

Complexity of digital resources: an analysis based on their conceptual networks

Crystiam K. P. Silva¹, Sean W. M. Siqueira(Advisor)¹, Bernardo P. Nunes(coadvisor)²

¹Graduate Program in Computer Science
Federal University of the State of Rio de Janeiro (UNIRIO) - Rio de Janeiro – RJ – Brazil

²Australian National University – Canberra – Australia.

{crystiam.kelle, sean.siqueira}@uniriotec.br, bernardo.nunes@anu.edu.au

Abstract. *Knowing the level of complexity of digital resources is crucial to delimit their use in the educational context. This paper summarizes the contributions of my thesis and focuses on strategies to build conceptual networks based on the content of digital resources; identifying metrics and features to measure complexity in conceptual networks accurately; and, proposes new approaches to level digital resources complexity. The contributions of this thesis are extensively evaluated with two large datasets containing resources in varied levels of complexity. The results show that the proposed metrics and features are suitable to estimate digital resources complexity and applicability in educational scenarios. The outcomes of this thesis have been published in high-impact venues.*

1. Introduction

One of the key factors for the selection of proper digital educational resources is the level of complexity. The adequate complexity of the digital resource can keep students engaged in learning tasks, provoke their curiosity [Wu and Miao 2013][Berlyne 1960], and affect their interest [Silvia 2005] and how they interact with the information [Zyngier et al. 2007].

Complexity is a multi-faceted phenomenon that has no predefined and unique set of dimensions and features to capture our intuitive ideas about what is meant by complexity [Gell-Mann 1995]. Moreover, a complexity feature can take different forms depending on the context and scientific domains. Therefore, there is often more than one way by which to measure any given complexity feature [Gell-Mann 1995].

Some researchers have investigated the complexity of digital resources through features such as readability, amount of information, coherence, content overlap and presentation [van der Sluis and van den Broek 2010] [Van der Sluis et al. 2014] [Benjamin 2012] [Collins-Thompson et al. 2011] [Sweller and Chandler 1994]. While all of these perspectives address relevant characteristics, they ignore the complexity of the concepts that need to be presented. Even studies that consider the concepts often underestimate the additional complexity caused by the relationships among concepts.

The central problem addressed in this research is focused on identifying appropriate (1) features of complexity and (2) strategies to deal with the complexity of a digital resource considering concepts included in it and their interrelatedness. The following activities were established to deal with these problems:

- to investigate ontologies, vocabularies, knowledge graphs available to build a conceptual network associated with digital resources;
- to identify the state-of-the-art metrics and features of complexity through a literature review;
- to propose features and metrics applicable to conceptual networks;
- to evaluate these features and metrics;
- to investigate strategies to adapt the complexity of digital resources;
- to propose strategies to deal with the complexity of digital resources based on their conceptual networks;
- to evaluate the impacts of the strategies on the features of complexity;
- to demonstrate multiple applicability in the real educational scenario.

Our approach supports educational stakeholders in recognizing, estimating and adapting the complexity of digital resources. Besides, its relevance for the Computers in Education has been already demonstrated through some publications obtained:

- a comprehensive systematic mapping about use of linked data in Education [Pereira et al. 2017b][Pereira et al. 2017a] produced to define the process for building the conceptual network;
- an analysis of differences in features of complexity according to the precedence of pairs of prerequisites [Pereira et al. 2019]
- an approach to structure a course's content as a conceptual network and to create visual representations to assist educational stakeholders to minimize the effects of oversimplification and compartmentalization of knowledge [Pereira et al. 2020].

2. State of the art

A literature review¹ was conducted on two main fields: (1) Complexity Theory and its employment in contexts such as complex systems, complex networks, tasks, and health, and (2) Education, in which the focus was on teaching and learning of complex educational subjects.

The definitions of complexity and complex systems, as well as their applications in task complexity and education context, indicate that factors such as the number of components of the systems, the interactivity between them, and the variety and diversity are studied as features that influence the complexity. These features are adapted and measured differently for each research field, emphasizing the subjectivity [Gell-Mann 1995] and context-dependency [Mainzer and Landauer 1995] of complexity.

From an educational perspective, studies focus on analyzing educational environments and structure as complex systems, including the relationship between educational stakeholders, political and pedagogical challenges. Few studies presented measurable and applicable features in real scenarios regarding the complexity of the educational domains to be learned or of digital resources. However, theoretical discussions indicate factors that influence the complexity and challenge the learning of complex domains.

