14th Annual International Conference on Industrial Engineering and Operations Management Dubai United Arab Emirates (UAE), February 12-14, 2024

Publisher: IEOM Society International, USA DOI: 10.46254/AN14.20240547
Published: February 12, 2024

Text Mining and Data Mining Applications in Industrial Engineering: Review and Extension

Mitra Bagherini
School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran
bagherini.mitra1@gmail.com

Ali Reza Hoseini
School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran
Ali_RezaHoseini@ind.iust.ac.ir/ORCID: 0000-0003-4705-7955

Mohammad Rabiei
Assistant Professor of Iranian Research Institute for Information Science and Technology, Iran
m.rabiei@irandoc.ac.ir

Morteza Baghepour
Department of industrial engineering, Antalya Bilim University, Antalya, Turkey
abagherpour81@gmail.com

Semail Ülgen
Department of industrial engineering, Antalya Bilim University, Antalya, Turkey
Sulgen@antalya.edu.tr

Abstract
Data mining and text mining models have been frequently carried out in a variety of engineering disciplines and sciences. In this paper, we have evaluated data mining and text mining models in industrial and systems engineering. We have categorized the results by journal, country, applications, etc. The results show that there is an increase growth on all aspects of data and text mining in industrial engineering and these novel methods will change the future of industrial engineering professions.

Keywords
Data mining, text mining, industrial engineering, journals, publications

1. Introduction
In the last decade, the volume of data has grown at an unprecedented rate (Qi 2020). According to an International Data Corporation (IDC) report, the digital universe would double every two years (Agrawal 2020, Lian, Zhang et al. 2020). This explosive growth in data generation has created a need for smart technologies to analyze this massive, fuzzy, and noisy data. Data mining (DM) is the process that can meet this need and extract useful information from data by providing algorithms, models, and sophisticated data analysis tools and techniques. This process is generally known as "knowledge mining from data". DM is an important process for finding meaningful new correlations, interesting patterns, interdependencies, and trends by sifting through huge data stored in data warehouses and databases (Han, Pei et al. 2011, Larose and Larose 2014). DM has a multidisciplinary nature and is influenced by machine learning, artificial intelligence, statistics, database and data warehouse systems, and other disciplines such as pattern recognition, information retrieval, visualization, algorithms, and high-performance computing (Han, Pei et al. 2011). According to the kind of knowledge extracted, different types of functions can be defined, such as classification,

prediction, regression, clustering, association, sequential pattern, and outlier detection. These functions are implemented and supported by various techniques (Yue, Wu et al. 2007, Sharma and Panigrahi 2013). Figure 1 presents a set of important and widely used functions and techniques in DM.

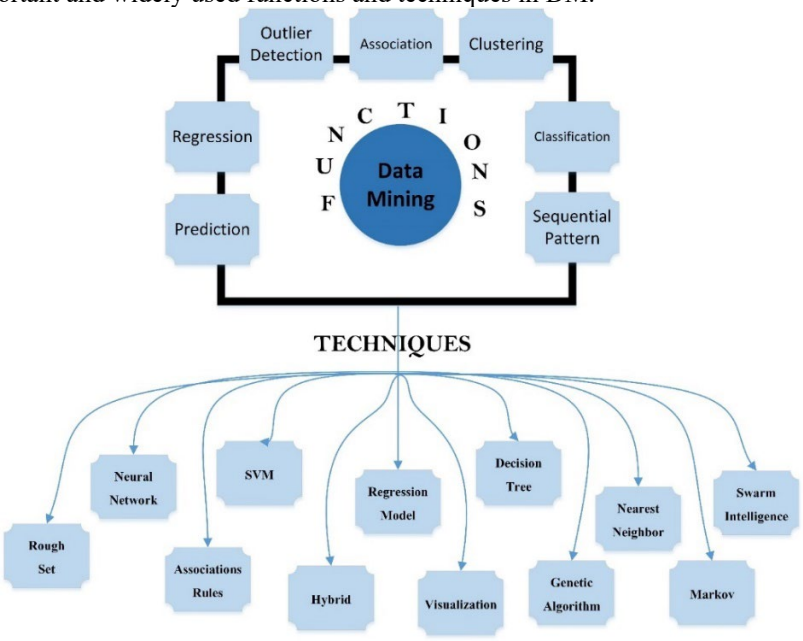


Figure 1. Data mining functions and techniques

Industrial Engineering (IE) is a broad and multidisciplinary science that dates back to the 19th century and its role in the manufacturing, government, and service organizations led to its growth in the 20th century (Zandin 2001). IE mainly focuses on how to improve processes that increase efficiency and reduce money, time, raw resources, and energy (Kosky, Wise et al. 2006). IE includes various fields, the most important sub-areas related to IE include the following: operation research and optimization, scheduling, quality and reliability engineering, supply chain management and logistics, financial, energy engineering, human factors and ergonomics (safety management), manufacturing, system simulation and stochastic process, and information system engineering and management (Salvendy 2001, Zandin 2001). In the recent years, not only the competitive environment of businesses has increased dramatically (Fernandes, Fitzgerald et al. 2019) but also the opportunity to gather and store huge amounts of data from different areas of the industry arose (Vazan, Janikova et al. 2017, Cattaneo, Fumagalli et al. 2018). due to the huge volume of data and the complexity of the relation between them, the analysis of data stored is very hard by traditional methods but DM by providing diverse tools and algorithms can be useful to extract valuable information from raw data (Cordeiro, Faloutsos et al. 2013). industrial engineers are getting increasingly dependent upon data for gaining visibility on the processes and systems behaviors and interactions with the environment. DM can be helpful to obtain a good understanding and knowledge by discovering hidden patterns and correlations and useful information. This can lead to improvement and optimization for their strategic decisions in different fields relevant to industrial procedures (Vazan, Janikova et al. 2017). Among the sections related to IE in which DM has been used, the following can be mentioned: product design (Kretschmer, Pfouga et al. 2017, Alkahtani, Choudhary et al. 2019), quality control (Ramana, Sapthagiri et al. 2016, Wei, Feng et al. 2017, Hirsch, Reimann et al. 2018), material handling (Demetgul, Yildiz et al. 2014), fault detection (Shao and Hou 2013), maintenance (Grabot 2020), supply chain (He and Song 2009, Shah-Hosseini 2013, Wu, Liu et al. 2019, Kara, Fırat et al. 2020, Ying, Chen et al. 2020), job cycle time prediction in systems and factories (Chen and Romanowski 2013, Chen 2016, Wang, Zhang et al. 2018), logistics (Kormann and Altendorfer-Kaiser 2017, van Cruchten and Weigand 2018, Knoll, Reinhart et al. 2019), marketing (Van Nguyen, Zhou et al. 2020). The dramatical growth of DM applications in IE over the last decade has led to an increase in the importance of research in this field, but existing limitations such as the uncertainly of the boundaries of IE have made research in this area difficult and it appears necessary to fully evaluate and provide a comprehensive picture of the current state of the DM application and increase productivity in the field of IE. This study aims to provide a comprehensive review of a sample of IE publications to identify and grasp the current state of the conducted research and literature in the last 10 years.

