VJE Tạp chí Giáo dục (2024), 24(15), 7-11 ISSN: 2354-0753

TỔNG QUAN PHÂN TÍCH VÀ KHAI PHÁ DỮ LIỆU TRONG GIÁO DỤC ĐẠI HỌC TIẾP CẬN THEO PHƯƠNG PHÁP TRẮC LƯỢNG THƯ MỤC TỪ CƠ SỞ DỮ LIỆU SCOPUS GIAI ĐOẠN 2004-2023

Đặng Ngọc Hùng[+], **Bùi Thanh Huyền**

Trường Đại học Công nghiệp Hà Nội

+Tác giả liên hệ ● Email: dangngochung@haui.edu.vn

Article history
Received: 03/4/2024
Accepted: 27/5/2024
Published: 05/8/2024

Keywords
Teaching, decision making, university, bibliometric analysis

ABSTRACT
The growth of analytical research and data mining in higher education have recently highlighted the importance of systematic synthesis. This article provides a bibliometric review of 4023 published articles from the Scopus database in the period from 2004 to 2023. The study focuses mainly on using analytical techniques and Data mining in advanced higher education to optimize models for predicting learner-learning outcomes and detecting learner behavior for timely intervention. Furthermore, the study shows that research on data analytics and mining in higher education came from researchers from many different countries. Most studies are the result of collaborations between multiple authors, and most authors collaborate with authors from the same country. The countries with the most publications on data analytics and mining in higher education are the United States, China, and others. The study demonstrates data mining and analytics techniques in higher education institutions to create new methods to improve educational outcomes, especially in building predictive models to develop student achievements.

1. Mở đầu

Ngày nay, công nghệ đã thay đổi căn bản cách thu thập thông tin, dữ liệu, phân tích, diễn giải để định hình thực tiễn, quy trình và ra quyết định trong nhiều lĩnh vực khác nhau, giáo dục cũng không ngoại lệ (Kar & Dwivedi, 2020; Baek & Doleck, 2020). Thật vậy, dữ liệu được ghi lại bằng công nghệ trong giáo dục là chủ đề nhận được nhiều sự quan tâm trong những năm gần đây bởi tiềm năng thúc đẩy đổi mới giáo dục và khiến một số phương pháp truyền thống trở nên lỗi thời (Chen et al., 2020). Do tầm quan trọng ngày càng tăng của phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học, việc tổng hợp và phân tích các nghiên cứu trước đây là rất cần thiết. Khả năng tiếp cận dữ liệu thư mục và phần mềm thân thiện với người dùng đã dẫn đến sự gia tăng số lượng nghiên cứu trắc lượng thư mục trong nhiều lĩnh vực khác nhau (Ellegaard & Wallin, 2015). Ở Việt Nam, một số nghiên cứu cũng đã sử dụng phương pháp trắc lượng thư mục để tổng quan các nghiên cứu liên quan đến lĩnh vực giáo dục (Đỗ Hồng Liên và cộng sự, 2022; Đinh Thanh Tuyến & Phùng Thị Thu Nghĩa, 2024). Tuy nhiên, chủ đề nghiên cứu phân tích và khai phá dữ liệu trong giáo dục đại học vẫn chưa được xem xét đến. Bài báo này trình bày tổng quan các nghiên cứu về phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học bằng số liệu thống kê và phân tích theo chủ đề.

