Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Understanding Learner Behavior Using Information Theory on Learning Analytics and Knowledge

Kensuke TAKII[a*], Changhao LIANG[a] & Hiroaki OGATA[a]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*kensuke.takii96@gmail.com

Abstract: It has been challenging for traditional Learning Analytics (LA) to accurately deal with learners' knowledge acquisition due to its data-driven nature. As a solution to this problem, this paper introduces Information Theory on Learning Analytics and Knowledge (ITLAK), a theoretical framework that quantifies the information value of learning behaviors. By defining the entropy of learning behavior and modeling learning as an information-theoretic process, ITLAK provides principled measures of knowledge engagement and entropy based on learner interactions with knowledge elements. The framework provides various applications, including personalized recommendations, while offering interpretable indicators of learning diversity and progress. It also supports the diagnosis of conceptual imbalance and contributes to theory-informed learning support. ITLAK advances LA by shifting focus from surface-level activity analysis to epistemically grounded models of knowledge acquisition. ITLAK may help bridge theory and practice toward more effective and interpretable learning support through future developments and research, such as empirical implementation and temporal simulation.

Keywords: Information theory, Learning Analytics, entropy, personalized learning

1. Introduction

Learning Analytics (LA) is a research field that aims to understand and optimize learning and the environments in which it occurs by collecting and analyzing data related to learners and their contexts (Ferguson, 2012). While LA has significantly contributed to the visualization of learning processes and the optimization of learning support systems, many existing approaches focus on building statistical or machine learning models based on learners' activity logs (Knight & Buckingham Shum, 2017). As a result, the quantitative analysis of the qualitative aspects of learning, such as what knowledge has been acquired and to what extent, remains insufficiently addressed in a data-driven manner. Consequently, it is challenging to accurately assess learners' knowledge acquisition status, making it difficult to provide appropriate guidance on what learning actions should be taken next.

Takii et al. (2024a) proposed the Open Knowledge and Learner Model (OKLM) to address this issue. OKLM provides a framework that bridges the gap between collected learning logs and the knowledge elements that should be acquired, enabling the estimation of learners' knowledge states. This model has been mathematically formulated (Takii et al., 2024b) and has contributed to linking learning behaviors with knowledge models. However, OKLM lacks a clear metric for quantitatively measuring learners' knowledge acquisition status, and no established method exists for assessing the information value of learning actions.

In this study, we propose Information Theory on LA and Knowledge (ITLAK), a theoretical framework that quantifies the information content of learning behaviors using information theory. ITLAK conceptualizes learning as a process in which learners recognize and engage with concepts, subsequently forming knowledge about them. By measuring changes in information content, ITLAK enables the quantitative evaluation of qualitative aspects of learning. In this paper, we present the mathematical formulation of ITLAK and discuss its theoretical properties and applicability.

2. ITLAK: Theoretical Framework

2.1 Open Knowledge and Learner Model: The Foundation

OKLM is a mechanism for associating the learning logs collected from the learning support systems by LA technology with the knowledge elements contained in the learning materials (Figure 1). The learning logs collected from the learning support systems correspond to each learning action the learner has taken in the systems. Each learning action has information such as the target learning material and its section. Therefore, it is possible to associate the learning log with the target knowledge element by linking it with the information about the knowledge element in the learning material.

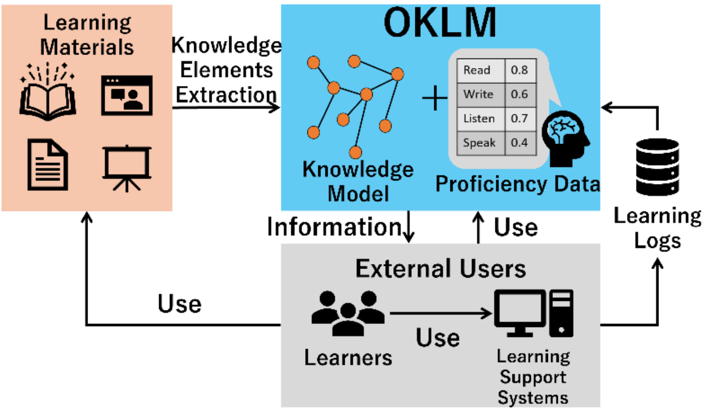

Figure 1. *Conceptual diagram of OKLM.*

2.2 Why Entropy?

Entropy, as a measure of uncertainty in information theory, offers a principled way to represent the diversity and unpredictability of learner behavior (Shannon, 1948). Let $p(x_i)$ be the probability that the random variable X takes the value x_i. The information content obtained when such an event occurs is defined as $-\log p(x_i)$. The measure of uncertainty, or entropy, of X is defined as $H(X) = -\sum_i p(x_i) \log p(x_i)$. Entropy shows the degree to which the probability distribution of an event is random or unpredictable. The smaller the entropy, the greater the certainty about the next event, and vice versa.