2. Literature review

One of the most important fields of IE is the supply chain, which due to its wide dimensions and its role in determining the sustainability of production, has attracted much attention in recent years, and various researches have been performed in this field. Fahimnia, et al. used bibliometric analysis to investigate research interests and topics in the field of green supply chain management, this study also identified top journals, authors, and countries in this field (Fahimnia, Sarkis et al. 2015). Also in another bibliometric study (Cancino, Amirbagheri et al. 2019), the forty years publications of Computers & Industrial Engineering (CIE) journal were reviewed and leading topics, institutions and countries in the field of IE were identified. Hosseini, developed a systematic literature review of articles focused on supply chain resilience and by using analysis of reviewed popular categories, journals, and organizations in this domain. In addition, they determined emerging trends in this area (Hosseini, Ivanov et al. 2019). In 2020, two important scientometric analyses have been performed in the field of supply chain disruption (SCD) and supply chain management. In the review article related to SCD, 1310 publications derived from the core collection of the Web of Science were examined which presents a comprehensive bibliometric overview and visualization of the field of SCD (Xu, Zhang et al. 2020). and in another study, papers from 1998 to 2017 in the field of supply chain management were examined, which analyzed issues related to supply chain management using scientometric analysis and determined the most productive authors, institutions, and countries (Yalcin, Shi et al. 2020).Some studies have investigated in other related fields of IE such as: logistics (Dolati Neghabadi, Evrard Samuel et al. 2019, Hu, Dong et al. 2019) and productivity (Jin, Leem et al. 2016). In an article published in 2020 (Rabiei, Hosseini-Motlagh et al. 2020), Rabiee has thoroughly examined the various fields of information technology, management and industrial engineering and identified important and trending issues in each field in recent years. This article shows that although computer science has been almost exclusively monopolized in the field of information technology in the past, it has been widely used in industrial engineering in recent years. Given the importance and application of this field in industrial engineering in the last few years, some studies have examined the role of data mining in some specific parts of IE. In 2009, an important study was conducted on the applications of data mining in manufacturing by Choudhary. this article provides an in-depth review of data mining applications in manufacturing knowledge areas. Choudhary's paper is a good reference for all researchers who seek to deeply understand the role of data mining in manufacturing (Choudhary, Harding et al. 2009). Also In a study in 2011(Köksal, Batmaz et al. 2011) with a review of the literature from 1997 to 2007, data mining applications in manufacturing with a focus on quality improvement issues were thoroughly investigated. In another study, papers and research work regarding DM applications in semiconductor manufacturing were examined by Espadinha-Cruz, et al (Espadinha-Cruz, Godina et al. 2021) , which analyzed issues related to semiconductor manufacturing using the structured review and bibliometric analysis to help the researchers in this field to obtain a general view on this field and discover its opportunities and gaps.

Although in the recent years, significant advances in DM applications in different fields of IE. have been reported, a thorough review of DM applications for the broad domain of IE does not exist in the literature and the majority of the studies focused on the narrow fields of IE. The present study has conducted a research based on bibliographic analysis and context analysis to fill this gap and will try to answer these questions: which sub-field of IE is leading in using DM functions and techniques? What trends can be observed with respect to countries publications in this fields? Which DM functions and techniques are commonly used for each sub-field of IE? And to identify overlooked and under examined sub-fields of IE in using DM applications.

3. Methodology

The current study employs a combination of bibliometric analysis and content analysis techniques to analyze the DM research in IE, we employed bibliometric analysis and topic modeling. The bibliometrics approach is a statistical analysis of books, articles to measure the "output" of research, institutions, and countries, to identify map the development of new fields of science (Merigó and Yang 2017). This paper will use the maps generated by VOS viewer (Van Eck and Waltman 2010) to conduct a systematic analysis of DM research. The reason for choosing Vos Viewer over other software is that it comfortably depicts large networks (Van Eck and Waltman 2014).

3.1. Data Acquisition

To collect data, the ISI Web of Science (WOS) repository was utilized. WOS is one of the reliable sources which has a comprehensive coverage of prominent and influential journals around the world (Shi and Liu 2019, Rabiei, Hosseini-

Motlagh et al. 2020). To exhaustively cover all papers related to applications of DM in IE, and minimize potential bias in selecting articles, a search string as a combination of DM keywords with Boolean operators in specific IE journals was performed. Since boundaries of IE are not clear, it is tough to specify the journals which can be regarded as IE source and to choose an appropriate data set. At first, the list of journals of IE was obtained from different articles (Kao 2009, Cancino, Amirbagheri et al. 2019), then the prepared list was presented to the experts for a survey. In addition to determining the prominent journals, experts could add self-recognized journals to the bottom of the list. Finally, all journals were ranked and 25 top journals were selected, the results were reported in Table1. After several tests, a retrieval search string was performed in the title, abstract, and keywords lists of papers to identify studies related to DM in IE. Figure 2 Illustrates the structure of the search string adopted in this study. We limited our analysis to documents published in journals that are listed in Table1 and the time span was set between 2011-2020 and applied the English language on the database retrieval procedure. A total of 21397 records were the recognized. The data was downloaded in XML and Plain Text formats. The Plain Text file was imported to VOS viewer to scientifically map. in order to increase the accuracy and quality of the analysis results, data must be preprocessed before data analysis. The initial data collected was made from 21397 documents. the retrieved articles from WOS were checked in terms of duplicates, errors and omissions and the dataset was corrected. Finally, the data reduced the total number of publications in the dataset to 20201.