2. Kết quả nghiên cứu

2.1. Phương pháp và dữ liệu nghiên cứu

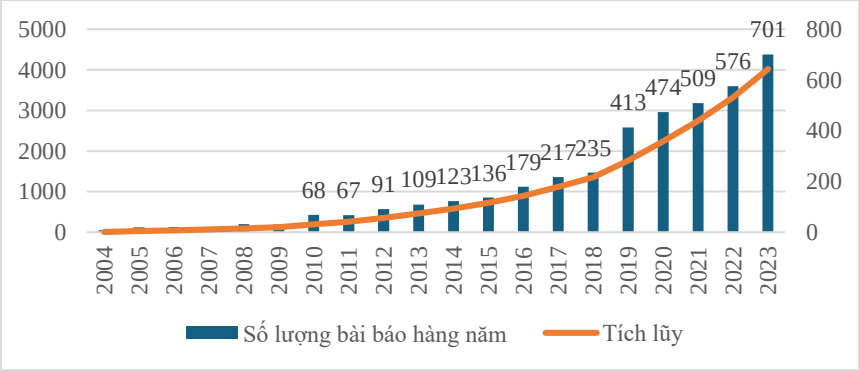
Trong khuôn khổ bài báo, các tác giả sử dụng kĩ thuật phân tích số lượng trích dẫn, quan hệ đồng từ (co-word) và phân tích đồng xuất hiện (co-occurence) của các từ khoá nghiên cứu nhằm thống kê số lượng ấn phẩm theo thời gian; số lượng xuất bản, trích dẫn của các quốc gia và mạng lưới cộng tác; các tạp chí nổi bật; đồng tác giả nổi bật và chỉ ra sự phát triển, bản đồ chủ đề nghiên cứu có liên quan đến chủ đề phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học. Quy trình làm sạch danh sách các công trình khoa học về chủ đề này gồm 4 bước: (1) *Xác định*: Chúng tôi sử dụng từ khoá như sau, cho kết quả 13.357 công trình, chứa ít nhất một cụm từ tìm kiếm (truy xuất từ cơ sở dữ liệu ngày 02/4/2024): "data mining" OR "data analysis " OR "data Analytics") AND ("Higher education" OR "University" OR "Universities") AND ("Student" OR "Students"); (2) *Lọc*: Chúng tôi tiến hành công tác sàng lọc các dữ liệu liên quan đến vấn đề lỗi, thiếu (trong đó có các tài liệu thiếu tóm tắt), về ngôn ngữ xuất bản (giới hạn ở tiếng Anh), khung thời gian xuất bản (2004-2023) ("data mining" OR "data analysis " OR "data Analytics") AND ("Higher education" OR "University" OR "Universities") AND ("Student" OR "Students" và phạm vi tài liệu (các

nước đang phát triển). Ở bước này, số lượng sau khi lọc chỉ còn 286 tài liệu; (3) *Kiểm tra tính hợp lệ*: Một trong những lí do loại các bài báo không phù hợp ở giai đoạn này là ở từng phần của bài báo đều có chứa từ khoá nhưng nội dung lại về lĩnh vực không liên quan; (4) *Tổng hợp*: Chúng tôi tổng hợp được danh sách gồm 4023 công trình khoa học phù hợp, bao gồm các bài báo và sách/chương sách để phục vụ giai đoạn phân tích trắc lượng thư mục theo mục tiêu của nghiên cứu.

Mục tiêu của việc sử dụng phương pháp nghiên cứu trắc lượng là để theo dõi sự phát triển về mặt nhận thức và lí thuyết của lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học. Việc tìm kiếm nâng cao các thuật ngữ liên quan đến và khai phá dữ liệu trong cơ sở giáo dục đại học đã được tiến hành trong cơ sở dữ liệu để xác định các ấn phẩm trên các tạp chí được lập chỉ mục trong danh mục Scopus. Bằng cách này, 4023 bài báo được xuất bản từ 2004 đến 2023 đã được tổng hợp để phân tích và nghiên cứu, thực hiện bằng chương trình RStudio và các câu hỏi nghiên cứu sau đây đã được giải quyết trong công trình này: (1) Khối lượng xuất bản hiện tại và xu hướng liên quan đến lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học là gì?; (2) Có thể thu được những hiểu biết gì từ xu hướng lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học?; (3) Những quốc gia nào có ấn phẩm phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học có ảnh hưởng nhất?; (4) Tạp chí nào hàng đầu đăng bài về phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học?; (5) Những chủ đề nào trong phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học có tiềm năng nghiên cứu trong tương lai?