This thesis proposes and investigates a new perspective of complexity analysis, which considers the intertwined conceptual network of a digital resource and, based on it, offers measurable features to estimate and deal with complexity.

¹Details and results from the literature review are under submission to an international journal.

3. The conceptual network of a digital resource

We propose to structure the digital resources' content as a conceptual network in which "nodes" of the network are concepts extracted from the resources, and the "edges" are semantic relationships among the concepts. This conceptual network can increase the understanding of digital resources and represent their complex structure formed by multiple intertwined concepts. The conceptual networks are composed of multiple and interrelated concepts that can have many attributes, and the relationships that can have diverse types, meanings and hierarchical levels.

To exemplify how the proposed process works, we used a topic extracted from Wikipedia - the Apartheid article ². The content of the Apartheid article is composed of unstructured information. They neither have an explicit representation of their concepts nor the structured design of how they are related. A comprehensive understanding of this article includes an understanding of varied concepts, such as ideologies, countries' history, notable leaders or institutions. For a meaningful understanding of this topic, such concepts have to be learned not only through their characteristics as isolated concepts; they inevitably need to be recognized as part of a network of concepts representing how they are related to each other. Therefore, a deep understanding of this subject can become complex since each concept must be learned with several other components.

To automatically extract structured information from digital resources, we proposed a framework called DRC-NR Framework ³, which takes advantage of named-entity recognition services, knowledge graphs, ontologies and taxonomies to extract a set of concepts included in digital resources, semantically enrich them, recognize their types and the relationships among them.

Figure 1 shows part of the conceptual network of a Wikipedia article about *Apartheid*, built through the DRC-NR Framework components. In this example, firstly, we submitted the article text to a Named Entities Recognition (NER) service to detect entities. Next, we used knowledge graphs to obtain attributes, relationships, classes and categories for each identified entity. Finally, we used ontologies/taxonomies to identify hierarchical relationships and group the entities ⁴.

After receiving digital resources constituted by unstructured information and modeling the digital resources' content as a semantically enriched conceptual network, the DRC-NR Framework provides: (1) an overview of the concepts inserted in digital resources, (2) a representation showing how concepts are related to each other, revealing connections and combinations, conceptual dependencies, and conceptual variations across contexts, (3) groups of concepts into different classes and categories, (4) analyzable dimensions, such as size, interrelatedness, clustering and diversity, (5) a comparative analysis among different digital resources in terms of the conceptual networks and features related to complexity, (6) adaptable conceptual networks based on features related to complexity and (7) user-friendly representations based on some instructional design principles to deal with complex topics.

²<https://en.wikipedia.org/wiki/Apartheid>

³Details from the steps included in DRC-NR can be seen in [Pereira et al. 2020][Pereira et al. 2019]

⁴In this example, we used the DBpedia Spotlight tool (<https://www.dbpedia-spotlight.org/>) to support the Named Entities Recognition task, and DBpedia Knowledge graph and ontology to identify their relationships and attributes