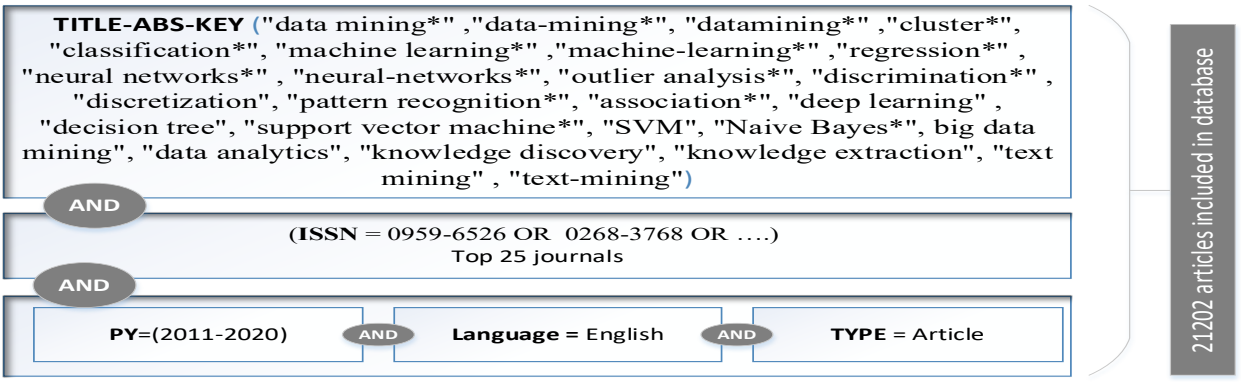


Figure 2. Overview of searching strategies

4. Results and Discussions

In this part of the study, the results of bibliometric analysis such as: publication and citation structure and the leading authors, institutions and countries are presented in Table 1. NP = Number of total Published in journal, TP = Total Publications in DM field, TC = Total Citation, TC/TP = Total Citation / Total Publications ratio (to measure the impact of each article) IF = a 5-year impact factor. It is worth noting that the importance of analyzing academic journals in any scientific field has been emphasized by many studies. Journals in Table1 are listed according to their Total number of Publications in Data mining (TP). It is seen in Table1 that the journal of Expert Systems with Applications has been the most productive journal in terms of TP, followed by Applied Soft Computing and Information Sciences. Interestingly, these three sources monopolize 42.31% of all of the available documents. Out of 25 journals, 14 are published in the United Kingdom. The rest of the journals are respectively published in Netherlands and United States. As for the journals with the highest number of citations per published article, the ranking is led by the Applied Energy (29.78) followed by the International Journal of Production Economics (26.4), and Management Science (25.8) in the third position.

Table 1. The selected journals based on the number of publications in DM: 2011-2020

Rank	Journal	Country	NP	TP	TC	TC/TP	%TP	IF
1	Expert Systems with Applications	UK	8440	4275	91192	21.33	19.98	5.448
2	Applied Soft Computing	Netherland	5656	2479	41747	17.3	11.59	5.390
3	Information Sciences	Netherland	6688	2300	48222	21.06	10.74	5.563
4	Journal of Cleaner Production	Netherland	19427	1753	29216	16.6	8.19	7.491
5	Advanced Manufacturing Technology	UK	13212	1207	12229	10.15	5.64	2.925
6	Energy	UK	14945	1150	23810	21.29	5.37	6.046
7	Applied Energy	UK	12807	1093	31902	29.78	5.11	9.086
8	European Journal of Operational Research	Netherland	6008	1015	17602	17.34	4.75	4.729
9	Engineering Applications of Artificial Intelligence	UK	2062	986	15829	16.05	4.50	3.810
10	IEEE Transactions on Industrial Informatics	US	3168	716	18418	25.72	3.35	9.008
11	Safety Science	Netherland	2640	546	8070	14.78	2.55	4.358
12	Energy Policy	UK	6879	527	12058	23.18	2.47	5.693
13	International Journal of Production Research	UK	4471	505	6645	13.2	2.36	4.145
14	Computers & Industrial Engineering	UK	3748	464	5766	12.56	2.16	4.296
15	Reliability Engineering & System Safety	Netherland	2635	443	7746	17.48	2.07	5.156
16	International Journal of Production Economics	Netherland	3080	320	8324	26.4	1.50	6.205
17	Construction Engineering and Management	US	146	264	2792	11.07	1.23	2.672
18	Management Science	US	2085	215	5486	25.8	1.01	5.469
19	Ergonomics	UK	1356	207	2121	10.3	0.97	2.548
20	Industrial Management & Data Systems	UK	907	202	3164	15.7	0.94	4.379
21	Computers & Operations Research	UK	2339	191	3091	16.18	0.89	3.804
22	Applied Ergonomics	UK	1741	186	2199	11.8	0.87	3.054
23	Journal of Manufacturing Systems	Netherland	758	134	1880	15.8	0.63	4.963
24	Operations and Production Management	UK	1231	121	1540	13	0.57	5.676
25	Production Planning & Control	UK	918	98	1059	11.03	0.46	3.930

4.2. Distribution of Publications by Countries

Table 2 shows the evolution of the number of articles on selected journals in the 20 countries with the highest number of articles published from 2011-2020. China as the country with the largest citation and output of academic papers produced 6352 (accounting for 30.35%) in total, followed by the USA (3090 documents, accounting for 14.77%) and England (1508 articles accounting for 7.21%). the presence of the developing country of Iran among the developed countries is significant. The results show that Asian countries were the main driving force behind publishing data mining's scientific development in IE. Table1 also gives the indicator such as PD to measure the influence of the top productive countries (production num/population*100000). Singapore (6) and Taiwan (4.2) have been founded to be the most productive country while a country like India that is 5th in terms of total publications, has the lowest total publications per capita.

Table 2. The top 20 countries based on publications per year

Rank	Country	Documents	Citations	Population	PD
1	China	6293	116887	1,439,323,776	4.37
2	USA	3057	62084	331,002,651	9.23
3	England	1494	32391	67,886,011	22.01
4	Spain	1416	26181	46,772,033	30.27
5	India	1247	26732	1,380,004,385	0.9
6	Iran	1030	21102	83,992,949	12.27
7	Taiwan	1024	21858	23,816,775	43.01
8	Australia	973	17667	25,693,059	37.87
9	Canada	924	17244	37,742,154	24.48
10	Italy	887	15911	60,461,826	14.67
11	South Korea	881	14740	51,269,185	17.18
12	Brazil	780	12013	212,559,417	3.66
13	France	724	14781	65,273,511	11.09
14	Turkey	688	17799	84,339,067	8.15
15	Germany	635	11739	83,157,000	7.34
16	Malaysia	410	11833	32,365,999	12.67
17	Japan	357	6645	126,476,461	2.82
18	Singapore	353	7256	5,850,342	60.34
19	Poland	351	7425	37,846,611	9.27
20	Netherlands	342	6527	17,134,873	19.96

4.3. Most Relevant Institutions

The top 20 most productive institutions for publications are indicated in Table 3. It is evident from the Table 3 that the Chinese Academy of Sciences out with the largest number of publications (352), followed by the Hong Kong Polytechnic University with 300 publications, and Shanghai Jiao Tong University with 245 documents. Among the top 20 institutions, fourteen were in China, which reflected the outstanding research achievements made by China in the field of Data Mining. The Chinese Academy of Sciences also has the highest number of citations, with a total of 9836. This institution holds the fourth position with 27.95 Average Citation. However, Beijing Institute of Technology was in the twentieth position according to the number of publications index, considering the average citation, it occupies the first place. Interestingly, Islamic Azad University, located in the developing country of Iran, it the second most productive institution with average citation 29.12.