2.2. Kết quả và thảo luận

2.2.1. Số lượng xuất bản theo thời gian



Hình 1. Số lượng nghiên cứu đã xuất bản phân tích và khai phá dữ liệu trong giáo dục đại học giai đoạn 2004-2023

Hình 1 cho thấy, lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học có tốc độ phát triển nhanh chóng, năm 2023 có số lượng xuất bản gấp hơn 10 lần so với năm 2011. Điều này phản ánh phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học với tư cách là một lĩnh vực nghiên cứu đang trở nên phổ biến trong giới học thuật và vẫn đang trong giai đoạn đang phát triển.

2.2.2. Số lượng xuất bản, trích dẫn của các quốc gia và mạng lưới cộng tác

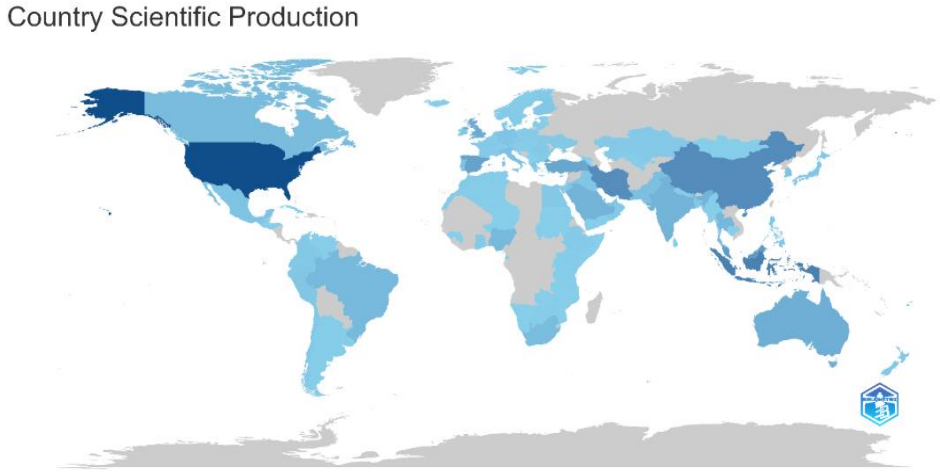
Khám phá các xu hướng nghiên cứu ở các khu vực địa lí khác nhau có thể giúp chúng ta hiểu bản chất và phạm vi của vấn đề nghiên cứu, từ đó đưa ra các giải pháp tiềm năng. Xét về chủ đề hiện tại, phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học là trọng tâm của các nước đang và phát triển, trong khi các nước kém phát triển ở châu Phi và châu Mỹ Latinh chưa thật sự chú trọng. Tuy nhiên, việc khám phá mạng lưới hợp tác của các quốc gia rất hữu ích trong việc xác định niềm tin chung, xây dựng chính sách và liên kết các dự án song phương và đa phương cũng như xác định quốc gia nào tham gia trao đổi kiến thức. Hoa Kỳ đứng đầu danh sách các quốc gia có số lượng bài viết về chủ đề này với tổng số bài báo là (1739), tiếp theo là Indonesia (1015), Trung Quốc (855). Ở châu Á, Trung Quốc, Indonesia, Iran và Malaysia đã tham gia vào lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học, trong khi Vương quốc Anh và Tây Ban Nha là những nước dẫn đầu về nghiên cứu về chủ đề này ở châu Âu (bảng 1).