In educational settings, such unpredictability is not noise but a reflection of the inherent heterogeneity in learners' knowledge, goals, and learning strategies. Even when engaging with the same learning materials, learners differ substantially in interpreting, prioritizing, and organizing knowledge elements. Consequently, the trajectory of what a learner might study next is often highly individualized and cannot be captured by aggregating behavioral metrics alone. Traditional indicators, such as activity counts, test scores, or response accuracy, are limited in this context. For example, even if a learner repeatedly learns a particular knowledge element, it is not intuitive to assume that all those learning behaviors are of equal value in learning. Conversely, even if a learner has learned the various knowledge elements evenly, the values of these learning behaviors should vary depending on the learner's past learning. Thus, they may reflect the quantity of learning but not its conceptual diversity or depth. This limitation highlights the need for metrics to assess structural distribution of the engagement in knowledge across multiple dimensions.

Information content and entropy address this need by quantifying how concentrated or dispersed a learner's engagement is across knowledge elements. This ability to represent diversity and uncertainty makes entropy especially suitable for modeling learner knowledge states, learning progress, and readiness for further instruction. The goal of ITLAK is thus to accurately capture the uncertainty and diversity of educational data and link it to learning support, and entropy is a central concept in fulfilling this goal.

Similar uses of entropy exist in other domains: in marketing, to model consumer behavior diversity (Lesser & Lusch, 1988); in ecology, to measure species richness and ecosystem resilience (Cushman, 2023). Thus, entropy is widely used to measure data diversity and information content. Likewise, in education, entropy provides a way to characterize learning diversity, detect imbalances, and support personalized interventions. By interpreting learning logs using entropy, learning logs are transformed into useful information that expresses the quality of the learner's learning and knowledge state, rather than just being an accumulation of data.

2.3 Information Content and Entropy of Learning Logs

This section formalizes the information quantity and entropy associated with learning logs to analyze the information inherent in learning behaviors. Although the analysis in this paper is based on the Experience API (xAPI) format (xAPI.com, 2011), which structures learning logs with specific components, this theoretical framework also applies to other formats.

2.3.1 Learning Logs and Knowledge Elements

To define the information content and entropy of the learning behavior, it is necessary to calculate the probability that the next learning log that the learner leaves behind is involved in learning a particular knowledge element. Here, let us assume learning logs are in the form of xAPI, a data specification to record and track learning experiences for e-learning systems. Each learning log l in xAPI format can be expressed as **a tuple of Actor, Verb, Object, Context, Result, and Timestamp**, capturing the learner's interaction with content. The set of all learning logs recorded by learner s up to time t is denoted as $L_{s,t} = \{l \colon \mathrm{Actor}(l) = s, \mathrm{Timestamp}(l) \leq t\}$.

To link learning logs to knowledge elements, we define $n_{s,t}(v)$ as the number of learning logs in $L_{s,t}$ that are associated with knowledge element v:

$$n_{s,t}(v) = \left|\left\{l \in L_{s,t} \colon v \in k(l)\right\}\right|,$$

where k is a function that maps a learning log to a set of knowledge elements.

2.3.2 Probability of Future Learning Behavior

Let $E^{V}_{s,t}(v)$ denote the event that the next learning log of learner s after time t is associated with knowledge element v from the set V. The probability of this event, $p\left(E^{V}_{s,t}(v)\right)$, is defined as follows to satisfy the axioms:

$$p\left(E^{V}_{s,t}(v)\right) = \frac{n_{s,t}(v) + \alpha}{\sum_{v' \in V}\left(n_{s,t}(v') + \alpha\right)},$$

where $\alpha > 0$ is a small smoothing parameter to address the zero-frequency problem. This probability reflects a learner's likelihood of engaging with a particular knowledge element based on their past learning logs.

2.3.3 Information Content and Entropy of Future Learning

The information content obtained when a learner engages with knowledge element v is defined as $i\left(E^{V}_{s,t}(v)\right) = -\log p\left(E^{V}_{s,t}(v)\right)$. This measure captures the amount of uncertainty reduction associated with the learner's engagement with v.

To quantify the overall uncertainty in the learner's future interactions with the knowledge elements in V, we define the entropy $H\left(\mathbb{E}^{V}_{s,t}\right) = -\sum_{v \in V} p\left(E^{V}_{s,t}(v)\right) \log p\left(E^{V}_{s,t}(v)\right)$. This entropy represents the diversity and distribution of a learner's potential interactions with knowledge elements in V. A higher entropy indicates a broader range of possible interactions, whereas lower entropy suggests focused engagement on specific knowledge elements.

These metrics provide a data-driven perspective on how learners interact with knowledge elements by quantifying the information and entropy of learning behaviors. These formulations lay the groundwork for analyzing learning behaviors in a principled manner and contribute to optimizing personalized learning pathways.

Unlike many conventional LA approaches that depend on surface-level indicators such as frequency, correctness, or time spent, ITLAK enables a structural understanding of how learning behaviors contribute to knowledge development. Its information-theoretic nature allows it to assess the diversity and specificity of knowledge engagement without reliance on predefined mastery models. This positions ITLAK as a theoretically grounded yet practical alternative to purely statistical modeling, enabling both personalized feedback and population-level diagnostics in a mathematically principled way.