Table 3. Top 20 institutions with the most highly published

Rank	Institutions	Country	documents	citations	Average Citation
1	Chinese Academy of Sciences	China	352	9836	27.95
2	Hong Kong Polytechnic University	China	300	7063	23.55
3	Shanghai Jiao Tong University	China	245	5590	22.82
4	Huazhong University of Science and Technology	China	244	4793	19.65
5	City University of Hong Kong	China	227	4948	21.8
6	Tsinghua University	China	227	3898	17.17
7	Tianjin University	China	191	3110	16.28
8	Islamic Azad University	Iran	186	5417	29.12
9	Nanyang Technological University	Singapore	172	3675	21.37
10	Dalian University of Technology	China	171	3335	19.5
11	Indian Institutes of Technology	India	159	3766	23.69
12	Zhejiang University	China	158	2962	18.75
13	Xi'an Jiaotong University	China	157	2668	16.99

4.4. Text mining

In order to assess documents from multiple viewpoints, a text mining-based approach has been applied on 21398 papers. In this approach, the data set is categorized based on the 3 categories of IE sub-areas, DM functions, and DM techniques. each category also comprises various clusters which will be discussed in more detail in the later sections. selecting the appropriate fields as representative terms that describe the whole of the text is very important to the result of categorizing. In this study, the combination of author keywords (AK), keywords plus (KP), title, and abstract were selected to construct a dictionary (Dic) set for each document:

$$Dici = \{AK \cup KP \cup Abstract \cup Title \\ if\ AK \cup KP \cup Abstract \cup Title \neq \emptyset\}$$

It is notable that although examination of abstracts provides satisfactory coverage of document, the accuracy of categorizing result increased when the analysis carried out on the abstract along with the title, AK, and KP because considering the combination of these sections together provides the highest amount of coverage and makes the better condition. Table 4 exhibits the amount of coverage of the AK, KP, title and abstract when placed as the text representative alone for each of the three categories.

Table 4. Top 20 institutions with the most highly published

	AK	KP	Title	Abstract	All
Sub-areas	48.9%	36%	40.5%	82.3%	85.8%
Functions	37.8%	28.4%	31.3%	76.8%	87.4%
Techniques	37.1%	23.8%	24.7%	64.8%	75.5%

Prior to data analysis, the dictionary is cleaned and pre-processed. the following 2 rules are carried out in this step:

1. Two different forms of the same words were corrected such as: data mining and data-mining
2. Abbreviation is replaced with the original words such as: SCM and supply chain management.

in this approach, we aim to examine the data set to identify patterns and trends in categories (G) of IE sub-areas, DM functions, and DM techniques.to this end, first, in consultation with the experts in the IE and DM fields, lists of words related to each cluster of categories were prepared, then sets of each cluster (c) were constructed based on the occurrence of the words of a list (L_c) in the dic's of papers:

$$categories = \{G: IE\ sub-areas, DM\ functions, DM\ techniques\}$$
$$G = \cup ci$$

$$ci = \{pk|\ i \in \{1,2,\dots,|G|\}, k \in \{1,2,\dots,|Dic|\},$$
$$\ni term, term \in Lci \wedge term \in Dicj\}$$

It is clear that the intersection of sets (Ci) of a category may not be empty because there is a possibility of overlapping between different clusters of a category. then in order to examine the interactions which exist between different clusters of categories, a co-word network analysis was employed. The co-word analysis is a powerful content analysis technique that received sufficient attention in recent years and identifies relationships among topics in texts (Hu and Zhang 2015, Leung, Sun et al. 2017, Nájera-Sánchez 2020). Since each set (Ci) represents the papers related to clusters of a category, the frequency of co-occurrence of papers between 2 sets is one of the good indexes which based on that the linkage among clusters can be analyzed. The clusters of the two categories will be connected to each other with a specific amount of relevancy. In other words, cluster y is connected to cluster x with the connection level of $\mathrm{p}\ (xi, yj)$. In this approach, we need to set a threshold for the connection level to examine the intensity of the relation between the two clusters. Actually, the aim is to find points that maximize coverage of the sets and maintain the integrity of the networks in the general state where the connection is established between all clusters of the two categories. Applying the threshold, it can be said that, set x is connected to set y if $\mathrm{p}\ (xi, yj) \geq th_1$.Therefore, for the cluster x and y, and threshold th_e, the set E contains all common papers between x and y. More formally:

$$E(ci(x), cj(y)) = \{pz \mid pz \in ci(x) \wedge pz \in cj(y) \wedge \mathrm{p}(ci(x), cj(y)) \geq the,$$
$$x, y \in categories,$$
$$x \neq y$$
$$i = \{1,2,\dots,|x|\},$$
$$j = \{1,1,\dots,|y|\},$$
$$z = \{1,2,\dots,|Dic|\}\}$$

4.5. Analysis by Sub-areas:

Assessing the application rates of DM functions and techniques in different areas related to IE, the present study tries to explore the under looked and overlooked areas in terms of DM applications. In this category, there are 10 clusters that are the principal areas on which IE focuses. As mentioned in Table 2, out of 21398 papers, 18400 papers (Almost 85.7%) appeared in 10 clusters. Figure 3 represents the number of papers appears in "operation research and optimization", "scheduling", "quality and reliability engineering", "supply chain management and logistics", "financial", "energy engineering", "human factors and ergonomics", "manufacturing", "system simulation and stochastic process", and "information system engineering and management". It is clear that there is an overlap between the areas due to the lack of a clear border and each can be considered a subset or part of the other field. operation research and optimization" has one of the largest clusters with the 7711 papers, the reason behind the growth of DM applications in this field can be linked with the rapid growth of data and the possibility of collecting and storing that in various industrial fields that led to discover knowledge for the explore the optimal solution to solve problems and decision-making process that are the principal objectives of the optimization process and operational research (Wang, Chaovalitwongse et al. 2010). It is found that areas such as human factors and ergonomics, scheduling and information system engineering and management have not gained sufficient attention in terms of data mining applications.

