Learning Analytics in Higher Education - A Literature Review

Philipp Leitner, Mohammad Khalil and Martin Ebner

Educational Technology, Graz University of Technology, Austria
{philipp.leitner, mohammad.khalil, martin.ebner}@tugraz.at

Abstract  Higher Education institutes share the same challenges as businesses. They need to increase financial and operational efficiency, expand local and global impact, establish new funding models during a changing economic climate and respond to the demands for greater accountability to ensure organizational success at all levels [1]. The pressing needs to provide solutions to various quandaries and improve the quality of educational establishments such as Higher Education, paved the road to emerge a field called Learning Analytics. This field has proven to provide appropriate tools for analyzing the academic progress of learners, to predict future performance and to identify potential problems at an early stage. This chapter looks into examining research studies of the last five years and presents the state of the art of Learning Analytics in the Higher Education arena. Further, it presents an overview over the used techniques and clusters the publications into their stakeholders. Finally, the authors tackled the limitations of the previous literature and discuss the most promising future lines and challenges.
Introduction

The amount of available data to explore has grown over the last years. Therefore, a new research field has been established under the term “Big Data”. It involves big amounts of data that cannot be evaluated with traditional methods of data processing because of their size, high complexity, rapid transience or weak structure. According to McKinsey Global Institute, Big Date is defined as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” [2]

The aim of Learning Analytics is to collect Big Data in the context of teaching and learning, further to analyze and interpret it to gain new insights and to provide the stakeholders with new models for improving teaching, learning, effective organization, and decision making [3]. A key fact is the return of the resulting knowledge to the teachers and students to optimize their teaching and learning behavior, to promote the development of skills in the area, and to better understand education as well as the connected fields, e.g. university business and marketing. Available resources can be used more efficiently to provide better support and individual care to develop potentials.
In the area of Higher Education, Learning Analytics has proven to be helpful to colleges and universities in strategic areas such as resource allocation, student success and finance. These institutions are collecting more and more data than ever before, to maximize strategic outcomes. Based on key questions data is analyzed and predictions are made to gain insights and set actions. Many examples of successful analytics and frameworks use are available across a diverse range of institutions and businesses [4]. Ethical and legal issues of collecting and processing students’ data are seen as barriers by the Higher Education institutions in Learning Analytics [5].

In this chapter, we present a literature review to evaluate the progress of Learning Analytics in Higher Education since its early beginning in 2011. Therefore, we searched through three popular libraries, including the Learning Analytics and Knowledge (LAK) conference, the SpringerLink, and the Web of Science databases. We then refined the returned results and settled on including 101 relevant publications. This chapter mainly contributes by analyzing them and list the used Learning Analytics methods, limitations and stakeholders. We believe the results of this review may assist universities to launch their own Learning Analytics projects or improve existing ones.

The next section gives a short introduction on the topic of Learning Analytics and describes Learning Analytics in Higher Education in detail. The subsequent sections are concerned with our research design, methodology and execution of the review. The outcomes of the research questions and the literature survey are presented in the fourth section. Afterwards we discuss the findings of the survey.
The last section gives a conclusion as well as a glance of future trends.

Background

Learning Analytics

Since its first mention in the Horizon Report 2012 [6], Learning Analytics has gained an increasing relevance. Learning Analytics is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts for purposes of understanding and optimizing learning and the environments in which it occurs" [7]. The Horizon Report 2013 identified Learning Analytics as one of the most important trends in technology-enhanced learning and teaching [8]. Therefore, it is not surprising, that Learning Analytics is the subject of many scientific papers. The research and improvement of Learning Analytics involves doing the development, the use and integration of new processes and tools to improve the performance of teaching and learning of individual students and of teachers. Learning Analytics focuses specifically on the process of learning [3]. Due to its connections with digital teaching and learning, Learning Analytics is an interdisciplinary research field with connections to the field of teaching and learning research, computer science and statistics [8]. The available data is collected, analyzed and the gained insights are used to understand the behavior of the students to provide them additional support [9].
A key concern of Learning Analytics is the gathering and anal-
yzation of data as well as the setting of appropriate interventions to im-
prove the learners learning experience [10]. These “actionable intel-
ligence” from data mining is supporting the teaching and learning
and provides ideas for customization, tutoring and intervention within
the learning environment [11]. According to [12] an analysis pro-
cess has five steps:

