VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2023, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 11, Number 4, October-December 2023 *pp:124-137*

The Impact of Data Mining on Digital Libraries – A bibliometric Study

Sana Alam[1*], **Shehnila Zardari**[2], **Umme Laila**[1], **Muhammad Abbas**[1], **Noor Ul Huda**[1], **Muhammad Asghar Khan**[1]

[1]Department of Computer Science and Information Systems, Institute of Business Management (IoBM), Karachi, Pakistan; [2]Department of Software Engineering, NED University of Engineering and Technology, Main University Road, Karachi, Pakistan

Keywords: *mining, bibliometric, virtual library, digital library*

Journal Info:
Submitted:
October 15, 2023
Accepted:
December 16, 2023
Published:
December 31, 2023

Abstract Purpose: The study provides a comprehensive bibliometric assessment of Data Mining in Digital/ Virtual Libraries to depict the importance of Data Mining technology with respect to Digital/ Virtual Libraries. Methodology: Our research work includes 215 studies from the past 23 years that are analyzed on the basis of carefully articulated 10 research questions. The tools used for performing analysis regarding visualization aspects include VOSviewer and Bibliometrix (R studio). Findings: The evaluation shows that the year 2017 has the highest number of publications. Our work also represents that the use of Data Mining in Digital/ Virtual Libraries has influenced multiple domains, however, it is frequently used in Computer Science. An assessment of the top 20 countries suggests that USA and China are the major contributors in terms of published articles for a time period of 23 years. In the study, we use VOSviewer for co-word analysis to represent the relatedness of documents based on keywords. Our study further explores the research themes and topic dendrogram with respect to Data Mining in Digital/ Virtual Libraries by using Bibliometrix (R studio). Our research findings also show that although most research work is published in the English language yet there are few major studies in other languages also. Originality: The research provides insight into the above-mentioned aspects of bibliometrics, enabling researchers and scholars to make better decisions regarding their research.

***Correspondence author email address:** sana.alam@iobm.edu.pk
DOI: 10.21015/vtse.v11i4.1700

1 Introduction

Recent studies show that proficiency in generating and collecting data has rapidly increased. Applications, including commercial management, government administration, and scientific and engineering data management, have used millions of databases. Due to the massive growth in users and data in databases, there is an urgent need for novel methodologies and technologies that can quickly and intelligently convert processed data into learning and data. As a result,

data mining has experienced rapid development [1], [2], and [3].

Data mining, also known as knowledge discovery in databases, is the process of removing implicit, previously unknown, and potentially useful information (such as knowledge norms and standards, limitations and boundaries, and regularities) from database data [4]. Data mining includes mathematical and statistical methods that can be used with data from a variety of sources [5]. Data mining is utilized to extract meaningful information from massive volumes of datasets. It involves revealing significant hidden patterns in a piece of data, to put it another way. The benefit of data mining is that it actively seeks industry trends and offers beneficial results to businesses that keep a lot of information [6]. We can access, store, and analyze vast amounts of data and potentially connect real-time data to assist our decision-making by utilizing big data technologies[7]. With the ongoing growth of automatization, artificial intelligence, and digitalization in the university library, the volume of information, is increasing. The data collected from academic institutions have had significant elements of big data[8] which is evident from the use of information resources and the current development of information technology.

Nowadays, many university libraries are using various technologies to manage their different data forms. Their database systems are efficient for the implementation of regular queries as well as other traditional functions [9] [10]. However, they cannot predict future trends when handling large amounts of data in the library. As a result, it is necessary to investigate ways to give users more customized services so that they can make more trustworthy decisions by utilizing data mining processes to extract information [8][10] .

Bibliometric assessment is a method for presenting a large number of quantitative facts on the basis of various diverse aspects [11] [12]. Bibliometric evaluation plays a key role in helping scholars, researchers and students make better decisions in their respective disciplines based on numerous features and encourages scholars, researchers, and students to conduct research in emerging areas. Web of Science (WoS) database is one of the most renowned, well-established, and highly significant databases for carrying out a bibliometric study.

The main goals of our research study include: (i) providing a comprehensive bibliometric assessment for Data Mining on Digital/ Virtual Libraries in accordance with the carefully designed 10 research questions. (ii) enabling the research scholars to make a better decision regarding the institution/ organization and country for their research work. (iii) evaluating the variation in research publication trends in accordance with the frequency of published works over the last 23 years. (iv) representing the top 20 countries, and institutions in terms of published studies. (v) representing the fact that Data Mining on Digital/ Virtual Libraries solidifies its position whose impact encompasses multi-disciplinary areas. (vi) depicting the relatedness amongst documents in terms of co-words (vii) representing the themes and topics in accordance with the published works (viii) representing the top funding agencies.

The significance of this study is that it would be beneficial to researchers, educators, and practitioners from the data mining and digital library industry to make a better judgment regarding various aspects of research in this area. The varied research areas illustrate the significant impact across multidisciplinary fields. Hence, this aspect helps researchers, educators, and practitioners in attaining a broader view to incorporate the knowledge, methodology, techniques, and applications of data mining in multidisciplinary or cross-disciplinary areas. Moreover, educators also benefit from cross-disciplinary research and modify the course outline by implementing data mining techniques, applications, and approaches based on cross-disciplinary or multidisciplinary fields representing the undeniable impact across various fields. The research production rate of different institutions and countries helps researchers and educators to make more informed decisions regarding various collaboration aspects. The most active institutions/ organizations in terms of frequency of publications

also have a great impact on academia-industrial collaboration. It helps the industrialists to make more informed decisions regarding the investment and establishment of research labs in the corresponding institutions/ organizations. Multi-national companies while analyzing various business aspects can go through the list of most active countries in accordance with the published articles to make a better decision regarding the research potential of the specific country. Moreover, the number of co-word occurrences also represents the published articles' research trends. These research trends are one the important aspects for the researcher, educators, and practitioners to suggest and incorporate innovations in the standard procedures and approaches of data mining applied to digital and virtual libraries. This further opens new dimensions of research in well-established fields. The analysis of highly cited scholarly works enables naive researchers to enhance their writing skills, thus, projecting their suggestions in a more effective way.

