Differentiating between Educational Data Mining and Learning Analytics: A Bibliometric Approach

Stevens Dormezil
Dept. of Research and Evaluation
3300 Forest Hill Boulevard
West Palm Beach, FL 33406
stevens.dormezil@palmbeachschools.org

Taghi Khoshgoftaar
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431
khoshgof@fau.edu

Federica Robinson-Bryant
Embry-Riddle Aeronautical University
600 South Clyde Morris Boulevard
Daytona Beach, FL 32114-3900
robinsof@erau.edu

ABSTRACT

Educational Data Mining and Learning Analytics are two relatively new research fields. Natural language techniques can be used to identify major research themes within each field. Similarities and differences between both domains are identified through the use of keyword analysis. Over 4,000 articles are analyzed and bibliometric techniques are used to select 60 articles that best represent major research themes within the intersection, as well as disjoint elements of both fields. Following keyword analysis, we conclude it is more accurate to describe what appears to be two domains (i.e. Educational Data Mining and Learning Analytics) as one domain (i.e. Learning Analytics) with one prominent subset (i.e. Educational Data Mining).

Keywords

educational data mining, learning analytics, bibliometrics, big data, machine learning, natural language processing

1. INTRODUCTION

Educational Data Mining (EDM) and Learning Analytics (LA) are two burgeoning disciplines within the fields of Machine Learning and Data Analytics. Although both fields are closely related and share a significant amount of overlap, there exists a subtle if not clear difference between them. Using bibliometric methods, this paper attempts to help highlight the overlap between EDM and LA while identifying key areas of distinction between both fields.

As defined by the Journal of Educational Data Mining, "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in." The Journal of Learning Analytics defines Learning Analytics as "… research into the challenges of collecting, analysing and reporting data with the specific intent to improve learning." The Society for Learning Analytics Research (SoLAR), which publishes the Journal of Learning Analytics, goes further to describe its role as "… an inter-disciplinary network of leading international researchers who are exploring the role and impact of analytics on teaching, learning, training and development."

Although Educational Data Mining and Learning Analytics have developed into two prominent communities, it is not prudent to consider the progress in one without considering the latest work being performed in the other. There is quite a bit of crossover between both fields with practitioners oftentimes making significant contributions in both fields. The synergistic and symbiotic relationship between the two, oftentimes leads many to use the terms Educational Data Mining and Learning Analytics interchangeably. This practice is problematic as both fields do have areas of unique research focus. The aim of this paper is to apply bibliometric approaches to explore prominent research themes within the fields of EDM and LA, as well as the intersection between the two, and identify influential sources within its literature.

This paper is organized as follows: Section 2 gives a brief overview of bibliometric techniques utilized in this study; Section 3 describes the methodology utilized to identify key areas of research within each targeted domain and key sources; Section 4 describes the results obtained; and Sections 5 and 6 captures final conclusions and opportunities for future work.

2. BIBLIOMETRIC TECHNIQUES

Bibliometrics is the use of statistical methods for the analysis of journal articles, books, publications and other works. Bibliometrics can be used as a key component in determining the impact of research in the scientific community and society [1] and can be used for both structural and conceptual purposes. Quantitative methods of bibliographic analysis include citation analysis, citation graphs, impact factors, Hirsch numbers, and altmetrics. However, both qualitative and quantitative measures are utilized to capture the productivity and quality of the work under study.

Often times, the relative importance of an article, author, or publication is measured through citation analysis. It can be used to study "…knowledge flows, the diffusion of ideas, intellectual structures of science, relevance of information resources, and evaluation of researchers and research institutions" [2]. Citations are used by a citing author to indicate use of the cited work. Therefore, citations can be viewed as an indicator of the relatedness of works. A citation relationship can manifest itself in

CEUR-WS.org/Vol-2592/short1.pdf

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

one of three ways: inter-citation counts, co-citation counts, or bibliographic coupling frequencies [2]. Inter-citation counts represent the frequency two objects have cited each other. Co-citation counts track the number of documents that cite two works together. Bibliographic coupling frequencies measure the number of cited references that two works have cited together. The overall number of times a piece of scientific literature is cited as well as its relationship relative to other cited scientific literature can be used to determine that literature's overall impact within the scientific community.