- **Capturing**
  Data is captured and collected in real-time from different sources
  (e.g. virtual learning environments, learning management sys-
tems, personal learning environment, web portals, forums, chat
rooms, and so on) and combined with student information.
  [13,14]

- **Reporting**
  The collected data is used to generate accurate models for identi-
fying and measuring the student’s progress. Often visualization is
used in Learning Analytics dashboards for a better understanding
of the data. [15]

- **Predicting**
  The data is used to identify predictors for student success, out-
comes and for identifying at risk students. Further, it is used for
decision making about courses and resource allocation which
then is used by the decision-makers of the institutions. [13]

- **Acting**
  The information gained from the data analyzation process is used
to set appropriate interventions in e.g. teaching or supporting
students who are at risk of failure or dropping out. [14]
Refining

The gathered information is used in a cyclical process for continuous improvements of the used model in teaching and learning.

Verbert and her colleagues [16] provided a guidance for the relevancy of datasets in learning and knowledge analytics research and identified a set of relevant objects for learning and knowledge analytics applications:

- Predicting learner performance and modeling learners.
- Estimate unknown values of variables which describe the learner, e.g. knowledge, grades, scores. [17]
- Suggestion of relevant learning resources.
- The learner data is analyzed for suggesting relevant learning resources, finding peer learners or learning paths.
- Increasing reflection and awareness.
- Focuses on the analysis and visualization of different learning indicators (such as: access to resources, time spend and knowledge level indicators) to foster awareness and reflection on the learning process [18].
- Enhancing social learning environments.
- Social interactions are analyzed and visualized to make people aware of their social context [19].
- Detecting undesirable learner behaviors
- Detect learners who may have problems, are at risk to drop out or fail, or show unusual behavior (e.g. misuse or cheating). [17]
- Detecting effects of learners
• Detect a student state of boredom, confusion, frustration, engagement [20].

Although research in the field of Learning Analytics in recent years celebrates boom, Learning Analytics is still in its infancy.

**Learning Analytics in Higher Education**

Higher Education looks forward to a future of uncertainty and change. In addition to the national and global as well as political and social changes, the competition on university level increases. Higher Education institutes share the same challenges as businesses - the need to increase financial and operational efficiency, expand local and global impact, establish new funding models during a changing economic climate and respond to the demands for greater accountability to ensure organizational success at all levels [1].

Higher Education must overcome these external loads in an efficient and dynamic manner, but also understand the needs of the student body, whom represents contributor as well as donor of this system [21].

In addition to the strong competition, universities have to deal with the rapidly changing technologies that have arisen with the entry of the digital age. In the course of this, institutions have entered the era of Big Data and collected enormous amounts of relevant data as a by-product. For instance, when students take an online course, use an intelligent tutoring system, play educational games or simply use an online learning platform.
In recent years, more universities use methods of Learning Analytics in order to obtain findings on the academic progress of students, predict future behaviors and recognize potential problems in an early stage. Further, Learning Analytics in the context of Higher Education is an appropriate tool for reflecting the learning behavior of students and provide suitable assistance from teachers or tutors. This individual or group support offers new ways of teaching and provides a way to reflect on the learning behavior of the student. Another motivation behind the use of Learning Analytics in universities is to improve the inter-institutional cooperation, and the development of an agenda for the large community of students and teachers [22].

On an international level, the recruitment, management and retention of students have become as high level priorities for decision makers in institutions of Higher Education. Especially improving the student retention starts and the understanding of the reason behind and/or prediction of the attrition has come in the focus of attention due to the financial losses, lower graduation rates, and inferior school reputation in the eyes of all stakeholders [23, 24].