Moreover, emerging research themes are also being identified that help the researchers. The continent-wise research contribution provides a global view of the research participation of various continents considering how often publications are made. The diversity of the languages of the published works encourages the researchers to contribute to the research by publishing articles in the language according to their fluency. In short, bibliometric assessment is beneficial in providing a comprehensive qualitative and quantitative analysis covering research aspects.

The remaining portion of the article is divided into a number of sections. Section 2 represents related work. The methodology is presented in Section 3. Section 4 covers the findings and discussion. Section 5 deals with the limitation and future work. Section 6 includes a conclusion and references are a part of Section 7.

2 Related Work

Table 1 provides an illustration of the research investigations based on the time periods, data sources, and parameters that were examined. Table 1 focuses on the bibliometrics works carried out over different time periods in the field of Data Mining. Table 2 highlights the related work in the field of the digital library. Table 3 lists our research contributions.

Table 1. Prominent Bibliometric studies in the field of Data Mining

Refs	Time Durations	Data Sources	Parameters Analyzed
[13]	2009-2018	WoS Scopus Science Direct	Publication rate of DM and ML papers in the field of public health, Systematic review, and Research topics
[14]	2011-2019	WoS	Latest advances, leading topics, current gaps, citation analysis, and research trends.
[15]	2014-2016	WoS	Survey of publications, Top Scholars, Top 10 articles, and organizations
[16]	2015-2019	WoS	Data analysis, Articles and Keyword analysis, and Percentage of Publications
[17]	1990-2016	Scopus	Author's analysis for scholarly publications for DM and Medical Imaging, countries collaboration, and Co-word Analysis
[18]	2001-2013	WoS	Technology forecasting, Citation analysis, network analysis, and Semantic patent analysis.
[19]	1990-2017	Pubmed, Embase, CINAHL, China Biology Medicine and Wang Fang Database	Trend changes in literature, Main areas of DM, and collaborative network analysis.

3 Methodology

3.1 STEPS OF METHODOLOGY

3.1.1 INFORMATION GATHERING AND ANALYSIS

1. Comprehensive Coverage

The WoS provides extensive coverage of scholarly literature, including peer-reviewed journals, conference proceedings, and research articles from various disciplines. Its comprehensive database ensures that researchers can access a wide range of high-quality sources relevant to their field of study [24].

2. Citation Indexing

One of the distinguishing features of the WoS is its citation indexing, which allows researchers to trace the influence and impact of a particular article over time. This feature is crucial for understanding the scholarly context and relevance of a given work [24].

3.Interdisciplinary Research

Researchers benefit from the interdisciplinary nature of the WoS, as it facilitates the exploration of connections and intersections between different fields of study [25]. This interdisciplinary approach enhances the depth and breadth of research projects

Table 2. Prominent Bibliometric studies in the field of Digital Library

Refs	Time durations	Data sources	Parameters Analyzed
[20]	2002 - 2016	WoS	Annual citation rate, the productivity of organizations and countries, citation analysis, and authors' productivity,
[[21]	1980 - 2017	WoS	Annual publication trend, citation trend, countries productivity, types of documents, research areas, analysis of keywords, top publication journals
[22]	2010 - 2019	[]WoS	[]Analysis of the documents published in Library Quarterly with respect to the trends in publications, authors' productivity, co-occurrence of words, highly cited articles, and collaboration amongst countries on the basis of bibliographic coupling
[23]	Till 2021	Scopus	Top contributing countries and organizations, occurrences of keywords, trends of publication, top-cited scholarly works, top contributing authors, trending topics, thematic analysis, and growth of the words.

4.Accurate and Reliable Data

The WoS is known for its commitment to providing accurate and reliable data, making it a trusted resource for researchers. Scholars can confidently rely on the information retrieved from this database for their academic work [24].

5.Time-saving Search Tools

The user-friendly interface and advanced search tools of the WoS contribute to efficient and effective literature searches. Researchers can save time and streamline their research process by utilizing these tools.

In conclusion, the use of the WoS is justified due to its comprehensive coverage, citation indexing, support for interdisciplinary research, provision of accurate data, and efficient search tools. Researchers can confidently rely on this database to access high-quality scholarly literature relevant to their research projects.

VOSviewer software is used to conduct various analysis and subsequently visualize the intellectual structure [26].

Table 3. Our Research Contributions

Our Research Contributions
Analysis based on the review timeline of 23 years (2000 – 2022)
Our work represents a comprehensive bibliometric assessment over a time duration of 23 years (2000 – 2022)
Analysis based on top20 institutions/ organizations and countries
Our research study lists down and analyzes the top 20 institutions/ organizations and countries in the field of Data Mining with respect to Digital/ Virtual Libraries
Analysis of multi-disciplinary research areas
The emergence of various diverse research areas re-affirms the fact that the impact of Digital Mining on Digital/ Virtual Libraries encompasses various multi-disciplinary areas besides Computer Science.
Language-based analysis of articles
To prevent any bias caused by the writing style, we have not only included research produced in other languages but also publications published in English.
Relationships between research works
The connectedness amongst research work is illustrated by the visualization software. The visual representation helps to interpret relationships amongst articles in terms of co-word much better.
Research Themes and Research Topics
The research themes and emerging research topics are presented by using the tool Bibliometrix (R studio) in the form of a thematic map and topic dendrogram.
Publication trend in terms of frequency of publications
Our study shows the variation in publication trends in terms of publications frequency over a time span of 23 years (2000 – 2022)

3.1.2 RESEARCH QUERY DESIGN AND IMPLEMENTATION

We have included the research query "("data mining") AND ("virtual library" OR "digital library" OR "e-library" OR "smart library" OR "library management" OR "library information" OR "library administration" OR "library authentication" OR "library science" OR "digital libraries" OR "smart libraries" OR "e-libraries" OR "virtual libraries")" which resulted in the total number of 227 records. After applying the Inclusion/ Exclusion criteria as shown in Table 4, a total of 215 records are finalized for analysis. All these 215 records are analyzed on the basis of 10 research questions as

presented in Table 5.