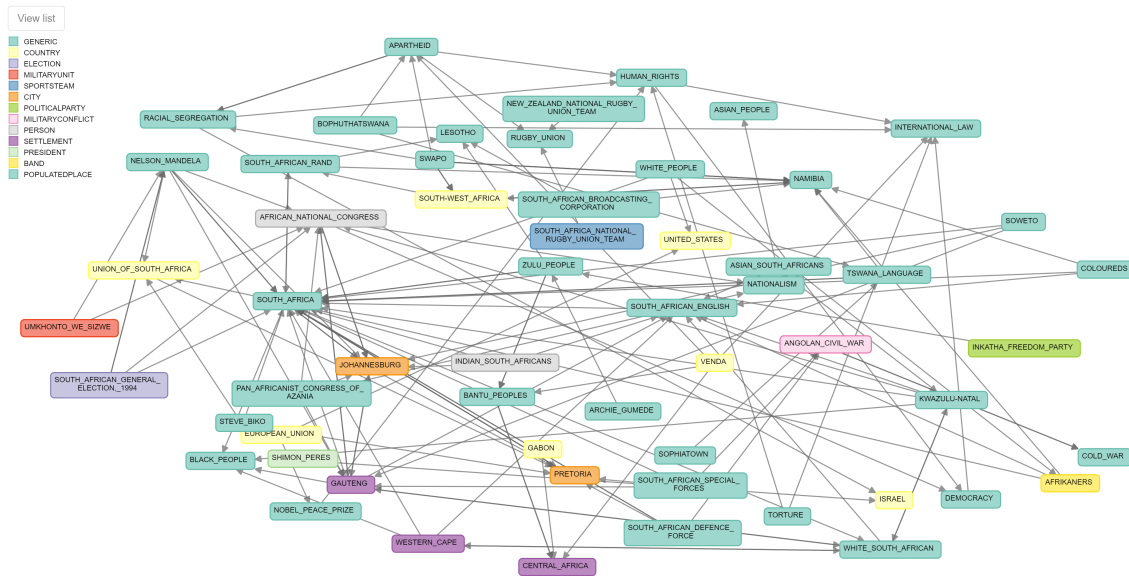


Figure 1. Example of the conceptual network about Apartheid

4. Features and metrics of complexity

To analyze and estimate the complexity of digital resources, considering concepts and their interrelatedness, we addressed two research questions:

- (RQ1): How can features and metrics of complexity be applied in conceptual networks of digital resources?
- (RQ2): What features applicable to conceptual networks best distinguish the complexity of digital resources?

4.1. Methodology

We performed an extensive literature review on Complexity Theory to understanding the diverse dimensions, features and metrics associated with complexity and their applicability in varied contexts. As a result of the literature review, we proposed a set of features and metrics that could also be applied in conceptual networks to estimate the complexity of digital resources.

We evaluated the proposed features and metrics by an experimental study. First, we selected equivalent articles (same topic) from two Wikipedia versions: Simple ⁵ and Regular English ⁶ and built conceptual networks using the DRC-NR Framework. Next, we calculated all the features/metrics to over 14,000 articles of diverse domains extracted from these two datasets with different complexity. After that, we statistically compared the results using significance tests to verify if there was statistical difference between features values in the two datasets (non-parametric tests, using the Mann-Whitney [Wilcoxon 1945] test) with a 95% confidence level ($p\text{-value} < 0.05$). We also measured how the features differ between samples (we used the Vargha and Delaney's A12 effect size).

⁵https://simple.wikipedia.org/wiki/Main_Page

⁶https://en.wikipedia.org/wiki/Main_Page

Finally, we applied a feature engineering process to identify the features' importance to distinguish between simple and complex networks. The features engineering process indicated the importance of the features analyzed and supported selecting the most appropriate features to adjust the complexity of the concept network. To support the features engineering process, we adopted the Boruta algorithm [Kursa et al. 2010]⁷.

4.2. Proposed features and metrics

We proposed using and adapting traditional complexity metrics and novel metrics applicable to conceptual networks to estimate complexity features based on features and metrics used in varied contexts (e.g., task complexity, complex network, complex systems). They were associated with four dimensions of complexity in digital resources context - *size, interrelatedness, clustering, and uncertainty*.

With this set of features and metrics, we answer the **research question RQ1**. Table 1 summarizes this set of features and metrics associated with four dimensions. The table also shows how these features and metrics can be applied or adapted to conceptual networks and their relationship and complexity (positive or negative) based on the literature reviewed.

4.3. Results

The experimental study showed a statistically significant difference in all proposed features and metrics, considering the datasets of resources with two distinct levels of complexity. Moreover, the process of featuring engineering highlighted some of the most important ones. The results answer the **research question RQ2**, showing that:

- All the features proposed were considered relevant. The *coverage, exploratory uncertainty (measured by diameter), number of concepts (measured by counting of concepts), the clusters of concepts distribution (measured by entropy by community) and number of relationships* are the five features with the highest levels of importance.

Briefly, some findings associated with RQ2 are:

- *the structure of the conceptual networks between articles with distinct complexity is different in terms of size, interrelatedness, clustering, and uncertainty of interactions and types of concepts;*
- *the density of simpler conceptual networks is greater than that of more complex conceptual networks*⁸. *However, it is worth mentioning that this is a measure influenced by the number of concepts - the portion of the potential connections in a network that are actual connections - therefore, the small number of concepts of simple networks can have influenced the density;*
- *there is a substantial difference in the values of complexity features concerning the domain;*
- *the simplification of digital resources decreases the coverage and the diversity of topics covered;*

⁷<https://www.rdocumentation.org/packages/Boruta/versions/7.0.0/topics/Boruta>

⁸There is a statistical difference, however, it was not possible to state that conceptual density is higher for Regular networks as presented in Table 1