Bảng 1. Tổng hợp 10 quốc gia có số bài báo và số lượng trích dẫn nhiều nhất

STT	Quốc gia	Số bài báo	Quốc gia	Tổng số trích dẫn	Trích dẫn trung bình trên bài báo
1	Hoa Kỳ	1739	Hoa Kỳ	6408	15,3
2	Indonesia	1015	Trung Quốc	3569	13,1
3	Trung Quốc	855	Tây Ban Nha	2738	21,1
4	Iran	829	Vương Quốc Anh	2058	15,5

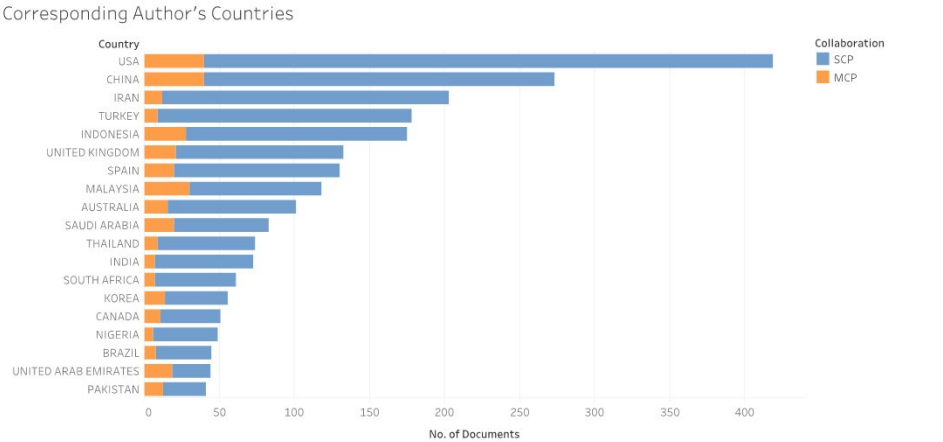
5	Malaysia	642	Thổ Nhĩ Kỳ	2048	11,5
6	Tây Ban Nha	543	Úc	1935	19,2
7	Thổ Nhĩ Kỳ	497	Saudi Arabia	1272	15,3
8	Vương Quốc Anh	440	Lebanon	1222	203,7
9	Australia	398	Iran	1220	6
10	Saudi Arabia	328	Ấn Độ	1119	15,3

Về trích dẫn, Hoa Kỳ tiếp tục là quốc gia đứng đầu trong tổng xếp hạng trích dẫn (bảng 1), hầu hết các quốc gia được trích dẫn đều tham gia vào mạng lưới và hợp tác quan trọng với nhau. Trung Quốc có số lượng bài báo xuất bản đứng thứ 3, nhưng tổng số trích dẫn lại xếp thứ hai, trong khi đó Indonesia mặc dù có số lượng công bố đứng thứ 2, nhưng lại không nằm trong 10 quốc gia có số lượng trích dẫn nhiều nhất, điều này phản ánh chất lượng của các bài báo của các tác giả Indonesia. Hình 2 cho thấy, nhiều khu vực, chẳng hạn như Trung Á và châu Phi, vẫn chưa tham gia vào các "cuộc tranh luận" khoa học về chủ đề phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học. Theo đó, chúng tôi cho rằng các nhà khoa học nên thực hiện nghiên cứu ở nhiều quốc gia, so sánh, phân tích các ngành, lĩnh vực khác và áp dụng cách tiếp cận quốc tế hơn (Màu càng đậm thể hiện các quốc gia có mạng lưới cộng tác càng lớn).



Hình 2. Mạng lưới cộng tác giữa các nước

Hình 3 cho thấy, 20 quốc gia có nhiều bài báo được xuất bản nhiều nhất và mô hình hợp tác giữa các quốc gia. 10 quốc gia hàng đầu bao gồm các quốc gia từ nhiều châu lục khác nhau như Trung Đông, Bắc Mỹ, châu Á, và châu Âu. Các tác giả từ Hoa Kỳ có nhiều ấn phẩm hơn đáng kể so với các quốc gia khác, quốc gia có nhiều ấn phẩm thứ hai (tức là Trung Quốc và Iran). Hình 3 cũng mô tả sự hợp tác quốc tế với MCP (ấn phẩm của nhiều quốc gia) cho biết số lượng bài báo có ít nhất một đồng tác giả từ một quốc gia khác; SCP (ấn phẩm của một quốc gia) cho biết số lượng bài báo có tất cả các tác giả đến từ cùng một quốc gia.