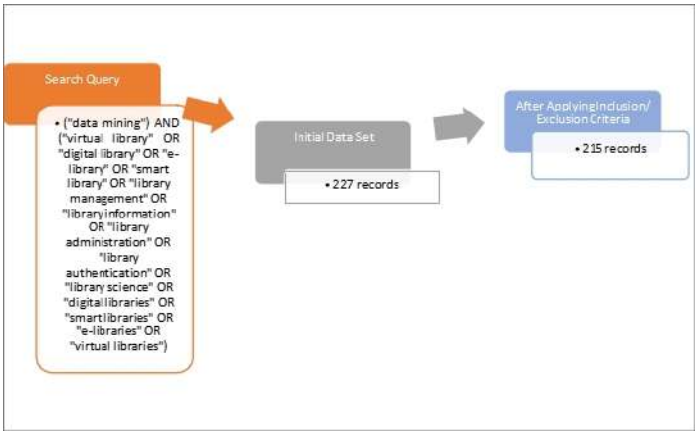

Figure 1. Methodological Process

Table 4. Dataset Inclusion and Exclusion Criteria

Inclusion/ Exclusion Criteria	Details of Criteria
Inclusion Criteria	6.Only the articles relevant to the topic are included. 7.Timespan for the dataset (2000 – 20 June 2022) 8.Following types of research publications are included. * Journal Articles * Conference Proceedings * Editorial Materials * Book Chapters * Review Articles
Exclusion Criteria	1-All the articles that don't fall into the relevant topic. 2-If the full text is not available for the article.

Table 5. Research Questions for our bibliometric study

Insights	Research Questions
Annual research publication	Q1. How many annual scholarly publications are there?
Types of Publication	Q2. Which type of documents are assessed?
Categories (WoS)	Q3. Which WoS categories are ranked among the top 20?
Research Areas	Q5. What is the impact of data mining on digital and virtual libraries in terms of research areas?
Organization Contribution	Q4. On the basis of publishing frequency, which organizations are included among the top 20?
Countries Contribution	Q6. Which nations are listed among the top 20 in terms of publications?
Languages Dialect	Q7. Using scientific publications that have been published, how are different languages ranked?
Continent-wise Research Contribution	Q8. What is the contribution of different continents in terms of frequency of publication?
Relation Amongst Documents	Q9. How do documents with similar words correlate with one another?
Emerging Research Topics and Research Themes	Q10. What are the hot subjects and developing research themes?
Funding Agencies	Q11. What are the top 20 funding agencies in the field of data mining in digital/ virtual libraries?

4 Research fundings and Discussion

This section provides a comprehensive description of the investigation.

4.1 YEAR-WISE SCIENTIFIC PRODUCTION

Annual research publication in Figure 2 displays the variation in publication trends in Data Mining for Digital/ Virtual Libraries over the past 23 years. The year 2017 shows that it has the highest publication count.

4.2 TYPES OF DOCUMENTS

Figure 3 shows the trend in the types of documents published. Analysis shows that a research article may be classified into more than one type of document. As an example, a research publication may b e categorized as a proceeding paper as well as an article. Proceeding's paper is the highest type of document these publications belong to.

4.3 TOP 20 WEB OF SCIENCE CATEGORY

Findings based on the top 20 WoS categories for the year 2000-2022 are depicted in Table 6 shows the WoS categories which are journal-based and mapped to one Research Area. The impact of Data Mining on Digital/ Virtual Library across various fields is apparent in the diverse WoS categories. The top three ranked categories are Computer Science Information

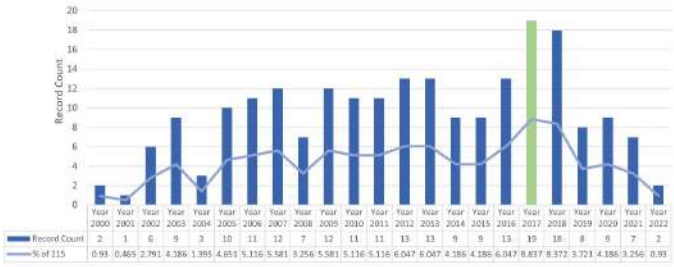

Figure 2. Publication trend in terms of the frequency of published articles over (2000 – 2022)

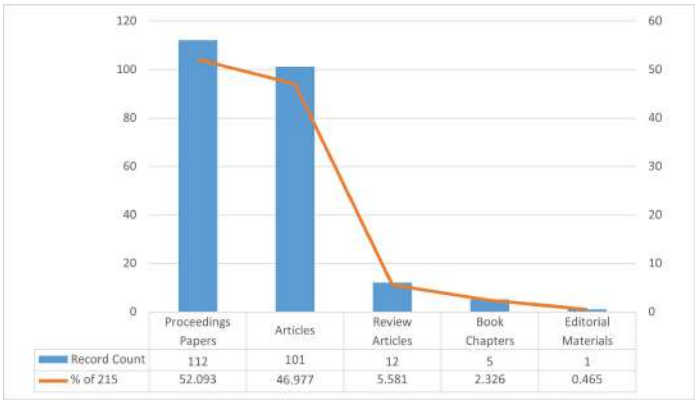

Figure 3. Publication count representing types of documents

Systems, Information Science Library Science, and Computer Science Theory Methods. The data mining of digital or virtual libraries done in these fields helped in extracting knowledge, finding patterns, and interacting with the statics of large data.

4.4 TOP 20 RESEARCH AREAS

Research areas are article-based. As a result, one can locate, access, and evaluate documents from numerous databases that fall under a given subject area. Table 7 represents the top 20 research areas obtained from the dataset (2000–2022). The analysis of research areas depicts that the influence and impact of Data Mining with respect to Digital/ Virtual libraries encompasses Business Economics, Engineering, Educational Research, Mathematics and Telecommunications and s not limited to only Computer Science or Information Science/ Library Science fields.

4.5 TOP 20 INSTITUTIONS

Institutional analysis provides a clear understanding regarding the contribution of different institutions with respect to the bibliometric assessment [27]. Table 8 depicts the ranking of the top 20 institutions with the most publication count for the 23-year timeframe (2000–2022). This criterion facilitates scholars to identify the highly contributing organizations/ institutions in the field of Data Mining for Digital/ Virtual Libraries. Figure 4 represents an analysis of the share of the countries to which the top 20 institutions/ organizations belong. The majority of the institutions are American and Chinese.

