SELWYN, N.; BULFIN, S. Exploring school regulation of students' technology use – rules that are made to be broken? **Educational Review**, v. 68, n. 3, 2016.

WENGER, E. **Communities of practice**: learning, meaning, and identity. Cambridge University Press, 1999.