Table 1. Complexity dimensions, features and metrics for digital resources

Dimension	Feature	Metric	Original/Applied/Adapted	Relationship with complexity
Size	Number of concepts	Counting of concepts on the network	Adapted number of elements [Wiesner and Ladyman 2020]	Positive
		Conceptual density $CD_{cn} = \frac{CN_c}{NT(c)} * 100$	Adapted Keyword density [Malaga 2009] [Bansal and Sharma 2015]	Positive
	Number of relationships	Counting of relationships on the network	Adapted number of interactions [Wiesner and Ladyman 2020]	Positive
	Information load	Conceptual information load $CIL_c =$ number of properties with literal values of a concept	Original	Positive
Network information load: Sum of conceptual information load $NIL = \sum_{c=1}^{c=n} CIL(c)$		Original	Positive	
Coverage	Domain conceptual coverage: ratio of number concepts of the conceptual network to conceptual network of the entire domain	Original	Positive	
Interrelatedness	Concept connectivity	Degree of a concept out-degree and in-degree of nodes	Applied Degree centrality	Positive
		Concept degree Frequency - Inverse Knowledge graphs Frequency $CDF - IKGF_c = \frac{\log(CD(c))}{\log(KGD(c))}$	Original	Positive
	Conceptual Network connectivity	Average degree $\bar{g} = \frac{1}{n} * \sum_{c \in CN} g(c)$	Applied Degree average	Positive
		Network density $DEN_{CN} = \frac{r}{n(n-1)}$	Applied Network density	Positive
Clustering	Number of communities	Counting of communities in the conceptual network	Applied Louvain method [Blondel et al. 2008]	Positive
		Clusters of concepts distribution	Size of communities: counting of concepts in each community	Applied Louvain method [Blondel et al. 2008]
	Diversity of types	Entropy using communities $H(p) = - \sum_{i=1}^{nc} P_i \ln(P_i)$ $P_i =$ probability of a random concept is of community i	Adapted Shannon Entropy index [Shannon 1948]	Positive
		Entropy using class $H(p) = - \sum_{i=1}^{nc} P_i \ln(P_i)$ $P_i =$ probability of a random concept is of class i	Adapted Shannon entropy index [Shannon 1948]	Positive
Exploratory uncertainty	Diversity of types	Entropy using categories $H(p) = - \sum_{i=1}^{nc} P_i \ln(P_i)$ $P_i =$ probability of a random concept is of category i	Adapted Shannon entropy index [Shannon 1948]	Positive
		Coefficient of variation $c_v = \frac{\sigma}{\mu}$	Applied	Negative
	Exploratory uncertainty	Average path length $L = \frac{1}{N(N-1)} \sum_{i,j=1, i \neq j}^N d(c_i, c_j)$	Applied	Positive
		Number of shortest paths Counting of number of shortest paths between all pairs of concepts	Applied	Positive
Diameter The longest of all the shortest paths in a network	Applied	Positive		

- *in simple and complex conceptual networks, it is possible to highlight more representative concepts by the degree of the nodes - power-law distribution;*
- *in simple and complex conceptual networks, prominent sets of concepts highly connected (clusters) are achievable by community detection algorithms.*

5. Strategies to adapt the complexity

In addition to a set of useful features and metrics to estimate the complexity of digital resources, we also focused on identifying strategies to deal with complexity. Often, educators and learning designers need to adjust the complexity of digital resources. It is essential that they can do it, minimizing comprehension problems such as oversimplification and compartmentalization of knowledge [Feltovich et al. 1993][Mandl et al. 1993]. Besides, it is relevant that they comprehend the impacts of the strategies on the complexity features to make decisions about appropriate simplification strategies for varied situations.

As part of the approach to address this problem, we focused on simplification strategies and proposed investigating two research questions:

- (RQ3): What strategies can be adopted to simplify the complexity of digital resources considering their conceptual networks?
- (RQ4): What is the impact of the simplification strategies on the complexity features of digital resources?

5.1. Methodology

We based on a literature review on Education to define simplification strategies to deal with teaching and learning complex educational subjects. The strategies are based and inspired by learning design principles to foster advanced knowledge acquisition and minimize the effects of oversimplification, and knowledge compartmentalization [Feltovich et al. 1993][Mandl et al. 1993].