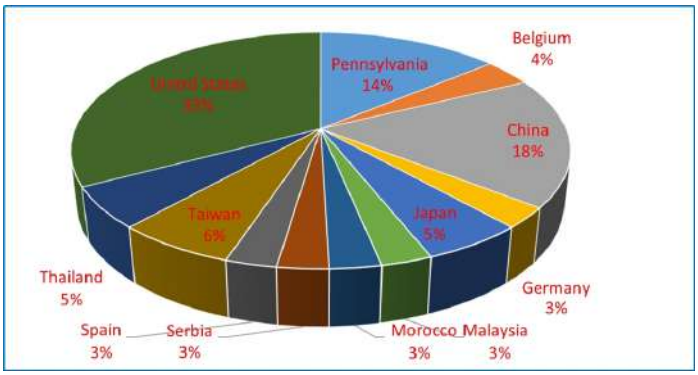

Figure 4. Percentage of the top 20 institutions/ organizations belonging to different countries

4.6 TOP 20 COUNTRIES

Countries analysis plays a pivotal role in bibliometric assessment [28] [29] [29][30]. Table 9 represents the 20 most contributing countries in the specified field in terms of publication count. USA and China lead the research contribution as these two countries are the most resourceful countries with highly advanced facilities for conducting research. However, many Asian and developing countries are also contributing effectively in terms of publication frequency.

4.7 LANGUAGES

This research criterion represents the use of various languages for publications in the field of Data Mining for Digital/ Virtual Libraries. Table 10 shows that English is the major language used for research publications. However, a smaller number of articles are published in Chinese and Spanish languages.

The articles published not just in English but in other languages proved to be a driving source for the spread of knowledge to non-English speakers. Figure 5 shows that almost 2% of articles are in languages other than English. This criterion also encourages the authors to document their research in their native languages.

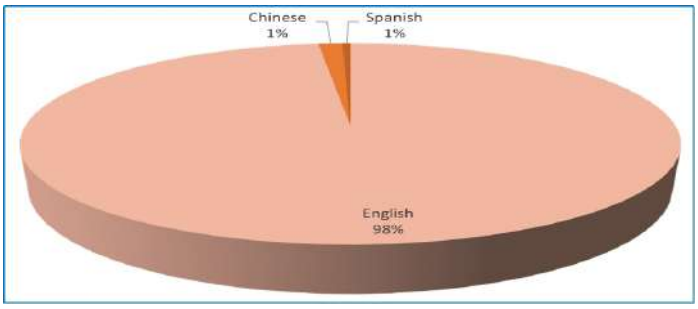

Figure 5. Percentage of the languages used for the publication for a time period (2000 – 2022)

4.8 CONTRIBUTION TO GLOBAL RESEARCH

Figure 6 shows the amount of research done on each continent determined by the number of studies that were published over a period of years 2000-2022. It is notable that in the case of an article having multiple authors, the contributions of the related authors belonging to all the countries based upon authors' associations are taken into account.

It is noticeable from Figure 6 that for the time frame of 23 years, Asia is the leading continent with the most research contributions in terms of the number of publications. For this time frame, North America is the second most contributing continent in terms of the number of research articles published in the field of Data Mining on Digital and Virtual Libraries. However, for the time duration (2000–2022) Europe also showed active participation in terms of publication frequency. However, Asian countries make a vital contribution of 41.3% of all scholarly works that have been published.

4.9 CO-WORD OCCURRENCE

The technique used to determine the pattern of keyword co-occurrence in a dataset is called co-word analysis [30][31]. According to the frequency of the term, co-word analysis indicates the relationships between

Figure 6. Continent wise distribution of publications for a time period (2000 – 2022)

the articles [31][32] [32][33]. The top 50 most frequent keywords are shown in Figure 7 as a word cloud using the software "bibliometrix (Studio R)" [33][34]. However, VOSviewer is used to do the co-word analysis by creating network visualization. Only the top 20 keywords are represented in Figure 8 for improved visibility. To create a map displaying keyword co-occurrence, we employed the full counting approach. Similar or related themes are displayed by keywords belonging to the same cluster. The total number of links among the top 20 keywords shows that each keyword is connected to every other term.

Figure 7. Word cloud comprising of top 50 keywords in accordance with the research publication for a time period (2000 – 2022)

4.10 RESEARCH THEMES

The relationships between themes, subjects, and trends are clarified by networks of co-occurrences or co-word analysis. It is the only strategy that applies verified study findings. Consequently, a study unit is a

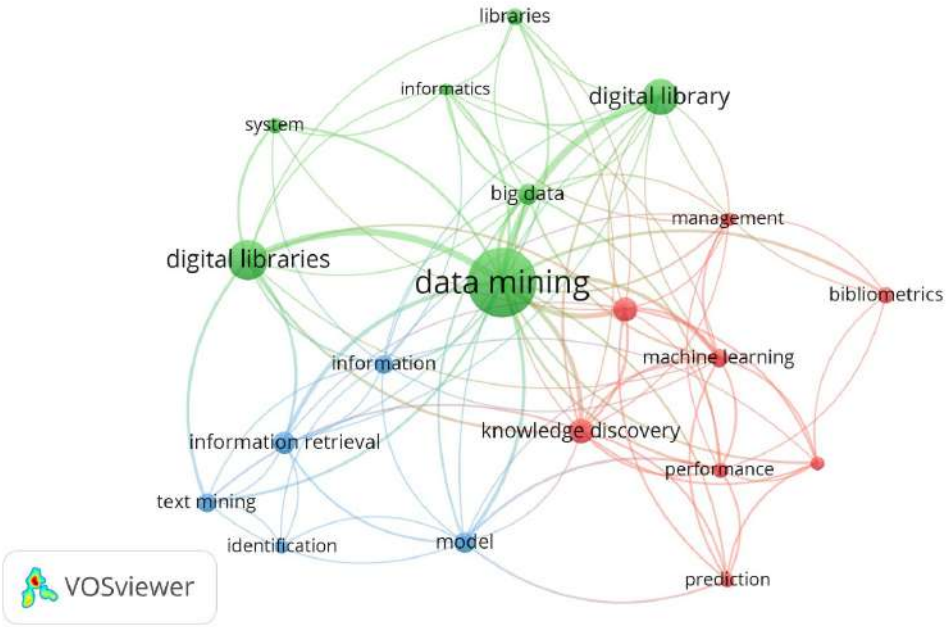

Figure 8. Co-word occurrence comprising the top 20 co-words in accordance with the research publication for a time period (2000 – 2022)

notion, phrase, or topic identified across the network as a whole.