Yet, this work focuses on the conceptual structure of the EDM and LA domains and therefore applies methods using keyword co-occurrences among the bibliographic collections. Through dimensionality reduction techniques such as Multidimensional Scaling (MDS), Correspondence Analysis (CA), or Multiple Correspondence Analysis (MCA), interactions within and across each topic unveil clusters of items that express common concepts (or thematic areas). This is accomplished by co-word analysis, where themes are determined by keywords and converted into clusters of keywords (or sublists). The definition of 'keyword' varies based on the needs of the researcher but can be limited to title of items, author keywords explicitly identified, abstracts, or combinations of the three determined by the database used. This work adapts the understanding that keywords and encompass all three fields.

Results from co-word analysis can be plotted on a two-dimensional map called a thematic map in order to visualize relationships among clusters. The conceptual structure depicted on such visualizations can show topics covered by researchers, relative similarity to other works, relative importance to the field, and the evolution of topics over a given period. Similar insights can be drawn from network diagrams.

Both thematic networks and citation graphs are natural visual representations of the respective networks. Network multiplication can be used to derive various network types [3]. Thematic networks are depicted by two-mode networks that represent links between a set of keywords and a set of corresponding works. Two-mode networks of this type can be represented with a matrix WK and can be computed through matrix multiplication, such as $WK^T * WK$. Other two-mode networks can be constructed, such as WA (works by authors), WJ (works by journals), and WC (works by classification) in a similar manner. Additional unique network types can be derived through matrix multiplication, such as $WJ^T * WA$ which gives the two-mode network AJ of authors by journals.

3. METHODOLOGY

There are basic steps that have been generalized for conducting bibliometric studies [4]. Figure 1 captures the five-step approach adapted in this study.

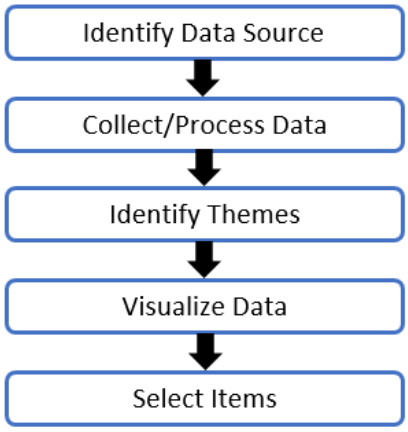

Figure 1. Process Flow for Bibliometric Approach

3.1 Data Source

All bibliographic information analyzed in this article was collected from the largest peer-reviewed citation database, Scopus. This abstract and citation database draws from three main sources: scientific journals, books and conference proceedings and integrates a patent database within the platform. Scopus is maintained by Elsevier and integrates titles from more than 5,000 publishers, with more than 20,000 serial titles, 150,000 books, and more than 70 million items. It was chosen as the most comprehensive data collection source available for the purpose of this research.

3.2 Data Collection

During the week of March 31 to April 6, 2019, three searches were conducted in the Scopus database. One search consisted of the term "learning analytics." The retrieved items were considered members of the LA dataset. A second search consisted of the terms "educational data mining." The retrieved items were considered members of the EDM dataset. Finally, a third search consisted of the two previous search terms combined with the Boolean operator AND. Items from this final group were considered members of the joint LA & EDM dataset. Table 1 gives summary statistics for the initial results of each search.