Hình 3. Cộng tác của các tác giả giữa các nước

2.2.3. Các tạp chí cốt lõi về phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học

Một nhà nghiên cứu phải làm quen với các tạp chí lớn trong lĩnh vực họ quan tâm để có được tài liệu có chất lượng cao và có tác động, sau đó, lập danh sách các tạp chí mục tiêu để công bố những kết quả nghiên cứu của họ. Do đó, các tạp chí hàng đầu xuất bản về chủ đề phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học đã được kiểm tra. 4023 bài viết trong cơ sở dữ liệu Scopus được trải rộng trên 1251 tạp chí thuộc nhiều lĩnh vực như Giáo dục, Máy tính và Xã hội học,... Trong số các tạp chí dẫn đầu về số lượng bài báo, tạp chí "Sustainability" xuất bản đứng đầu danh sách với 89 bài nghiên cứu. Sự tồn tại của một tạp chí cốt lõi như "Sustainability" và vị trí hàng đầu của nó cho thấy tầm quan trọng và sự đa ngành của lĩnh vực nghiên cứu này.

2.2.4. Phân tích đồng tác giả

Phần này xác định các tác giả tích cực nhất trong lĩnh vực này xét về số lượng ấn phẩm và tác động của họ được đo bằng số trích dẫn nhận được từ các bài báo khác. Đứng đầu tiên trong danh sách các tác giả có ảnh hưởng nhiều nhất là bài viết của Samaha và Hawi (2016) với chủ đề về mối quan hệ giữa nghiện điện thoại thông minh, căng thẳng, kết quả học tập và sự hài lòng với cuộc sống, đăng trên Tạp chí "Computers in Human Behavior", có số lượng trích dẫn nhiều nhất với 703, tỉ lệ trích dẫn hằng năm là 78,11. Việc nghiên cứu các học giả tích cực và có ảnh hưởng nhất trong lĩnh vực họ quan tâm có thể giúp các nhà nghiên cứu phát triển các dự án hợp tác mang lại nghiên cứu có tác động và chất lượng cao.

2.2.5. Sự phát triển và bản đồ chủ đề nghiên cứu

Khi nghiên cứu về phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học được thực hiện vào khoảng năm 2004 - 2016, các chủ đề liên quan đến là sử dụng các công cụ thống kê truyền thống để nghiên cứu về đặc tính của nhân khẩu học, giới tính, môi trường xã hội đến kết quả học tập của sinh viên. Tuy nhiên, từ năm 2017 cho đến nay, các nghiên cứu thường liên quan đến việc sử dụng công nghệ như máy học, khai phá dữ liệu để phân tích và đánh giá kết quả học tập, giảng dạy và hỗ trợ cho việc ra quyết định.