Despite that Learning Analytics focuses strongly on the learning process, the results still in the beneficial for all stakeholders. Romero & Ventura [25] divided the involved stakeholders based on their objectives, benefits and perspectives in the following four groups:

- **Learners**
  Objective examples: Support the learner with adaptive feedback, recommendations, response to his or her needs, for learning performance improvement.
• **Educators**
  Objective examples: understand students’ learning process, reflect on teaching methods and performance, understand social, cognitive and behavioral aspects.

• **Researchers**
  Use the right data mining technique which fits the problem, evaluation of learning effectiveness for different settings.

• **Administrators**
  Evaluation of institutional resources and their educational offer.

**Research Design, Methodology and Execution**

This research aims at the elicitation of the state of the art on the advancement of the Learning Analytics field in Higher Education since it emerged in 2011. The research questions are:

• **RQ1:** What are the research strands of the Learning Analytics field in Higher Education (between January 2011 and February 2016)?

• **RQ2:** What kind of limitations do the research papers and articles mention?

• **RQ3:** Who are the stakeholders and how could they be categorized?

• **RQ4:** What methods do they use in their papers?

In accordance to this objective, we performed a literature review following the procedure of Machi and McEvoy [26]. Figure 1 displays the steps of the used process.
After we selected our topic, we identified data sources based on their relevance in the computing domain:

- the papers of the Learning Analytics and Knowledge conference published in the ACM Digital Library,
- the SpringerLink, and
- the Thomson Reuters Web of Science database

and the following search parameters:

In the LAK papers, we didn’t need to search for the “Learning Analytics” term because the whole conference covers the Learning Analytics discipline. We searched the title, the abstract and the author keywords for “Higher Education” and/or “University”.

In the SpringerLink database, we searched for the “Learning Analytics” term in conjunction with either “Higher Education” or “University” (“Learning Analytics AND (Higher Education OR University)”).
In the Web of Science database, we searched for the topic “Learning Analytics” in conjunction with either “Higher Education” or “University” and in the research domain “science technology”.

The defined inclusion criteria of the fetched papers from the libraries were set to be: a) written in English, and b) published between 2011 till the February 2016. We superficially assessed the quality of the reported studies, considering only articles that provided substantial information for Learning Analytics in Higher Education. Therefore, we excluded articles that did not meet the outlined inclusion principles.

The literature survey was conducted in February and March 2016. In the initial search, we found a total of 135 publications (LAK: 65, SpringerLink: 37, Web of Science: 33). During the first stage, the search results were analyzed based on their titles, author keywords and abstracts. After this stage, 101 papers remain for the literature survey. We fully read each publication and actively searched for their research questions, techniques, stakeholders, and limitations. Regular meetings between the authors were set on a weekly basis to discuss the results. Additionally, we added to our spreadsheet the Google Scholar (http://scholar.google.com) citation count as a measurement of article’s impact.
Analysis of Learning Analytics in Higher Education

Publication

In order to present our findings, we analyze each of the research questions separately.

RQ1: What are the research strands of the Learning Analytics field in Higher Education (between January 2011 and February 2016)?

In order to answer this question, we tried to extract the main topics from the research questions of the publications. We identified that many of the publications do not outline their research questions clearly. Many of the examined publications described use cases. This concerns in particular the older publications of 2011 and 2012, and is probably resulting from the young age of the scientific field of Learning Analytics. As a result, we did a brief text analysis on the fetched abstracts in order to examine the robust trends in the prominent field of Learning Analytics and Higher Education. Word cloud was our selection of representing the terms as shown in figure 2.
Fig. 2. Word cloud of the prominent terms from the abstracts

In order to ease reading the cloud, we adopted four levels of representation depicted in four colors. The obtained list of words that have been used were classified into singular phrases, bi-grams, tri-grams and quad-grams. The most cited singular words were “university”, “academic”, “performance”, “behavior” and “MOOCs”. “Learning Analytics”, “Higher Education”, “learning environment”, “case study” and “online learning” were the most repeated bi-grams. The highest tri-grams used in the abstracts were “learning management systems”, “Higher Education institutions” and “social network analysis”. While quad-grams were only limited to “massive open online courses” which were merged at the final filtering stage with the “moocs” term.