Table 1. Initial bibliometric results

Variable	Description	LA	EDM	LA & EDM
Items	# of unique documents	3008	1351	295
Sources (Journals, Books, etc.)	# of unique sources among documents	736	600	151
Keywords Plus (ID)	Keywords taken form titles, authors and abstracts by Scopus	6899	3892	1128
Author's Keywords (DE)	Keywords explicitly identified by authors	4885	2453	695

Period	Time period between earliest document and most recent document	2010 - 2019	2003 - 2020	2011 - 2019
Average Citations per Items	Average #of citations	5.97	8.544	12.07
Authors	Total # of *unique* authors across all documents	5373	3048	831
Author Appearances	Total # of authors	10200	4408	1063
Authors of Single-Authored Items	# of authors publishing alone	234	98	22
Authors of Multi-Authored Items	# of authors publishing collaborative works	5139	2950	809
Single-Authored Items	# of documents with a single author	404	155	28
Items per Author	# of documents per author	0.56	0.443	0.355
Authors per Item	Ratio of total # of documents and total # of authors	1.79	2.26	2.82
Co-Authors per Item	Average # of co-authors per document	3.39	3.26	3.6
Collaboration Index	Total authors of multi-authored articles/total multi-authored articles	1.97	2.47	3.03

All data files were downloaded as BibTeX bibliography files inclusive of all available Scopus data related to each item. Reference entries were stored in a style-independent, text-based file format [5] similar to the example entry in Figure 2.

```
@Book{todeschini+baccini
  author = "Robert {Todeschini} and Alberto {Baccini}",
  title = "Handbook of bibliometric indicators :
  quantitative tools for studying and evaluating research",
  publisher = "Wiley-VCH Verlag",
  year = 2016,
  address = "Weinheim, Germany",
  edition = "First"
 }
```

Figure 2. BibTeX Format

In each data set, entries with missing authors were removed. These entries generally correspond to conference proceeding papers that summarized collections of papers presented at that associated conference. Also, documents that were within the LA and joint LA & EDM datasets, as well as the EDM and joint LA & EDM datasets were removed to ensure that all three datasets were mutually exclusive. Documents that were missing keyword terms were also eliminated, as two-mode networks generated from works by keywords were to be used as the primary tool for generating thematic maps [3]. Search terms such as 'learning analytics' and 'educational data mining' were removed from each respective dataset to limit the chance of formulating dominant clusters that were centered around search terms.

Following the processing of all three data sets, 1,952 documents remain in the LA dataset (-35%), 783 observations in the EDM dataset (-40%), and 226 observations in the joint EDM and LA datasets (-20%). Table 2 gives summary statistics for the initial results of each search.

Table 2. Bibliometric results for included items

Variable	LA	EDM	LA & EDM
Items	1952	783	226
Sources (Journals, Books, etc.)	438	389	105
Keywords Plus (ID)	6436	3465	1180
Author's Keywords (DE)	3645	1592	574
Period	2010-2019	2003-2019	2011-2019
Average citations per items	5.057	8.558	7.947
Authors	3962	1942	719
Author Appearances	7068	2604	878
Authors of single-authored items	122	52	12
Authors of multi-authored items	3840	1890	707
Single-authored items	136	57	15
Items per Author	0.403	0.493	0.314
Authors per item	2.48	2.03	3.18
Co-Authors per Item	3.33	3.62	3.88
Collaboration Index	2.6	2.11	3.35

3.3 Theme Identification

The primary focus of the bibliometric analysis conducted within this research was to identify the key research themes within the fields of LA and EDM. Following preparation of the data sets, the R package, Bibliometrix, was used to further process raw BibTex files, perform co-word analysis and generate thematic maps [6]. Key areas of research focus within each respective dataset were identified as clusters formed from keywords extracted from each dataset. Clusters are identified by co-word analysis where keywords that occur frequently together within a research domain are grouped together. Clustering also shows subgroups of keywords that are linked to each other and the degree of those relationships. Therefore, the final clusters selected as representative samples for the three datasets were those clusters that demonstrate a higher relative degree of density and centrality when compared to other clusters. Centrality represents a cluster's relative interaction with other clusters, while density represents the relative interaction of members within a cluster.

Parameters were selected to maximize the amount of clusters generated from keywords within each of the three datasets. The relative frequency of the occurrence of a keyword within a cluster was set to three for all three domains: LA, EDM, and the combined LA & EDM dataset. The number of keywords were varied from 0 to the maximum amount of keywords in a collection's dataset by increments of 50. This process enabled optimization of the selected parameters in order to maximize the number of keyword clusters in each collection.