3. Potential Applications

ITLAK provides a unified framework for quantifying the informational content of learning actions based on their association with specific knowledge elements. This enables a wide range of applications in learning support and educational research.

3.1 Understanding Learning Status and Diversity

The entropy $H(\mathbb{E}_{s,t}^{V})$ serves as an indicator of how predictable a learner's next knowledge engagement is. A low entropy suggests focus or imbalance in knowledge, while high entropy may reflect exploratory but fragmented learning. These interpretations enable fine-grained diagnosis of learning behavior and support strategies such as broadening a narrow focus or scaffolding dispersed attention.

3.2 Recommendation System

By recommending knowledge elements with higher information content $i\left(E_{s,t}^{V}(v)\right)$, ITLAK can guide learners toward less familiar but potentially fruitful areas of learning. Since these values are dynamically updated, the system can provide real-time personalized recommendations. Furthermore, comparing the probability distributions $p\left(E_{s,t}^{V}(v)\right)$ across learners allows for collaborative filtering: identifying similar learners and suggesting knowledge based on successful learning patterns within peer groups.

Besides, ITLAK allows for clustering learners by their information profiles, which enables grouping them by learning styles or proficiency levels. This supports the design of the recommendations for interventions. Applying clustering to knowledge elements based on their informational role for learners can also provide a suitable curriculum design and learning sequence.

4. Conclusion and Future Work

This study proposed ITLAK, a theoretical framework that enables the quantification of learning behaviors through their informational contributions. By formalizing the information content and entropy of learner interactions with knowledge elements, ITLAK shifts the analytical focus from mere behavioral frequency to a principled evaluation of knowledge acquisition.

The strength of ITLAK lies in its theoretical clarity and its potential to reinterpret learning as an informational process. Unlike traditional LA approaches that often rely on coarse performance metrics or static modeling of learner behavior, ITLAK offers a continuous, interpretable, and adaptable framework. This opens a new dimension for understanding how knowledge is acquired, reinforced, or forgotten over time.

Looking ahead, ITLAK is expected to contribute to various layers of educational practice and theory:

- **At the system level**, it can be the backbone of new learning support systems that offer real-time, interpretable feedback based on learners' knowledge trajectories.
- **At the pedagogical level**, ITLAK enables instructors to design data-informed, adaptive curricula sensitive to learners' cognitive states and conceptual balance.
- **At the research level**, future research may lead to constructing a new entropy-based learning theory and its validation.

Furthermore, ITLAK may contribute to theory-informed analysis by relating entropy-based patterns to conceptual models such as the Zone of Proximal Development (Vygotsky & Cole, 1978) or constructivist learning (Bruner, 1997), though further empirical work is needed. In practical terms, entropy-based indicators may help educators identify learners who are conceptually stagnant or overly narrow in their engagement, supporting timely and targeted interventions.

For future work, several directions will be pursued. For example, we can develop analytical tools using ITLAK, which enable us to visualize entropy-based indicators to support teachers and learners in real-time. This can contribute to further research comparing ITLAK to existing methods from predictive or explanatory perspectives. Another example is agent-based or temporal simulation studies to explore the effect of various learning strategies on entropy and knowledge development. This can also pave the way to optimizing learning paths targeting individual learners and peer group dynamics. Through these developments, ITLAK aspires to become a tool for understanding learning and a theoretical foundation for redefining education as an information process.

Acknowledgements

This work was supported by CSTI SIP Program Grant Number JPJ012347, JST Grant Number JPMJFS2123, and JSPS KAKENHI Grant Number 23H00505 and 25K21357.

References

Bruner, J. S. (1997). The culture of education. In *The culture of education*. Harvard university press.

Cushman, S. A. (2023). Entropy, ecology and evolution: toward a unified philosophy of biology. *Entropy*, *25*(3), 405.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International journal of technology enhanced learning*, *4*(5-6), 304-317.

Knight, S., & Buckingham Shum, S. (2017). Theory and learning analytics. *Handbook of learning analytics*, *1*, 17-22.

Lesser, J. A., & Lusch, R. F. (1988). Entropy and the prediction of consumer behavior. *Behavioral Science*, *33*(4), 282-291.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Takii, K., Koike, K., Horikoshi, I., Flanagan, B., & Ogata, H. (2024a). OKLM: A universal learner model integrating everyday learning activities with knowledge maps. In *Companion Proceedings 14th International Conference on Learning Analytics & Knowledge, Japan* (pp. 191-193).

Takii, K., Liang, C., & Ogata, H. (2024b). Open Knowledge and Learner Model: Mathematical Representation and Applications as Learning Support Foundation in EFL. In *Proceedings of 32nd International Conference on Computers in Education*. (pp. 595-604)

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

xAPI.com. (2011). *What is an LRS? Learn more about Learning Record Stores*. https://xapi.com/learning-record-store/