The word cloud shows a glance about the general topics when Learning Analytics is ascribed with Higher Education. Learning Ana-
lytics researchers focused on utilizing its techniques towards enhancing performance and students’ behaviors. The popular adopted educational environment was MOOC platforms. Furthermore, Learning Analytics was also used to perform practices of interventions, observing dropout, videos, dashboards and engagement.

Figure 3 displays the collected articles from the library data sources. Results show an obvious increase in the number of publications since 2011. For instance, there were 32 papers in 2015, incremented from 26 articles in 2014 and 17 articles in 2013. However, there were 5 articles only in 2011 and 12 articles in 2012. Because February 2016 was the date of collecting the publications in this study, the 2016 year was not indexed with many papers. On the other hand, the figure shows the apparent involvement of the journal articles from the SpringerLink and Web of Science libraries from 2013.

![Figure 3. Collected articles distribution, by source and year.](image-url)
We cross-referenced the relevant publications with Google Scholar to derive their citation impact. Table 1 shows the 10 most cited publications.

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>Year of Publication</th>
<th>No. of Google Citations (Feb. 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course Signal at Purdue: Using Learning Analytics to Increase Student Success [27]</td>
<td>2012</td>
<td>164</td>
</tr>
<tr>
<td>Social Learning Analytics: Five Approaches [28]</td>
<td>2012</td>
<td>94</td>
</tr>
<tr>
<td>Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers [29]</td>
<td>2011</td>
<td>52</td>
</tr>
<tr>
<td>Where is Research on Massive Open Online Courses Headed? A Data Analysis of the MOOC Research Initiative [31]</td>
<td>2014</td>
<td>46</td>
</tr>
<tr>
<td>Course Correction: Using Analytics to Predict Course Success [32]</td>
<td>2012</td>
<td>36</td>
</tr>
<tr>
<td>Improving retention: predicting at-risk students by analyzing clicking behavior in a virtual learning environment [33]</td>
<td>2013</td>
<td>34</td>
</tr>
<tr>
<td>Learning designs and Learning Analytics [34]</td>
<td>2011</td>
<td>33</td>
</tr>
<tr>
<td>The Pulse of Learning Analytics Understandings and Expectations from the Stakeholders [35]</td>
<td>2012</td>
<td>30</td>
</tr>
<tr>
<td>Inferring Higher Level Learning Information from Low Level Data for the Khan Academy Platform [36]</td>
<td>2013</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1. Citation impact of the publications

RQ2: What kind of limitations do the research papers and articles mention?

We identified three different limitations, either clearly mentioned in articles or being tacitly within the context.
1. Limitations through time

Some of the publications stated that continuous work is needed [37]. Either a longitudinal study would be necessary to prove hypotheses [38], or because of the shortage of the project [39].

2. Limitations through the size

Other publications talked about the need for more detailed data [40], the small group sizes [41], the unsure scalability, possible problems in wider context [42] and the problem of the generalization of the approach or method [43].

3. Limitations through the culture

Many of the publications mention that their approach might only work in their educational culture and is not applicable somewhere else [44]. Additionally, the ethics differ strongly around the world, so cooperation projects between different universities in different countries needs different moderation [45]. Also, the use of data could be ethically questionable [28].