Figure 9 shows the typological pattern which is known as a "thematic map" [34][35] [35][36]. The Figure includes the motor themes, an important research idea that is gaining traction, in the first quarter (top right area); The second quarter (top left) features highly developed, isolated elements that are currently the subject of specialized investigation; the emerging theme is represented in the third quarter (lower left region); The information regarding the basic components, which are crucial to the field, have not been developed in depth but frequently make references to important research areas can be found in the fourth quarter (bottom right side). The topic dendrogram displays the hierarchical order and connectedness between the keywords generated via hierarchical categorization as shown in Figure 10.

Figure 10 display word clusters in red and blue that correspond to terms in the data that are related to one another. The most common subjects, their connections to other subjects, and the color-coded grouping of these subjects are shown in the topic

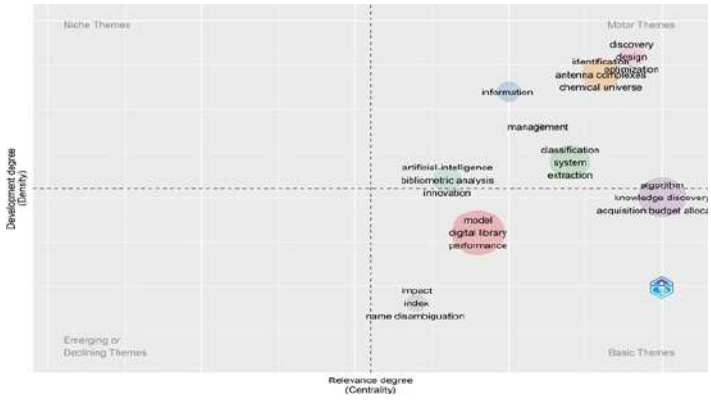

Figure 9. Thematic Map in accordance with the research publication for a time (2000 – 2022)

dendrogram (Figure 10). Based on how frequently they are related to one another, as seen by the red hue in Figure 10, the common subjects are clustered under a single cluster. This further demonstrates that, in contrast to the themes in the red color classification, those in the blue color classification are less frequently used. After that, each of them is divided into numerous sets, each set into numerous sub-sets, and so on until the topic has been used, at which point

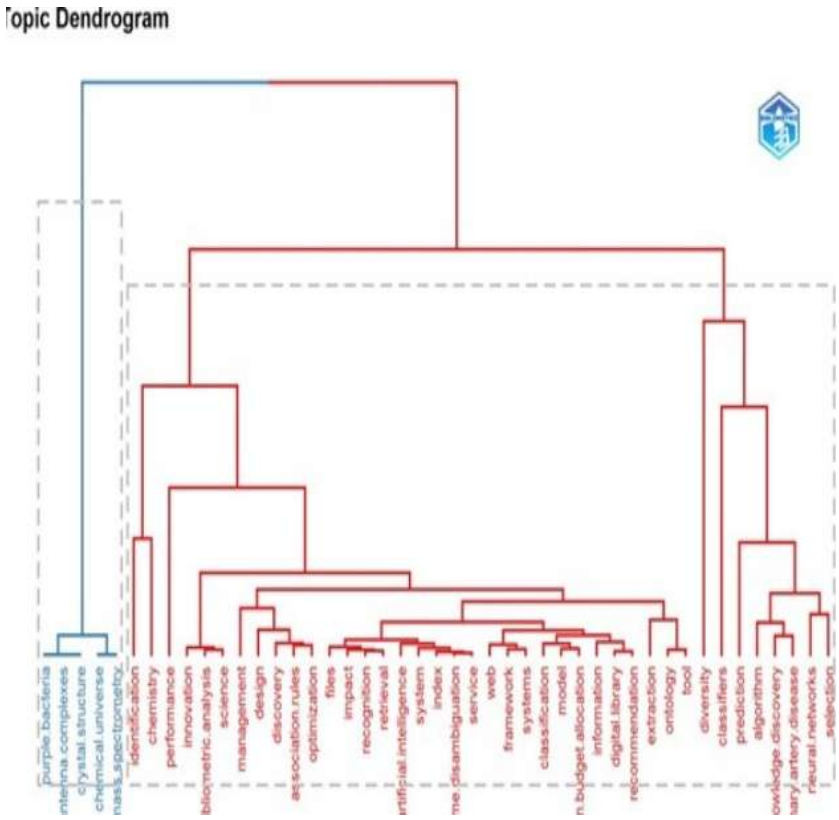

Figure 10. Topic Dendrogram in accordance with the research publication for a time (2000 – 2022)

a large number of topics are incorporated into one group, demonstrating the correlation across some topics in study articles in the discipline of data mining on digital and virtual libraries.

4.11 TOP 20 FUNDING AGENCIES

It is important to know the agencies which fund the studies (Table 11) in specific research areas so that the researchers can apply for the grant which is required to carry out meaningful research. The National Science Foundation, the National Natural Science Foundation of China, and the European Commission are the top three funding agencies that support research studies using data mining in the field of digital libraries.

5 Data Mininig applications in Digital Library

A Digital library is digitalized information backed by sophisticated technology spreading knowledge from one part of the globe to another. It uses multiple technologies to store, update and retrieve data to enable the user to access information quickly. It is more than just a library's website on the Internet; it is a distributed, object-oriented, and platform-independent collection of digital content. Based on different processing items, data mining of digital libraries can be categorized into three categories. The main areas of application are shown in Figure 11.

a.Digital Library Structure Mining: Data mining allows one to gain and learn knowledge from organizational structure and links. Mining the hyperlinks and the URL of documents helps understand the internal and external structures of a document. Interconnections between documents help in mining user digital libraries' complete information and document content.

b.Digital Library Content Mining: Content mining based on the digital library is used to learn profound knowledge via pattern recognition and analysis of the information in the textual and multimedia content. It includes five aspects: data organization, automatic feature extraction, text summarizer, automatic

Table 6. Top 20 WoS categories over a time span (2000–2022)