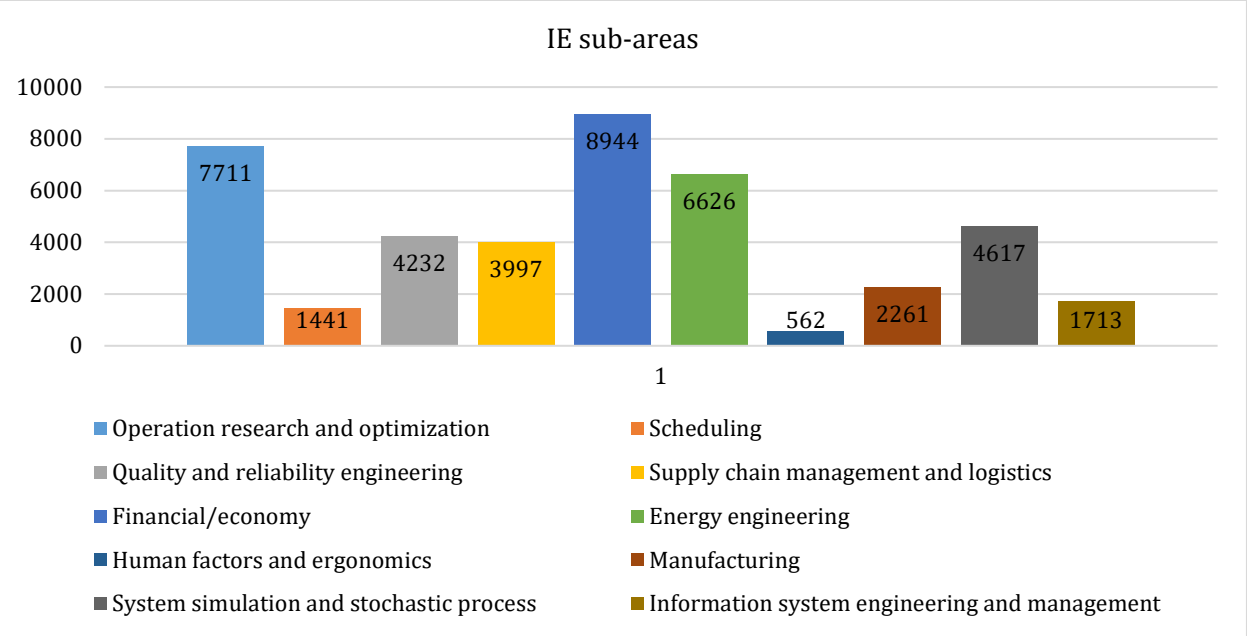
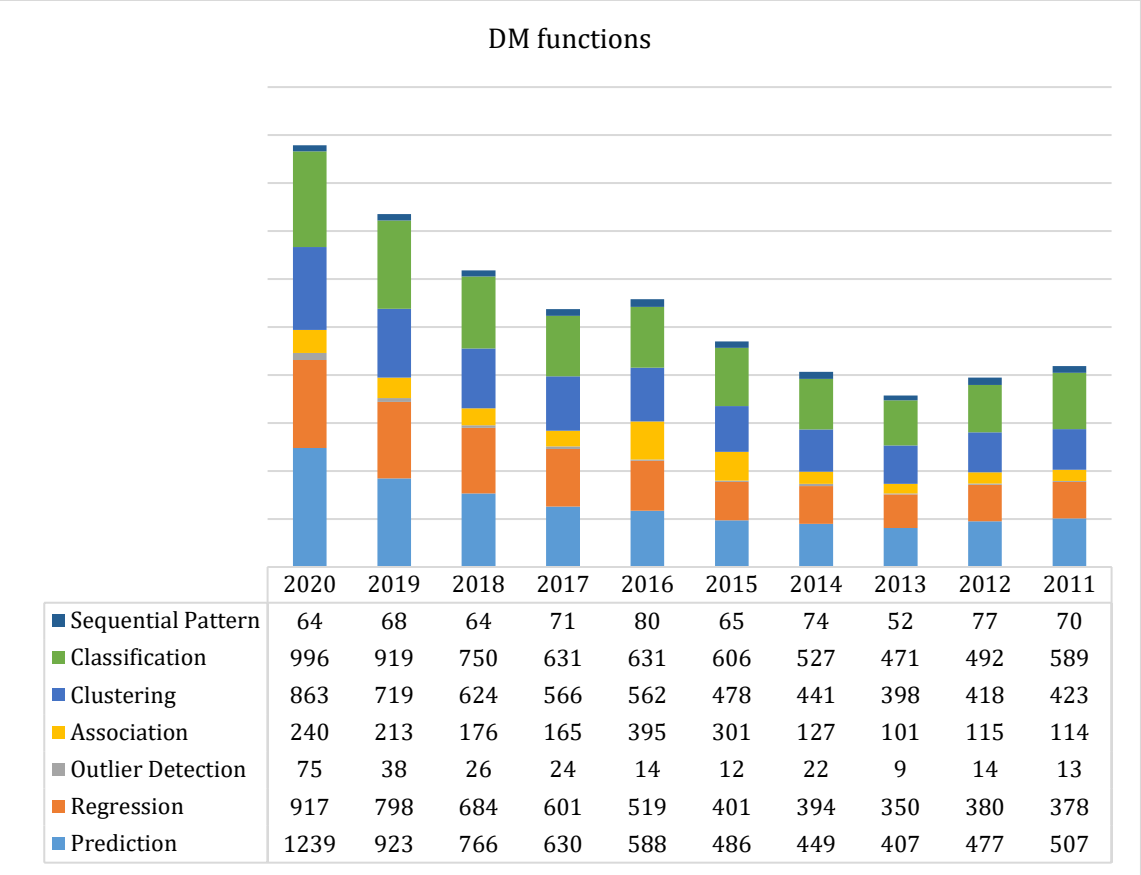