We evaluated the strategies by an experimental study. After proposing strategies, we applied them in the conceptual networks associated with Wikipedia articles. After that, we recalculated these metrics and compared the result of the Original Conceptual network and the Simplified Conceptual network by the strategies. In this experiment, we considered a set of Regular Wikipedia articles with about 1,400 articles. We hypothesized that the results would be statistically different for all the features after applying the simplification strategy. We conducted hypothesis tests (Wilcoxon-Mann-Whitney test with a 95% confidence level) to compare the results and Vargha and Delaney's A_{12} to measure the effect size.

5.2. Proposed strategies to deal with the complexity of digital resources

We introduced four strategies to deal with the complexity of digital resources based on their conceptual networks. These four strategies are not exhaustive, but they indicate suitable approaches that consider the conceptual network structure. The strategies focus on reducing the conceptual networks maintaining the most relevant concepts and their relations to understand them as an intertwined network of concepts. We also considered strategies that allow expanding the conceptual network incrementally while the introductory learning is achieved. The following four strategies are an answer to the third **research question (RQ3)**:

- **Focusing on clusters** - proposes the simplification of the conceptual networks through their largest communities. By this strategy, we aim to reduce the conceptual network and focus on the most connected concepts on the network;
- **Focusing on the central cluster and most representative concepts** - considers the community composed of concepts strongly connected to a central concept. Therefore, its feasibility depends on distinguishing a central concept in the conceptual network. This strategy includes the main cluster - containing concepts densely related to the main concept - and other most representative concepts.
- **Focusing on paths to explore concepts** - explores sequences of concepts and aims to simplification by eliminating redundant or long paths between concepts. We identify all the shortest paths between pairs of concepts in the network that crosses the central concept. The selection of the shortest paths to explore the concepts reduces the conceptual network through two approaches. First, the strategy eliminates paths that do not include the central concept. Second, it eliminates alternative, longer paths to understand how two pairs of concepts are related;

- **Focusing on patterns of concepts and their interactions** - focuses on selecting concepts that follow patterns with high occurrence in the conceptual network. The patterns are established through the relationships between types (classes) of concepts.

5.3. Results

The results of statistic tests answered the fourth (**research question RQ4**), showing that *all strategies can be applied to simplify the complexity of digital resources through their conceptual network, since each strategy could reduce some features that contribute to the digital resource's complexity, such as the number of concepts, relationships, and length of paths. The analysis also revealed they vary in terms of the impact on the features, advantages, and disadvantages.* Some additional important findings concerning the fourth question, RQ4, are:

- *Strategy 1 - focusing on large clusters - has a low impact on interrelatedness features since the concepts continue highly intertwined and interdependent;*
- *Strategy 2 - focusing on central and the most representative concepts - has a more powerful impact on the number of concepts and also by relationships between them than Strategy 1;*
- *Strategy 3 - Focusing on paths to explore concepts - has a significant impact on the simplification of the concept network, mainly concerning the number of paths among concepts;*
- *Strategy 4 - Focusing on patterns of concepts and their interactions - did not have a significant reduction potential concerning the dimension and interrelatedness among concepts.*

6. Contributions

This research produced conceptual, methodological and technical contributions. The main ones are highlighted in what follows:

- a set of features that contributes to estimating the complexity of digital resources;
- strategies that allow adapting the complexity following learning design principles to support the learning process;
- a framework that includes components, services, tools, techniques, methods, ontologies and datasets to structure the information of digital resources as a conceptual network;
- a method for creating incremental conceptual networks of entire courses to assist educational stakeholders (including learning designers) to minimize the effects of oversimplification and compartmentalization of knowledge during the progress of a course [Pereira et al. 2020];
- an approach to find the ideal order of presentation of information [Pereira et al. 2019] using an initial set of features of complexity applied to a dataset with pairs of prerequisites to analyze differences in each feature according to the precedence of the concepts;
- an extensive systematic mapping about use of Linked Data in Education, including ontologies, vocabularies and knowledge graphs [Pereira et al. 2017b];
- a graph database with about 14,000 graphs representative of Wikipedia articles and graphs created from Computer Science courses that can be used as a benchmark for comparative analysis of Linked Data and Knowledge Representation.

7. Conclusion

The outcomes of this research benefit several educational stakeholders in dealing with the complexity of digital resources and entire courses as well as open new research avenues in the areas of Sequencing[Manrique et al. 2018], Knowledge Tracing[Piech et al. 2015], and Searching as Learning[Gimenez et al. 2020].