Web of Science Categories	Record Count	% of 215
Computer Science Information Systems	69	32.093
Information Science Library Science	66	30.698
Computer Science Theory Methods	50	23.256
Computer Science Artificial Intelligence	39	18.14
Computer Science Interdisciplinary Applications	36	16.744
Engineering Electrical Electronic	30	13.953
Computer Science Software Engineering	10	4.651
Engineering Multidisciplinary	8	3.721
Education Educational Research	7	3.256
Mathematics Applied	7	3.256
Telecommunications	7	3.256
Business	5	2.326
Education Scientific Disciplines	5	2.326
Materials Science Multidisciplinary	5	2.326
Operations Research Management Science	5	2.326
Chemistry Medicinal	4	1.86
Chemistry Multidisciplinary	4	1.86
Management	4	1.86
Multidisciplinary Sciences	4	1.86
Radiology Nuclear Medicine Medical Imaging	4	1.86
Social Sciences Interdisciplinary	4	1.86

document classifier, and automatic information organization. Data mining prioritizes the best search path and organizes search keywords based on logic corresponding to the target.

c.Digital Library User's History Mining: Data mining technology helps examine log files and mine access patterns and provides website management decisions and personalized services. User history mining implies server logs, cookies, user registration data, e-mail queries, and web purchase data. Track of user log helps analyze individual user preferences to keep the corresponding customized services per the user access log. Moreover, data analysis of the mined data helps digital libraries predict customer behavior.

6 Future work and Limitations of the Research study

6.1 FUTURE WORK

This study leverages the Web of Science database, delving into bibliometric research to examine the utilization of data mining in digital/virtual libraries.

For a more comprehensive analysis, future studies could consider expanding their database repertoire to encompass platforms such as PubMed or Scopus. Despite the thorough exploration encapsulated within our work, which encompasses eleven well-defined research questions, there exists the potential for future investigations to introduce additional inquiries. These supplementary research questions could facilitate a more exhaustive exploration of various facets related to the application of data mining in the context of digital and virtual libraries.

Table 7. Top 20 Research Areas over the time period (2000 – 2022)

Research Areas	Record Count	% of 215
Computer Science	133	61.86
Information Science Library Science	66	30.698
Engineering	47	21.86
Education Educational Research	10	4.651
Business Economics	8	3.721
Mathematics	7	3.256
Telecommunications	7	3.256
Chemistry	5	2.326
Materials Science	5	2.326
Operations Research Management Science	5	2.326
Pharmacology Pharmacy	4	1.86
Physics	4	1.86
Radiology Nuclear Medicine Medical Imaging	4	1.86
Science Technology Other Topics	4	1.86
Social Sciences Other Topics	4	1.86
Automation Control Systems	3	1.395
Biochemistry Molecular Biology	3	1.395
Health Care Sciences Services	3	1.395
Medical Informatics	3	1.395
Biophysics	2	0.93
Biotechnology Applied Microbiology	2	0.93
Imaging Science Photographic Technology	2	0.93
Public Administration	2	0.93
Public Environmental Occupational Health	2	0.93

6.2 LIMITATIONS OF THE STUDY

Limited timespan: Research published in 23 years (2000–2022) timespan is included. **Use of ISI Web of Science (WoS):** WoS is the most commonly used database but other databases such as Scopus and IEEExplore may also be included.

7 Conclusion

Data science is revolutionizing the way digital libraries manage and store information. Digital libraries allow users to access information quickly and easily. However, traditional methods of organizing and managing data are becoming increasingly outdated. Data science is providing a way to improve the efficiency of digital libraries. Data science can be used to analyze large amounts of data in order to optimize data storage, search algorithms, and access control. Data science

Table 8. Top 20 institutions from 2000 to 2022 ranked by the quantity of publications

Institutions/ Organizations	Countries	Record Count	% of 215
University of Illinois System	United States	8	3.721
Pennsylvania State University	Pennsylvania	5	2.326
Dalian University of Technology	China	4	1.86
National Yang Ming Chiao Tung University	Taiwan	4	1.86
University of North Carolina	United States	4	1.86
Beijing Institute of Technology	China	3	1.395
Bell Labs	United States	3	1.395
Chiang Mai University	Thailand	3	1.395
Chinese Academy of Sciences	China	3	1.395
Fraunhofer Gesellschaft	Germany	3	1.395
Hebei University of Engineering	China	3	1.395
Indiana University Bloomington	United States	3	1.395
Indiana University System	United States	3	1.395
Jilin Agricultural University	China	3	1.395
Kyushu Institute of Information Sciences	Japan	3	1.395
Mohammed V University in Rabat	Morocco	3	1.395
National Changhua University of Education	Taiwan	3	1.395
North Carolina State University	United States	3	1.395
Thammasat University	Thailand	3	1.395
Universiti Teknologi Malaysia	Malaysia	3	1.395
University of Belgrade	Serbia	3	1.395
University of Murcia	Spain	3	1.395
University of Tokyo	Japan	3	1.395
Virginia Polytechnic Institute State University	United States	3	1.395

can also be used to customize user experiences and create personalized recommendations based on user preferences. Data science can even be used to identify and prioritize information that is most relevant to a user's needs. By utilizing data science, digital libraries can become more user-centric, making it easier for users to find the information they need quickly and efficiently. As the use of data science increases, digital libraries are becoming more efficient and effective, providing users with a better experience.

This research study evaluates 215 publications from 2000 to 2022 according to the 10 research questions. The article presents an extensive bibliometric assessment in the field of Data Mining applied to Digital/ Virtual Libraries. Our findings conclude that the impact of Data Mining for Digital/ Virtual Libraries is not only limited to the Library Science and Computer Science field but extends to multi-disciplinary areas such as Engineering, Education Educational Research, Business Economics, Mathematics, and Telecommunications. The research publication trend shows that the year 2017 has the maximum number of research contributions in terms of publication frequency.