Figure 3. Industrial engineering sub areas

4.6. Functions and techniques

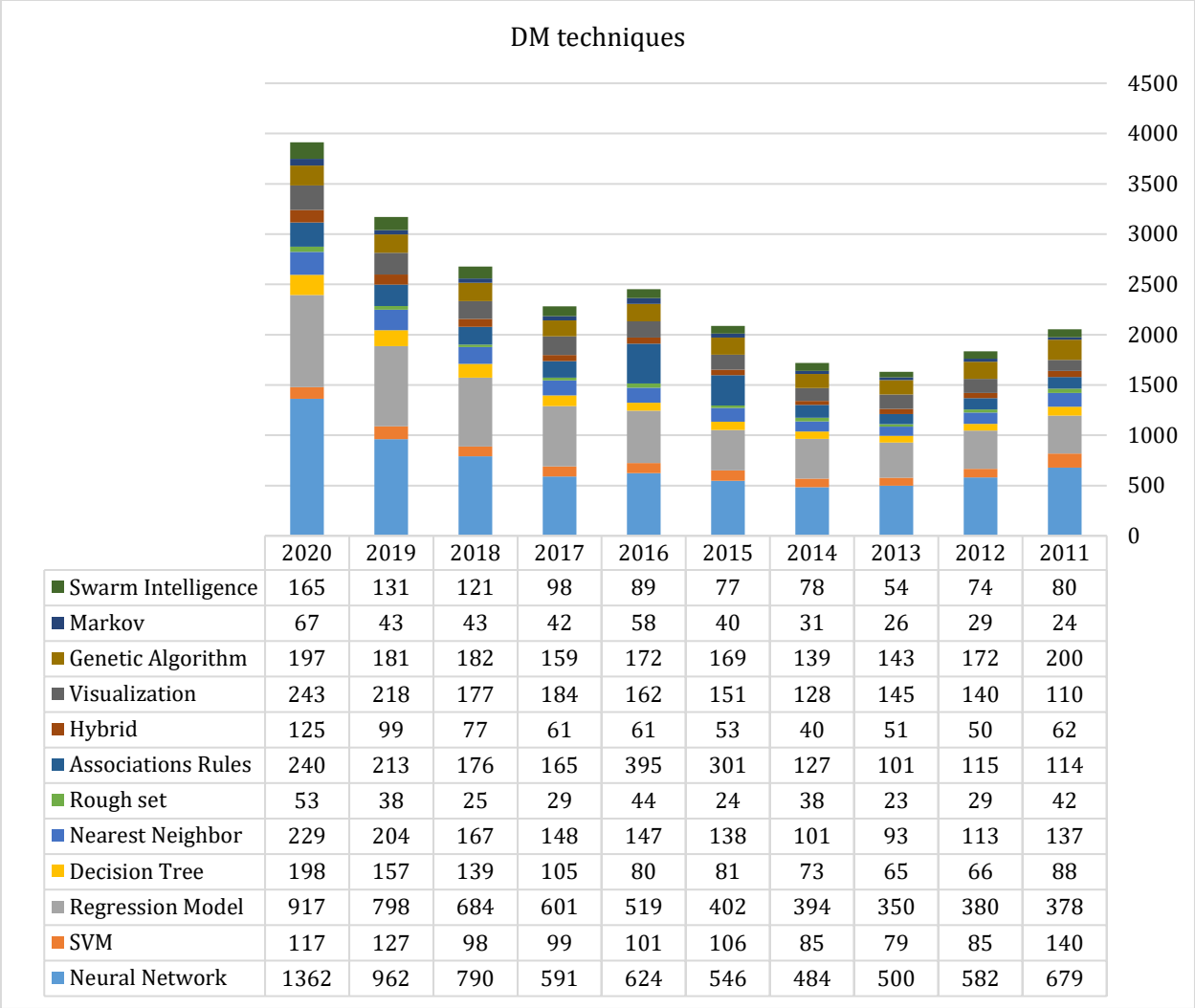
Figures 4, 5 and 6 show the applications of each of the data mining functions and techniques in the database, respectively. As mentioned in the previous section, a little part of the data does not fall under the specified functions and techniques, however, it is clear that the sum of functions and techniques used is greater than all documents because some articles have used more than one type of function and technique. The major functions considered include: classification, clustering, prediction, regression, association, sequential pattern and outlier detection. The results show that classification is one of the powerful functions that plays an essential role in the various domains.in fact, classification induces a process that helps organize data sets by classifying data into target classes (Ngai, Hu et al. 2011). Prediction, clustering, and regression are other functions that have widely used, and a large part of the data allotted to them each year, conversely, association, sequential pattern, and outlier detection have received the least attention each year.



	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011
Sequential Pattern	64	68	64	71	80	65	74	52	77	70
Classification	996	919	750	631	631	606	527	471	492	589
Clustering	863	719	624	566	562	478	441	398	418	423
Association	240	213	176	165	395	301	127	101	115	114
Outlier Detection	75	38	26	24	14	12	22	9	14	13
Regression	917	798	684	601	519	401	394	350	380	378
Prediction	1239	923	766	630	588	486	449	407	477	507

Figure 4. Data mining functions

The major techniques examined include: neural networks, SVM, decision tree, nearest neighbor, genetic algorithm, visualization, association rules, Markov, swarm intelligence, regression models, hybrid and rough set. The text mining experiments showed that techniques like: neural networks, regression models, association rules, genetic algorithm and visualization are mostly used every year while techniques like: Markov, hybrid and rough set have the least popular and application. It is worth noting that neural networks are widely applied in various domains and have the highest number of applications per year. neural networks are a set of algorithms that investigate to find the relationship between the data set by performing a function similar to the human brain (Liao, Chu et al. 2012, Saad 2020), the use of this technique has grown exponentially due to its excellent performance year by year (Huang 2009).



	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011
Swarm Intelligence	165	131	121	98	89	77	78	54	74	80
Markov	67	43	43	42	58	40	31	26	29	24
Genetic Algorithm	197	181	182	159	172	169	139	143	172	200
Visualization	243	218	177	184	162	151	128	145	140	110
Hybrid	125	99	77	61	61	53	40	51	50	62
Associations Rules	240	213	176	165	395	301	127	101	115	114
Rough set	53	38	25	29	44	24	38	23	29	42
Nearest Neighbor	229	204	167	148	147	138	101	93	113	137
Decision Tree	198	157	139	105	80	81	73	65	66	88
Regression Model	917	798	684	601	519	402	394	350	380	378
SVM	117	127	98	99	101	106	85	79	85	140
Neural Network	1362	962	790	591	624	546	484	500	582	679

Figure 5. Data mining techniques

4.8. Relationships between IE sub-areas and DM functions and techniques

As shown in Figures 7 and 8, application rates of DM functions and techniques in each of IE sub-areas can be discerned through the strength of linkage between the clusters of IE sub-areas and DM functions and techniques in the network. on the top of the Figure 7 and 8, there are some nodes to which IE sub-areas belongs to, and it includes 10 main clusters that IE focuses on them and the bottom are nodes that belong to the DM functions and techniques, respectively which consist of 7 and 12 clusters, as mentioned in previous sections. each cluster in a category is considered as a node and the clusters related to one category are represented through the nodes of the same shape and color but their size is proportional to the cluster's share of the total data related to its category. Edges depict the connection relationship between two clusters, which the thickness of that is proportional to the frequency of co-occurrence between words related to 2 clusters. The relation between IE sub areas and a set of important DM functions is shown in detail in Figures 6 to 16. Based on the results in these Figures, we can infer that "prediction" function is the most popular and widely used in many IE sub-areas such as: operation research and optimization, financial, energy engineering, system simulation, quality and reliability engineering and manufacturing and scheduling. As shown in these figures, classification along with clustering can be considered one of the noteworthy functions in IE sub-areas, it also clears that " information system engineering and management" and "human factors and ergonomics " have the strongest association with the node of classification, and in the field of "operation research and optimization" and "quality and reliability engineering" after prediction has attracted the most attention. We can also find that in all fields, "sequential pattern" and "outlier detection" are relatively isolated. it can be said that all nodes related to the clusters

of IE sub-areas except "quality and reliability engineering" have the fewest association with the "outlier detection", so the link between them is eliminated by applying Threshold and can be ignored, which are shown in dash in figures (Figures 7, 8, 9, 10, 11, 12, 13, 14 15 and 16). Nevertheless, this function has received more attention in the field of "quality and reliability engineering" compared to other fields. The results also indicate that the node of association consistently has a moderate association with other nodes related to IE sub-areas, while plays an important role in the field of "human factors and ergonomics" and is one of the most widely used functions in this field. In Figure 7, the regression function appeared to have received the most attention in the field of "supply chain management and logistics", also has one of the strongest connections with the many other fields such as "financial", "energy engineering", " system simulation and stochastic process", "manufacturing " and " human factors and ergonomics".