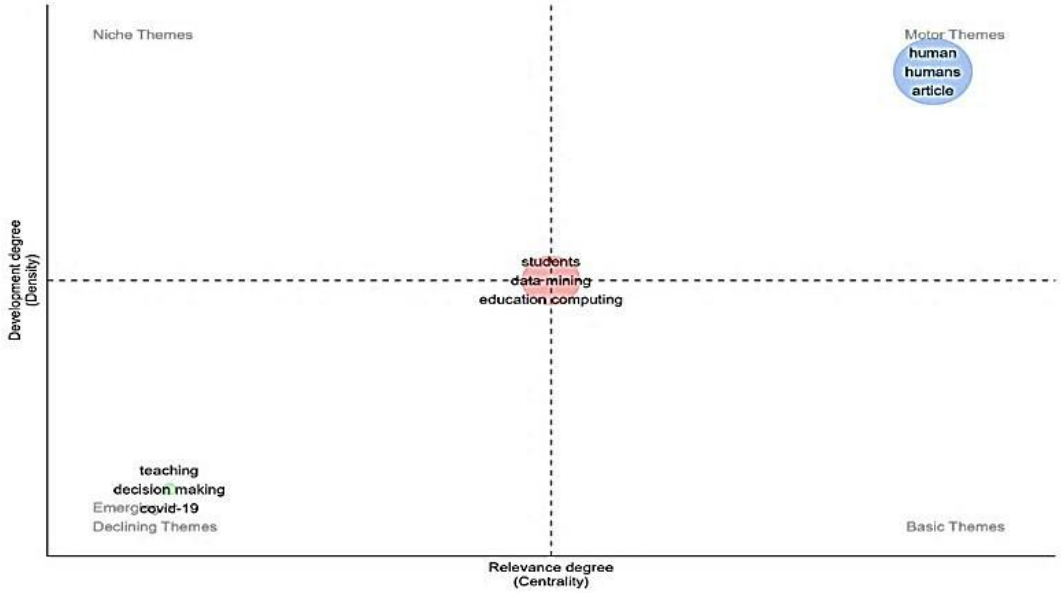
Bản đồ chuyên đề có khả năng diễn giải cao và đặt nền tảng cho việc phân tích bốn góc phần tư trong đó các chủ đề được xác định. Việc nhóm các chủ đề thành các góc phần tư giúp xác định chủ đề nào trong một lĩnh vực được phát triển hơn các chủ đề khác cũng như chủ đề nào đang nổi lên và cần thiết và do đó cần được chú ý nhiều hơn. Các chủ đề được lấy từ các cụm từ khóa và mối liên hệ giữa chúng. Hai biến số xác định tất cả các chủ đề nghiên cứu được tạo ra trong quá trình này là "mật độ" và "tập trung". Để phân loại các chủ đề, có thể sử dụng các phép đo xu hướng trung tâm như giá trị trung bình và giá trị trung bình cho mật độ và tính trung tâm. Các từ khóa và các mối liên kết của chúng tạo thành một biểu đồ "mạng chuyên đề" theo một chủ đề, với "trung tâm" trên trục hoành và "mật độ" trên trục tung. Nếu một nút có mật độ kết nối cao với các nút khác trong mạng thì nút đó có tính trung tâm cao hơn và quan trọng trong mạng. Nút có mật độ lớn hơn biểu thị cường độ liên kết nội tại mạnh hơn hoặc mạnh hơn giữa các nút. Mật độ của một lĩnh vực nghiên cứu cho thấy khả năng duy trì và phát triển của nó.

Trong hình 4 (trang bên), một bản đồ chuyên đề về lĩnh vực phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học, với bốn góc phần tư Q1 - Q4, được cung cấp; góc phần tư phía trên bên phải (Q1) đại diện cho các chủ đề vận động, góc phần tư phía dưới bên phải (Q4) đại diện cho các chủ đề cơ bản, góc phần tư phía trên bên trái (Q2) đại diện cho các chủ đề chuyên biệt/ngách và góc phần tư phía dưới bên trái (Q3) đại diện cho các chủ đề mới nổi/suy giảm. Các chủ đề trong Q1 đều đã chín muồi và rất quan trọng đối với cấu trúc của lĩnh vực nghiên cứu về con người trong giáo dục. Các chủ đề trong Q3 đang phát triển yếu và cận biên, chủ yếu cấu thành các chủ đề mới nổi hoặc đang suy thoái, chẳng hạn như dạy học trực tuyến, Covid-19 và hỗ trợ ra quyết định.

Phát hiện của chúng tôi cho thấy, số lượng ấn phẩm đã tăng lên những năm gần đây, đặc biệt là năm 2023, số lượng ấn phẩm nhiều tối thiểu gấp mười lần so với 20 năm trước. Xu hướng này phản ánh sự chú ý ngày càng tăng đối với nghiên cứu phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học. Nguyên nhân có thể là do các công nghệ và kĩ thuật mới nổi được phát triển cũng như các nhà nghiên cứu từ các ngành khác tham gia nghiên cứu liên quan đến phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học. Như vậy, ngày càng có nhiều bộ dữ liệu, công cụ miễn phí sẵn có và sự quan tâm đến các ứng dụng, môi trường giáo dục mới cũng tăng lên (Romero & Ventura, 2020).