The evaluation of the top 20 institutions/organizations reveals that 33% of these institutions/organizations

Table 9. Top 20 institutions/organizations in terms of publication count for a time span (2000 – 2022)

Countries	gray**Record Count**	% of 215
lightgray USA	58	26.977
Peoples R China	53	24.651
lightgray Taiwan	14	6.512
England	13	6.047
lightgray Germany	11	5.116
France	8	3.721
lightgray India	7	3.256
Italy	6	2.791
lightgray Japan	6	2.791
Spain	6	2.791
lightgray Pakistan	5	2.326
Saudi Arabia	5	2.326
lightgray Australia	4	1.860
Malaysia	4	1.860
lightgray Mexico	4	1.860
Thailand	4	1.860
lightgray Canada	3	1.395
Cuba	3	1.395
lightgray Greece	3	1.395
Morocco	3	1.395
lightgray Serbia	3	1.395
South Korea	3	1.395
lightgray Brazil	2	0.930
Egypt	2	0.930
lightgray Macedonia	2	0.930

Table 10. Languages used for the research publications over 2000 – 2022

Languages	Record Count	% of 215
English	211	98.14
Chinese	3	1.395
Spanish	1	0.465

Table 11. Top 20 funding agencies for the research publications over 2000 – 2022

Funding Agencies	Record Count	% of 155
National Science Foundation	**18**	5.488
National Natural Science Foundation of China	**7**	2.134
European Commission	**6**	1.829
Spanish Government	**4**	1.220
Air Force Office of Scientific Research	**3**	0.915
National Institute of Health USA	**3**	0.915
United States Department of Defense	**3**	0.915
United States Department of Energy	**3**	0.915
United States Department of Health and Human Services	**3**	0.915
Chinese Knowledge Center Of Engineering Science And Technology	**3**	0.915
[HTML]F1F1F1 Consejo Nacional De Ciencia Y Tecnologia	**2**	0.610
Coordenacao De Aperfeicoamento De Pessoal De Nivel Superior	**2**	0.610
European Commission Joint Research Centre	**2**	0.610
German Research Foundation	**2**	0.610
Institute Of Museum And Library Services	**2**	0.610
National Basic Research Program Of China	**2**	0.610
National Key Research And Development Program Of China	**2**	0.610
Nih National Library Of Medicine	**2**	0.610
Nsf Directorate For Social Behavioral Economic Sciences	**2**	0.610
Uk Research Innovation	**2**	0.610
Universiti Tun Hussein Onn Malaysia	2	0.610

belong to the USA for the timespan 2000-2022. However, the other prominent country as per research output is China to which 18% of the institutions/ organizations belong. This further affirms the fact that the USA and China lead the research in the field of Data Mining with respect to Digital/ Virtual Libraries.

The results of the publication medium study show that while English makes up a sizable portion of the published papers, there are also a modest number of scientific articles that are published in languages besides English. Co-word occurrences, research themes, and emerging research topics with respect to the area of Data Mining for Digital/ Virtual Libraries are also part of our study.

Our research study's major objectives are (i) to provide a thorough bibliometric evaluation for Data Mining on Digital/Virtual Libraries in accordance with the nine carefully crafted research questions. (ii) to empower the researchers to select the institution/organization and countries for their respective research area in a more informative manner. (iii) to compare the variations in research publishing patterns over the last 23 years depending on the frequency of published works. (iv) to represent the top 20 countries and institutions or organizations in terms of published scientific studies. (v) to illustrate the impact of Data Mining on Digital/ Virtual Libraries in various diverse fields and domains. (vi) to display

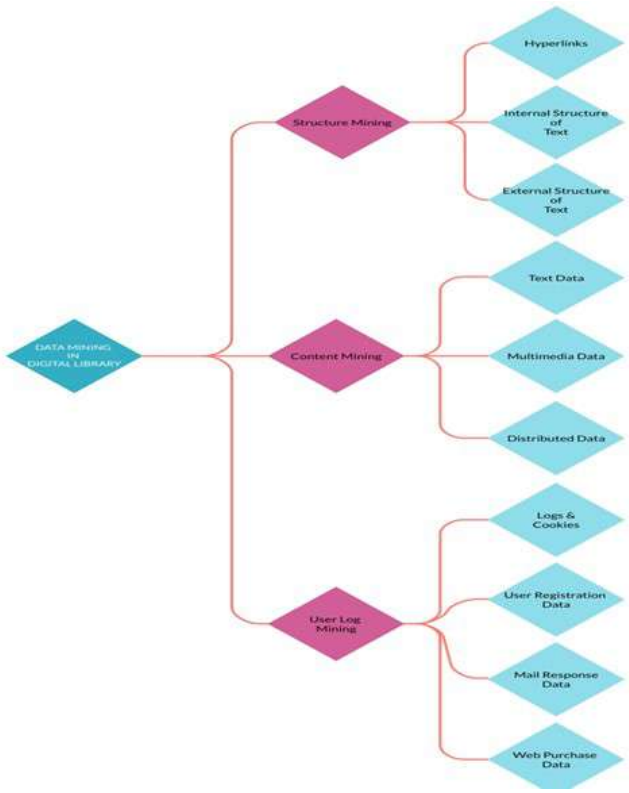

Figure 11. Application Areas of Data Mining In Digital Libraries

relatedness among various scientific studies in terms of co-words. (vii) to present the concepts, topics, and research themes for Data Mining on Digital/Virtual Libraries.

Author Contributions

Sana Alam: Conceptualization, Methodology, Software **Shehnila zardari**: Data curation, Writing- Original draft preparation. **Umm-e-Laila**: Visualization, Investigation. **Muhammad Abbas**: Supervision.: **Muhammad Asghar Khan**: Software, Validation. **Noor Ul Huda**: Writing- Reviewing and Editing

Compliance with Ethical Standards:

It is declared that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining." American Association for Artificial Intelligence, 1996.

[2] G. Patetsky-Shapiro, "Discovery, analysis and presentation of strong rules," *Knowledge Discovery and Databases*, 1991.

[3] A. Silberschatz, M. Stonebraker, and J. Ullman, "Database research: achievements and opportunities into the 1st century," *ACM sIGMOD record*, vol. 25, no. 1, pp. 52–63, 1996.

[4] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[5] D. C. Weaver, "Applying data mining techniques to library design, lead generation and lead optimization," *Current opinion in chemical biology*, vol. 8, no. 3, pp. 264–270, 2004.