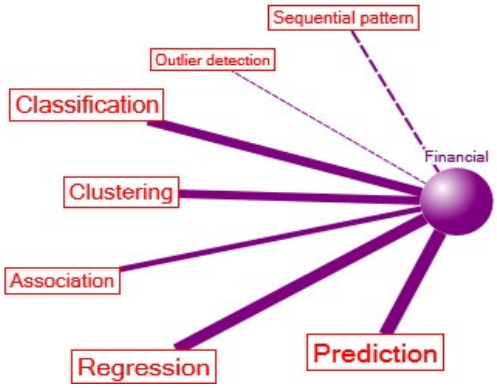


Figure 6. Relationship between Financial and DM functions

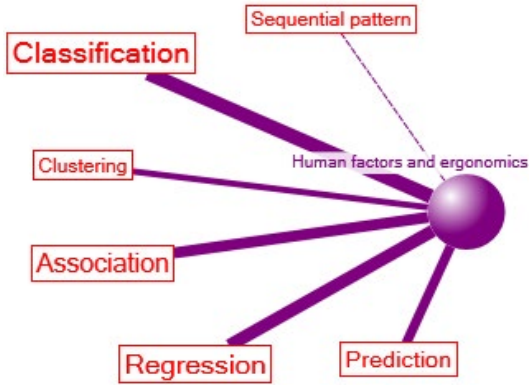


Figure 7. Relationship between Human factors and DM functions

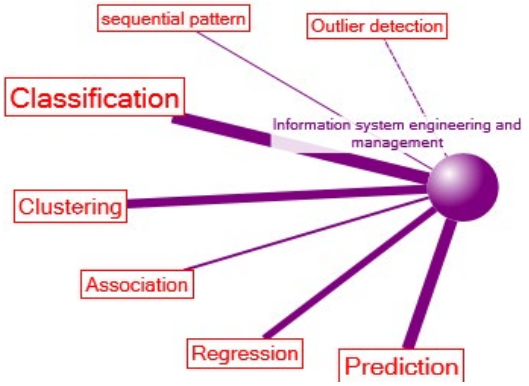


Figure 8. Relationship between information system and DM functions

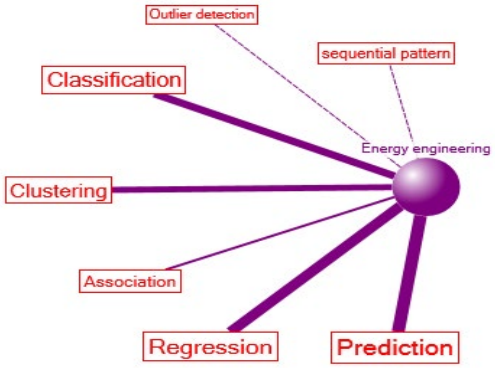


Figure 9. Relationship between Energy engineering and DM functions

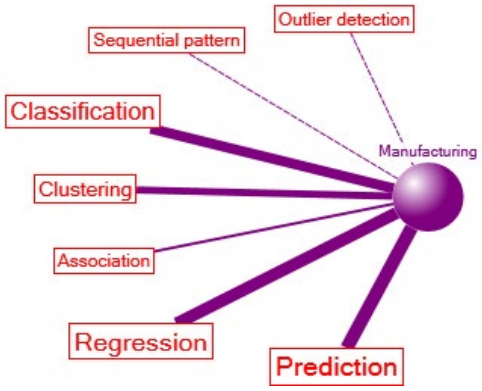


Figure 10. Relationship between Manufacturing and DM functions

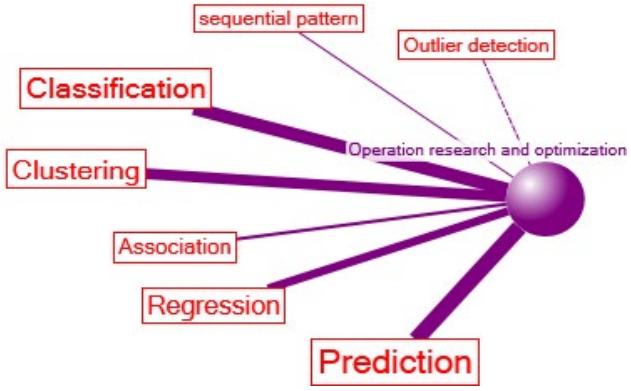


Figure 11. Relationship between Operation research and DM functions

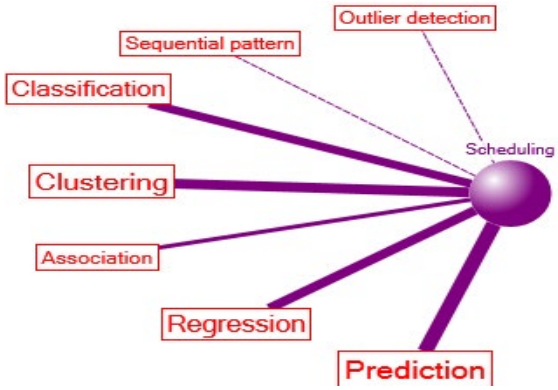


Figure 12. Relationship between Scheduling and DM functions

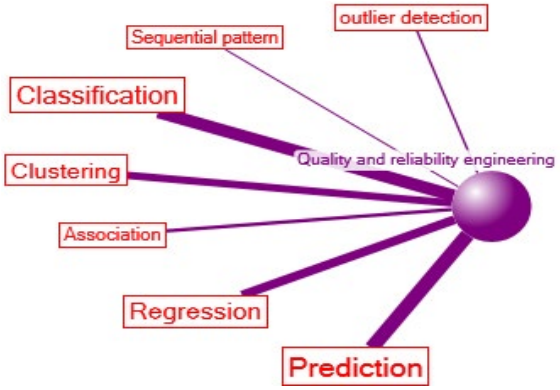


Figure 13. Relationship between Quality and reliability and DM functions

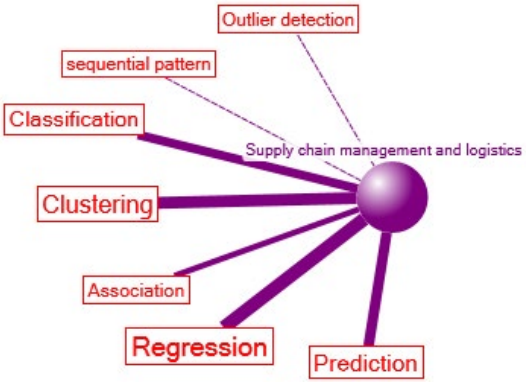


Figure14. Relationship between SCM and logistic and DM functions

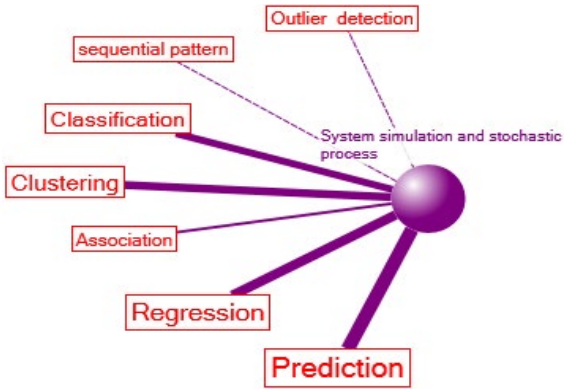


Figure15. Relationship between system simulation and DM functions

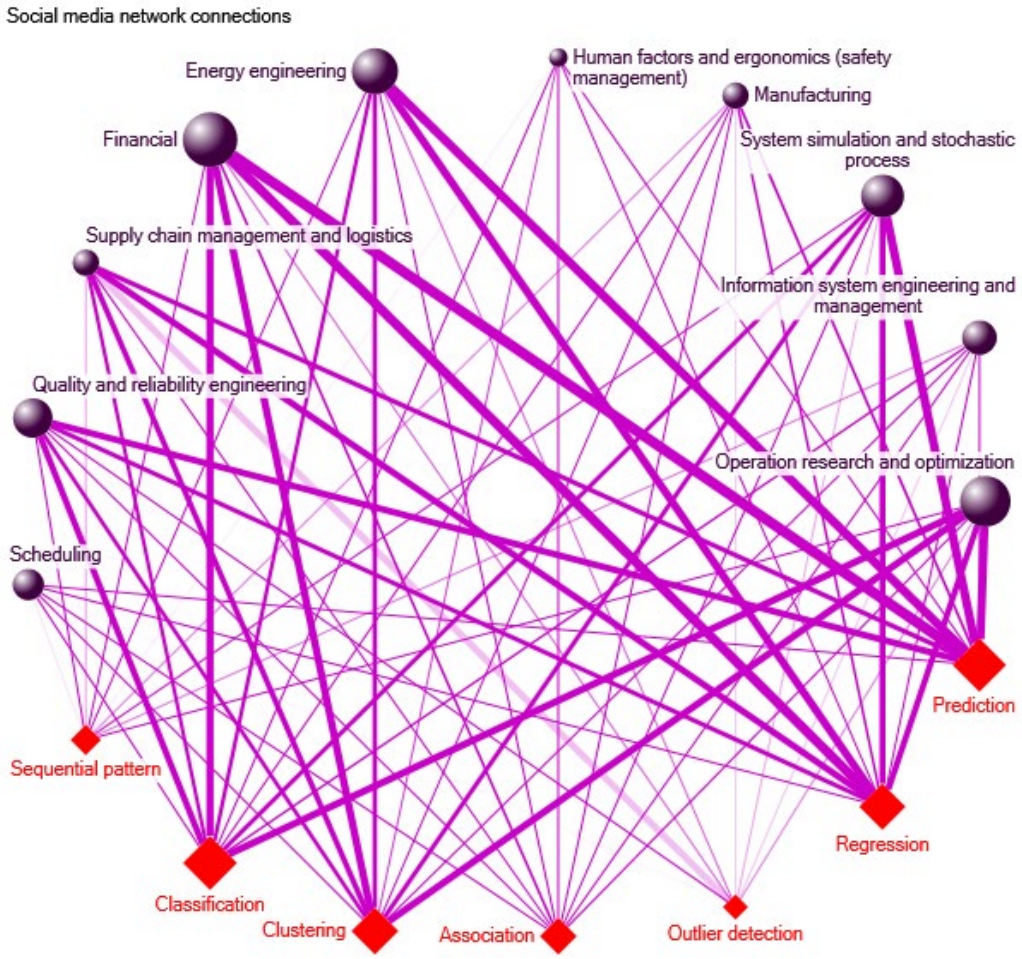


Figure 16. Relationship between IE sub-areas and DM functions

4.10. IE sub-areas and techniques

As can be seen in Figure 8, all of the nodes related to IE sub-areas have the strongest association with the nodes of "neural networks" and "regression models", it can be said that these two techniques are the most prominent techniques in all field of IE.

$E(sx, ty) = \{ pi \mid pi \in sx \wedge pi \in ty \wedge \mathrm{p}(sx, ty) \geq th2\}$