In this paper we summarized the main scientific and technical contributions of the thesis, including new approaches to level digital resources complexity; metrics and features to determine the complexity of digital resources based on conceptual networks; and datasets that can be used as benchmark. We also developed tools to visualize and analyse the complexity of digital resources (available in the full thesis).

References

- Bansal, M. and Sharma, D. (2015). Improving webpage visibility in search engines by enhancing keyword density using improved on-page optimization technique. *International Journal of Computer Science and Information Technologies*, 6(6):5347–5352.
- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.
- Berlyne, D. E. (1960). Conflict, arousal, and curiosity. page 303p.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Collins-Thompson, K., Bennett, P. N., White, R. W., De La Chica, S., and Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412.
- Feltovich, P. J., Spiro, R. J., and Coulson, R. L. (1993). Learning, teaching, and testing for complex conceptual understanding. *Test theory for a new generation of tests*.
- Gell-Mann, M. (1995). What is complexity? <i>Remarks on simplicity and complexity by the Nobel Prize-winning author of</i> The Quark and the Jaguar. *Complexity*, 1(1):16–19.
- Gimenez, P., Machado, M., Pinelli, C., and Siqueira, S. (2020). Investigating the learning perspective of searching as learning, a review of the state of the art. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 302–311, Porto Alegre, RS, Brasil. SBC.
- Kursa, M. B., Rudnicki, W. R., et al. (2010). Feature selection with the boruta package. *J Stat Softw*, 36(11):1–13.
- Mainzer, K. and Landauer, R. (1995). Thinking in Complexity: The Complex Dynamics of Matter, Mind, and Mankind. *American Journal of Physics*.
- Malaga, R. A. (2009). Web 2.0 techniques for search engine optimization: Two case studies. *Review of Business Research*, 9(1):132–139.

- Mandl, H., Gruber, H., and Renkl, A. (1993). Misconceptions and knowledge compartmentalization. In *Advances in psychology*, volume 101, pages 161–176. Elsevier.
- Manrique, R., Sosa, J., Marino, O., Nunes, B. P., and Cardozo, N. (2018). Investigating learning resources precedence relations via concept prerequisite learning. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 198–205. IEEE.
- Pereira, C. K., Medeiros, J. F., Siqueira, S. W., and Nunes, B. P. (2019). How complex is the complexity of a concept in exploratory search. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 17–21. IEEE.
- Pereira, C. K., Nunes, B. P., Siqueira, S. W., Manrique, R., and Medeiros, J. F. (2020). ‘a little knowledge is a dangerous thing’: A method to automatically detect knowledge compartmentalization and oversimplification. In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, pages 140–144. IEEE.
- Pereira, C. K., Siqueira, S., and Nunes, B. P. (2017a). Dados conectados na educação. In *6º DesafIE! - Workshop de Desafios da Computação aplicada à Educação, 2017, São Paulo. Anais do XXXVII Congresso da Sociedade Brasileira de Computação - CSBC*.
- Pereira, C. K., Siqueira, S. W. M., Nunes, B. P., and Dietze, S. (2017b). Linked data in education: a survey and a synthesis of actual research and future challenges. *IEEE Transactions on Learning Technologies*, 11(3):400–412.
- Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., and Sohl-Dickstein, J. (2015). Deep knowledge tracing. *arXiv preprint arXiv:1506.05908*.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Silvia, P. J. (2005). What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89.
- Sweller, J. and Chandler, P. (1994). Why some material is difficult to learn. *Cognition and instruction*, 12(3):185–233.
- van der Sluis, F. and van den Broek, E. L. (2010). Using complexity measures in information retrieval. In *Proceedings of the third symposium on information interaction in context*, pages 383–388. ACM.
- Van der Sluis, F., Van den Broek, E. L., Glassey, R. J., van Dijk, E. M., and de Jong, F. M. (2014). When complexity becomes interesting. *Journal of the Association for Information Science and Technology*, 65(7):1478–1500.
- Wiesner, K. and Ladyman, J. (2020). Measuring complexity.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.
- Wu, Q. and Miao, C. (2013). Curiosity: From psychology to computation. *ACM Computing Surveys (CSUR)*, 46(2):18.
- Zyngier, S., Van Peer, W., and Hakemulder, J. (2007). Complexity and foregrounding: In the eye of the beholder? *Poetics Today*, 28(4):653–682.