[6] C.-C. Chang and R.-S. Chen, "Using data mining technology to solve classification problems: A case study of campus digital library," *The Electronic Library*, vol. 24, no. 3, pp. 307–321, 2006.

[7] Y. Cheng, Y. Kuang, X. Shi, and C. Dong, "Sustainable investment in a supply chain in the big data era: An information updating approach," *Sustainability*, vol. 10, no. 2, p. 403, 2018.

[8] Y. Li, "Analysis of the construction of the library information resources from the perspective of big data," *Applied Mechanics and Materials*, vol. 631, pp. 1067–1070, 2014.

[9] M. Zhang, "Application of data mining technology in digital library." *J. Comput.*, vol. 6, no. 4, pp. 761–768, 2011.

[10] N. Will, "Data-mining: Improvement of university library services," *Technological Forecasting and Social Change*, vol. 73, no. 8, pp. 1045–1050, 2006.

[11] J. M. Merigó and J.-B. Yang, "A bibliometric analysis of operations research and management science," *Omega*, vol. 73, pp. 37–48, 2017.

[12] S. Alam, S. Zardari, and J. Shamsi, "Comprehensive three-phase bibliometric assessment on the blockchain (2012–2020)," *Library Hi Tech*, vol. 41, no. 2, pp. 287–308, 2023.

[13] B. S. dos Santos, M. T. A. Steiner, A. T. Fenerich, and R. H. P. Lima, "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018," *Computers & Industrial Engineering*, vol. 138, p. 106120, 2019.

[14] Y. Hu, Z. Yu, X. Cheng, Y. Luo, and C. Wen, "A bibliometric analysis and visualization of medical data mining research," *Medicine*, vol. 99, no. 22, 2020.

[15] S.-F. Tseng, Y.-L. Won, and J.-M. Yang, "A bibliometric analysis on data mining and big data," *International Journal of Electronic Business*, vol. 13, no. 1, pp. 38–69, 2016.

[16] C. Baek and T. Doleck, "Educational data mining: A bibliometric analysis of an emerging field," *IEEE Access*, vol. 10, pp. 31 289–31 296, 2022.

[17] S. A. Abd Karim and P. N. Nohuddin, "Bibliometric analysis of data mining on medical imaging," in *Journal of Physics: Conference Series*, vol. 1997, no. 1. IOP Publishing, 2021, p. 012017.

[18] F. Madani, "'technology mining'bibliometrics analysis: applying network analysis and cluster analysis," *Scientometrics*, vol. 105, no. 1, pp. 323–335, 2015.

[19] Q. Xiao, J. Wang, Y. Wang, and Y. Wu, "Data mining in nursing: A bibliometric analysis (1990–2017)," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*. IOS Press, 2019, pp. 1616–1617.

[20] K. Ahmad, Z. Jian Ming, and M. Rafi, "Assessing the digital library research output: bibliometric analysis from 2002 to 2016," *The Electronic Library*, vol. 36, no. 4, pp. 696–704, 2018.

[21] J. O. Hodonu-Wusu and G. N. Lazarus, "Major trends in lis research: A bibliometric analysis," *Library Philosophy and Practice*, p. 1, 2018.

[22] M. Naveed, N. Ali, S. Aslam, and N. Siddique, "Research output of the library quarterly: A bibliometric analysis during 2010-2019," *Library Philosophy and Practice*, pp. 1–15, 2021.

[23] P. Waghmare, "Global research trends on library management and administration: A bibliometric analysis," *Library Philosophy and Practice*, pp. 1–15, 2022.

[24] C. Birkle, D. A. Pendlebury, J. Schnell, and J. Adams, "Web of science as a data source for research on scientific and scholarly activity," *Quantitative Science Studies*, vol. 1, no. 1, pp. 363–376, 2020.

[25] C. Burgers, B. C. Brugman, and A. Boeynaems, "Systematic literature reviews: Four applications for interdisciplinary research," *Journal of Pragmatics*, vol. 145, pp. 102–109, 2019.

[26] X. Ding and Z. Yang, "Knowledge mapping of platform research: a visual analysis using vosviewer and citespace," *Electronic Commerce Research*, pp. 1–23, 2020.

[27] L. Ensslin, A. Dutra, S. R. Ensslin, E. A. Moreno, L. C. Chaves, and A. A. Longaray, "Sustainability in library management in higher education institutions: a bibliometric analysis," *International Journal of Sustainability in Higher Education*, vol. 23, no. 7, pp. 1685–1708, 2022.

[28] D. Guleria and G. Kaur, "Bibliometric analysis of ecopreneurship using vosviewer and rstudio bibliometrix, 1989–2019," *Library Hi Tech*, vol. 39, no. 4, pp. 1001–1024, 2021.

[29] P. R. Bhagat, F. Naz, and R. Magda, "Artificial intelligence solutions enabling sustainable agriculture: A bibliometric analysis," *PloS one*, vol. 17, no. 6, p. e0268989, 2022.

[30] B. Hamadicharef, "Scientometric study of the ieee transactions on software engineering 1980-2010," in *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science: Volume 1*. Springer, 2012, pp. 101–106.

[31] F. G. de Freitas and J. T. de Souza, "Ten years of search based software engineering: A bibliometric analysis," in *Search Based Software Engineering: Third International Symposium, SSBSE 2011, Szeged, Hungary, September 10-12, 2011. Proceedings 3*. Springer, 2011, pp. 18–32.

[32] S. Alam, S. Zardari, and M. Bano, "Software engineering and 12 prominent sub-areas: Comprehensive bibliometric assessment on 13 years (2007–2019)," *IET Software*, vol. 16, no. 2, pp. 125–145, 2022.

[33] M. Aria and C. Cuccurullo, "bibliometrix: An r-tool for comprehensive science mapping analysis," *Journal of informetrics*, vol. 11, no. 4, pp. 959–975, 2017.

[34] S. Zardari, S. Alam, H. A. Al Salem, M. S. Al Reshan, A. Shaikh, A. F. K. Malik, M. Masood ur Rehman, and H. Mouratidis, "A comprehensive bibliometric assessment on software testing (2016–2021)," *Electronics*, vol. 11, no. 13, p. 1984, 2022.

[35] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *Journal of the American Society for information Science and Technology*, vol. 62, no. 7, pp. 1382–1402, 2011.