Furthermore, ethical discussions about data ownership and privacy have recently arisen. Slade & Prinsloo [46] pointed out that Learning Analytics touches various research areas and therefore overlaps with ethical perspectives in areas of data ownership and privacy. Questions about who should own the collected and analyzed data were highly debated. As a result, the authors classified the overlapping categories in three parts:

- the location and interpretation of data,
- informed consent, privacy and the de-identification of data, and
- the management, classification and storage of data.
These three elements generate an imbalance of power between the stakeholders which they addressed by proposing a list of 6 grounding principles and considerations: Learning analytics as moral practice, Students as agents, Student identity and performance are temporal dynamic constructs, Student success is a complex and multidimensional phenomenon, Transparency, Higher education cannot afford to not use data. [46]

**RQ3: Who are the stakeholders and how could they be categorized?**

In order to answer this question, we determined the stakeholders from the publications and categorized them into three types. As a basis, we took the four stakeholders introduced in [26] and merged the Researchers and Administrators from the original classification into one distinct group. Therefore the institutional perspective (Academic Analytics) is separated from the learners’ and teachers’ one (Learning Analytics).

Figure 4 depicts the defined Learning Analytics stakeholders as a VENN-Diagram. The figure shows that there had been more research conducted concerning the Researchers/Administrators with overall 65 publications and 40 of them only concerning themselves, than in the field of Learners with a total of 53 publications and 21 single mentions. Also, it seems that Teachers are only a “side-product” of this field with only 20 mentions and only 7 dedicated to them alone.
Fig. 4. VENN-diagram of stakeholders in the publications

Most of the combined articles addressed Researchers/Administrators together with Learners (20 publications). Only 8 articles can be found with an overlap between Learners and Teachers, which should be one of the most researched and discussed combinations within Learning Analytics in Higher Education. Nearly no work has been done by combining Researchers/Administrators with Teachers (in 1 publications) and only 4 paper combined all 3 stakeholders. This lack of research will be a matter of debate in the discussion section.

**RQ4: What techniques do they use in their papers?**

By analyzing the selected studies, we identified the techniques used in Learning Analytics and Higher Education publications. We took into account the methods presented by Romero & Ventura [25],
Khalil & Ebner [47], and Linan & Perez [48]. We propose an overview of the used techniques of the different articles in table 2.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Goal/Description</th>
<th>Key applications</th>
<th>Citations Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>To infer a target attribute or single aspect of the data from some combination of other aspects of the data.</td>
<td>Predicting student performance and detecting student behaviors.</td>
<td>36</td>
</tr>
<tr>
<td>Clustering</td>
<td>To identify groups of instances that are similar in some respect.</td>
<td>Grouping similar materials or students based on their learning and interaction patterns.</td>
<td>18</td>
</tr>
<tr>
<td>Outlier Detection</td>
<td>To discover data points that are significantly different than the rest of the data.</td>
<td>Detection of students with difficulties or irregular learning processes.</td>
<td>29</td>
</tr>
<tr>
<td>Relationship Mining</td>
<td>To identify relationships between variables and normally to encode them in rules for later use.</td>
<td>Identifying relationships in learner behavior patterns and diagnosing student difficulties.</td>
<td>10</td>
</tr>
<tr>
<td>Social Network Analysis</td>
<td>To understand and measure the relationships between entities in networked information.</td>
<td>Interpretation of the structure and relations in collaborative activities and interactions with communication tools.</td>
<td>15</td>
</tr>
<tr>
<td>Process Mining</td>
<td>To extract process related knowledge from event logs recorded by information system to have a clear visual representation of the whole process.</td>
<td>Reflecting student behavior in terms of its examination traces, consisting of a sequence of course, grade and timestamp.</td>
<td>4</td>
</tr>
<tr>
<td>Text mining</td>
<td>To derive high-quality information from text.</td>
<td>Analyzing the contents of forums, chats, web pages and documents.</td>
<td>6</td>
</tr>
<tr>
<td>Distillation</td>
<td>To represent data in intelligible</td>
<td>Helping instructors to</td>
<td>33</td>
</tr>
</tbody>
</table>
Data for Human Judgment ways using summarization, visualization and interactive interfaces to highlight useful information and support decision-making. visualize and analyze the ongoing activities of the students and the use of information.