3.4 Theme Visualization

Resulting themes were then mapped to a visualization demonstrated by Cobo et al. [4], namely the thematic map. Clusters enable visualization relative to one another based on density and centrality in regions of the diagram in Figure 3. Once plotted, the largest clusters located in Quadrant I (labeled clockwise), or the Highly Developed, Motor Themes quadrant, were selected as the representative themes of the subject area domain. These themes have a high density and high centrality, and are the fundamental themes of the field.

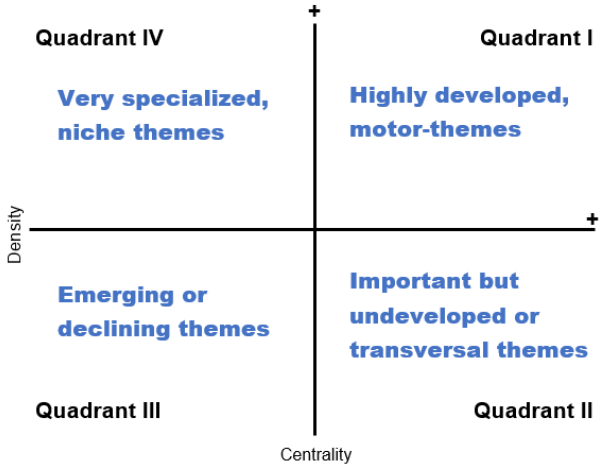

Figure 3. Thematic Map Quadrants

3.5 Item Selection

Following the selection of the clusters as the major themes for this collection of documents within each domain, a "bag of words model" was used to help select items that should be considered as most influential for the field [7]. A universal vector was created using all possible keywords in the collection of articles for each collection dataset. Subsequently, each thematic cluster and each document was converted to a vector by completing a one-hot encoding of the keywords within a cluster or the keywords associated with a document. Each document was compared to the thematic cluster, and a cosine similarity score was generated to assess how similar the documents were.

To select the representative sample of items within a thematic cluster, a combination of cosine similarity scores and total citations per item was utilized. Items with the highest cosine similarity score were selected within each cluster and represent those in the "core zone" described by Bradford's law [8]. Bradford's law serves as a good rule of thumb for describing the exponentially diminishing returns of retrieving articles for a given domain within a database. The three zones of Bradford's law are show in Figure 4.

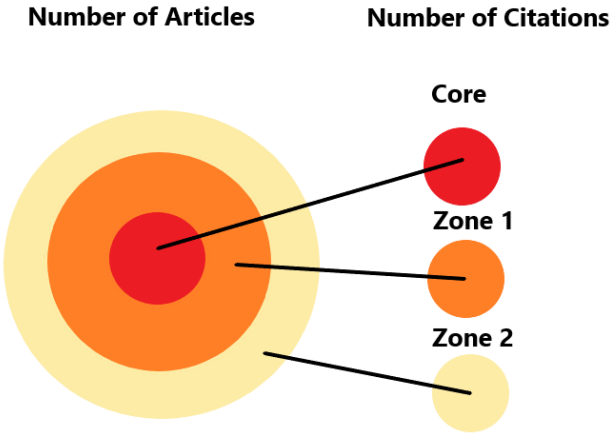

Figure 4. Bradford's Law Zones

Each item within the core zone were then ranked according to total citations. The top ten articles based on total citations were selected as the representative sample for each respective keyword cluster.