Về nguồn mà các bài báo phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học trích dẫn hàm ý một số xu hướng nghiên cứu của lĩnh vực này, đầu tiên, phụ thuộc rất nhiều vào một nguồn cụ thể như giáo dục, công nghệ thông tin. Một xu hướng khác của cấu trúc nguồn hàm ý trọng tâm cốt lõi của nghiên cứu phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học là phát triển và cải tiến các kĩ thuật phân tích dữ liệu bằng cách sử dụng các công nghệ tiên tiến cũng như sự tham gia của nghiên cứu phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học vào

bối cảnh sư phạm và học tập. Ngoài ra, một số nguồn được trích dẫn thường xuyên tập trung vào các kĩ thuật khai thác dữ liệu (Ví dụ: Học máy) trong khi một số nguồn khác liên quan đến việc xuất bản nghiên cứu kết nối phương pháp sư phạm và lí thuyết với công nghệ (Ví dụ: Máy tính trong hành vi con người). Xu hướng nghiên cứu phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học trong việc kiểm tra các kĩ thuật đổi mới và kết nối phương pháp sư phạm với công nghệ cũng là một trong những nguồn được trích dẫn nhiều nhất.



Hình 4. Các chủ đề nghiên cứu

3. Kết luận

Kết quả nghiên cứu cho thấy xu hướng gần đây của lĩnh vực khai thác dữ liệu giáo dục; phản ánh một xu hướng tích cực trong lĩnh vực khai thác dữ liệu giáo dục, đặc biệt là trong cấp độ giáo dục đại học. Phân tích phản ánh rằng việc sử dụng các phương pháp phân tích và khai phá dữ liệu trong lĩnh vực này đang trở nên nghiên cứu ngày càng phổ biến. Mục tiêu của các nghiên cứu là tạo ra các phương pháp mới nhằm cải thiện hiệu suất giáo dục, đặc biệt là thông qua việc xây dựng các mô hình dự đoán để đánh giá thành tích của sinh viên, theo đó thể hiện sự quan tâm và nỗ lực đáng kể từ cộng đồng nghiên cứu trong việc áp dụng công nghệ thông tin để tối ưu hóa quá trình học tập và giảng dạy trong giáo dục đại học. Tuy nhiên, kết quả cũng đặt ra yêu cầu một cộng đồng các nhà nghiên cứu phân tích và khai phá dữ liệu trong cơ sở giáo dục đại học đa dạng và hòa nhập trước thực trạng sự hợp tác quốc tế còn tương đối hạn chế.

Tài liệu tham khảo

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278.

Đinh Thanh Tuyến, Phùng Thị Thu Nghĩa (2024). Nghiên cứu chủ đề đọc viết ở giai đoạn mầm non tại các nước đang phát triển trên cơ sở dữ liệu Scopus trong giai đoạn 1994-2021: Xu hướng và hợp tác quốc tế. *Tạp chí Giáo dục*, *24*(6), 1-5.

Đỗ Thị Hồng Liên, Nguyễn Lê Vân An, Nguyễn Tiến Trung (2022). Xu hướng nghiên cứu về chủ đề quốc tế hoá chương trình đào tạo: một nghiên cứu trắc lượng. *Tạp chí Giáo dục*, *22*(17), 1-7.

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*(3), 1809-1831. https://doi.org/10.1007/s11192-015-1645-z

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research - Moving away from the "What" towards the "Why." *International Journal of Information Management, 54*, 102205. https://doi.org/10.1016/j.ijinfomgt.2020.102205

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355

Samaha, M., & Hawi, N. S. (2016). Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in Human Behavior, 57*, 321-325. https://doi.org/10.1016/j.chb.2015.12.045