Discovery with Models To use a previously validated model of a phenomenon as a component in another analysis such as prediction or relationship mining. Identification of relationships among student behaviors and characteristics or contextual variables. Integration of psychometric modelling frameworks into machine-learning models.

Gamification To use game elements in a non-gaming system to solve problems and improve user experience as well as user engagement. Include possibilities for playful learning to maintain motivation; e.g. integration of achievements, experience points or badges as indicators of success.

Machine Learning To automate analytical model building using algorithms that iteratively learn from data. Find hidden insights in data automatically (based on models who are exposed to new data and adapt itself independently).

Statistic To numerical state facts in any department of inquiry placed in relation to each other. (A.L. Bowel) Analysis and interpretation of quantitative data for decision making.

Table 2. Overview of the 12 techniques used

The results of table 2 show that the research is focused mainly on prediction [27, 49, 50] with a total of 36 citations. Outlier detection for pointing out at-risk or dropping out students [33, 51, 52] with a
citation count of 29. Distillation of data for human judgment in form of a visualization [30, 52] with a citation count of 33 than in all other parts including rarely used techniques like gamification [54] or machine learning [55] with a total amount of 102 counts.

Conclusion & Future Trends

In this chapter, we examined over 1000 pages during this study. We presented the state of the art of Learning Analytics and Higher Education based on analyzing articles from three major library references of the Learning Analytics field.

First, we identified the research strands of the relevant publications. Most of the publications described use cases rather than comprehensive research - especially the prior publications, which is comprehensible because at the time, the universities had to figure out how to handle and harness the abilities offered by Learning Analytics for their benefit. Further, we discussed the limitations of the publications and categorized them into time, size and cultural limitations. Thereafter, we clustered the publications into stakeholders and showed that especially at an infant stage of Learning Analytics in Higher Education the difference between the use of Learning Analytics and Academic Analytics was very similar. At last, we gave an overview of the techniques used in the publications.

At the end, we are going to tackle the future development in the field of Learning Analytics in Higher Education, which can be divided into short-term (1-2 years) and long term (3-5 years) trends.

Over the next 1 to 2 years, universities must adjust to the social and
economic factors, which postulated the change in the capabilities of the students [56].

The tuning of the areas analysis, consultation, examination of individual learning outcomes and the visualization of continuously-available, aggregated information in dashboards are gaining more and more importance. Students expect real-time feedback during learning with critical self-reflection on the learning progress and learning goal which strengthens their expertise in self-organization.

If adequate quantities of data from students are available, they can be carried out for subsequently, predictive analytics. [56]

The relevance of Learning Analytics in Higher Education will mint even more over the next 3 to 5 years. This trend is promoted by the strong interest of students for individual evaluations and care. To serve this market, dashboards and analysis applications that specifically address the needs of each customer will develop stronger.

This approach offers many advantages: Accessing your own data in an appropriate form allows better self-reflection and a healthy rivalry among the fellow students.

The teachers can survey a large amount of students and precisely recognize those who need their help. University and college drop-outs can be better detected by appropriate analyzing and with targeted interventions they remain in the university system. [21]

To master the associated problems, the Learning Analytics market will have to change. Currently, many different systems and analytical approaches are used. The fragmentation of the market will grow even further in the future, which makes the interuniversity comparison very difficult or even impossible. Therefore, the creation of standards is essential. [21]
Furthermore, a change in the type of analysis is foreseeable. Most current and past data have been used to measure the success of students. Today, advances in predictive analytics (predictive analysis) are more important. By using the analysis of existing data sets of many students, predictive models can be developed and warn thus students who are at risk not to meet their learning success.

[21]

Acknowledgments This research project is co-funded by the European Commission Erasmus+ program, in the context of the project 562167-EPP-1-2015-1-BE-EPPKA3-PI-FORWARD

References