4. RESULTS

Six representative keyword clusters were extracted from the three datasets, two from each dataset. Words that represent each cluster for the selected keyword clusters are indicated in Table 3 – 5 and grouped according to one of four categories. 'Analysis/Tools' captures keywords related to the detailed examination or processing of inputs or the application of specific devices, software, or methods to perform particular analysis functions. 'Context (Environment)' features keywords related to the conditions that influence the setting where work is performed while 'Context (Target Group)' encompasses keywords related to specific individuals, groups, or levels. The final category, 'Teaching/Learning' features keywords related to the act of teaching (or providing instruction to learners) or learning (acquiring knowledge and skills by studying, experiencing, or being taught)

There are several keywords present across multiple datasets. For example, the keyword 'education computing' is shared between the LA dataset and the combined LA & EDM dataset. The keyword 'student performance' is shared between the dataset of EDM articles and the combined LA & EDM dataset.

Also noteworthy, words within the two LA keyword clusters focus primarily on instruction and communication. Within the first cluster, keywords such as "curricula", "mobile learning", "ontology", and "conceptual frameworks" dominate. However, the second cluster of LA, focuses on words such as "natural language processing", "computational linguistics", and "linguistics". Other notable terms within this cluster include "students' behaviors" and "knowledge building."

Within the EDM word cluster, the focus appears to be more on student performance and technical details for methods of predicting performance. Within the first cluster, keywords such as "recommender systems", "cognitive tutors", and "factorization," appear. Within the second cluster, reinforcement of many of these same themes in keywords such as "association rules", "supervised learning", and "decision support systems" are present.

Table 3. EDM Keyword Cluster Categories

Category	Keywords
Analysis/Tools	algorithms, association rules, classifiers, data sets, factorization, information management, matrix algebra, matrix factorizations, recommender systems, supervised learning
Context (Environment)	cognitive tutors, decision support systems, knowledge management, learning systems
Context (Target Group)	student models, students, students' performance, student's performance, university students
Teaching/Learning	N/A

Table 4. LA Keyword Cluster Categories

Category	Keywords
Analysis/Tools	computational linguistics, natural language processing, natural language processing systems, statistics
Context (Environment)	Computer-aided instruction, information systems, mobile applications, mobile learning, ubiquitous learning
Context (Target Group)	students' behaviors
Teaching/Learning	assessment, conceptual frameworks, curricula, design, education computing, information science, knowledge building, learning dispositions, linguistics, ontology

Table 5. LA & EDM Keyword Cluster Categories

Category	Keywords
Analysis/Tools	codes (symbols), forecasting, linear regression, performance prediction, predicting modeling, process mining
Context (Environment)	computer-based assessment, educational learning environment, e-learning, learning management system
Context (Target Group)	personalizations, student performance
Teaching/Learning	education computing

Keywords were further divided into four categories: analysis/tools, context (environment), context (target group), and teaching/learning. The category context is those keywords that are related to specific individuals, groups, or levels. Context (environment) is keywords related to the condition that influences the setting where learning or academic work is performed. The category teaching/learning are those keywords that are related to the act of teaching (e.g. providing instruction to learners) or learning (e.g. acquiring knowledge and skills by studying, experiencing or being taught). And finally, the category analysis/tools is keywords related to the detailed examination or processing of inputs or the application of specific devices, software, or methods to perform particular analytic functions. The categories context (target group) and analysis/tools are more closely aligned with the definition of Educational Data Mining while the categories context (environment) and teaching/learning are more closely related to Learning Analytics. In Table 3, 14 of the 19 keywords extracted from the EDM dataset are associated with EDM categories. In table 4, 11 of 20 keywords extracted from the LA dataset associated with LA categories. In Table 4, 10 of 13 keywords extracted from the combined LA & EDM dataset are associated with EDM categories. Although the keywords with the EDM and combined LA & EDM datasets differ, it is clear that when viewed from a framework of larger encompassing categories that the EDM and LA & EDM datasets share more similarities than differences.

Ultimately, 43 items were identified in the core zone for each cluster. After being ranked based on its influence (or total citations), the top item from each cluster was identified and captured in Table 6.

Table 6. Top item from each cluster

Item	Cluster	# Citations
Temporal predication of dropouts in MOOCS: Reaching the low hanging fruit through stacking generalization	LA 1	48
Learning dashboards: An overview and future research opportunities	LA 2	111
Recommender system for predicting student performance	EDM 1	80
A review on predicting student's performance using data mining techniques	EDM 2	84
Participation-based student final	LA &	43

performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory	EDM 1	
MOOCS: So many learners, so much potential	LA & EDM 2	102

5. CONCLUSION

Following keyword cluster extraction, it becomes clear that major research themes within LA focus primarily on student-focused learning objectives. Keywords such as "curricula", "student's behaviors", and "knowledge building" suggest a focus on using technology to help students learn and understand how they learn. There is also emphasis on words within the natural language learning domain such as "linguistics", "computational linguistics", and "natural language processing" that suggest an emphasis on bridging the gap between human and computer interaction through natural language processing.

The keyword clusters for EDM suggest a focus on the algorithms behind predicting student performance. Keywords such as "student performance", "cognitive tutors", "student models", and "learning algorithms" are aligned with this focus. Other keywords such as "recommender systems" and "performance classifiers" further suggest EDM's focus on predicting student preferences based on their performance.

Within the joint LA and EDM, the keywords such as "regression analysis", "linear regression", and "predictive modeling" highlight common algorithms within both domains. This focus on algorithms within the intersection of LA & EDM is similar to the focus and direction of major themes in EDM. With this consideration, it appears that EDM can be considered a subset of LA. What appeared to be two domains with a significant amount of overlap can be better described as one domain (i.e. LA) with one prominent subset (i.e. EDM). Figure 5 depicts both relationships.

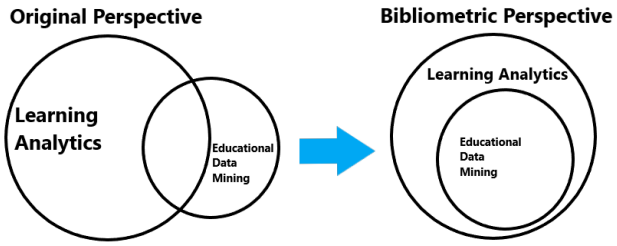

Figure 5. LA and EDM Venn diagrams

6. FUTURE WORK

Managing bias is a concern when conducting literature reviews, yet the application of a bibliometric approach helps to mitigate such risks. Future work could include literature reviews based on the items chosen by this bibliometric approach or adaption of such approach to alternative topics. Further information about the distinction between LA and EDM based on research production can be gleaned from the preliminary findings herein. Also, while the technique outlined in this paper focuses on identifying key concepts within Quadrant I of the thematic map, it could be modified to explore other quadrants that focus on emerging themes or isolated themes, i.e. Quadrants III and IV respectively. The identification of promising research areas can be beneficial for active researchers. Furthermore, exploration of the evolution of keyword themes over time is possible by dividing the same data into different consecutive groups of years for analysis over time. Finally, more advanced natural language techniques, such as word2vec models could be utilize to generate more robust results overall.

7. REFERENCES

[1] L. Bornmann, "Do altmetrics point to the broader impact of research? An overview of benef...: .," *J. Informetr.*, vol. 8, no. 4, pp. 1–24, 2014.Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[2] C. Hurter, "Analysis and Visualization of Citation Networks," Synth. Lect. Vis., vol. 3, no. 2, pp. 1–127, 2015.

[3] V. Batagelj and M. Cerinšek, "On bibliographic networks," *Scientometrics*, vol. 96, no. 3, pp. 845–864, 2013.

[4] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field," *J. Informetr.*, vol. 5, no. 1, pp. 146–166, 2011.

[5] J. Fenn, "Managing Citations and Your Bibliography with \bibtex," Pr. J., vol. 1, no. 4, pp. 1–19, 2006.

[6] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," *J. Informetr.*, vol. 11, no. 4, pp. 959–975, 2017.

[7] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 794–804, 2018.

[8] N. Desai, L. Veras, and A. Gosain, "Using Bradford ' s law of scattering to identify the core journals of pediatric surgery," *J. Surg. Res.*, vol. 229, pp. 90–95, 2019.