$R = \{(x, y) \mid x \in \{1,2, \dots, |Sub - areas|\}, \mathrm{p}(sx, fy) \geq th1 \}$

$y \in \{1,2, \dots, |Techniques|\}$

It is also remarkable that in the field of "human factors and ergonomics", "associations rules" technique is also one of the most significant techniques, conversely "rough set" and "Markov" have the fewest relationship with other nodes of IE sub-areas, but in the fields of "human factors and ergonomics" and "information system engineering and management", additional attention has been allotted to the "Markov" and "rough set", respectively. The results show that although little attention has been attached to "swarm intelligence" in all fields, this technique plays a crucial role in "operation research and optimization". As shown in Figure 17, "association rules" and "genetic algorithm" appeared to have received considerable attention in all clusters of IE but "genetic algorithm" has a little relationship with the cluster of "human factors and ergonomics", also a strong relationship between "financial", which is the largest sub-areas of IE, and "association rule" indicates that "association rules" play a significant role in this field.

in general, "decision tree" along with the "nearest neighbor" has a moderate association with other clusters of IE sub-areas, but it is notable that "decision tree" has received the least attention in "operation research and optimization" compared with other sub-areas. In Figure 8, little association between the nodes of "hybrid" and "swarm intelligence" and other clusters of IE sub-areas, reflects that these 2 techniques are relatively isolated and need more attention.

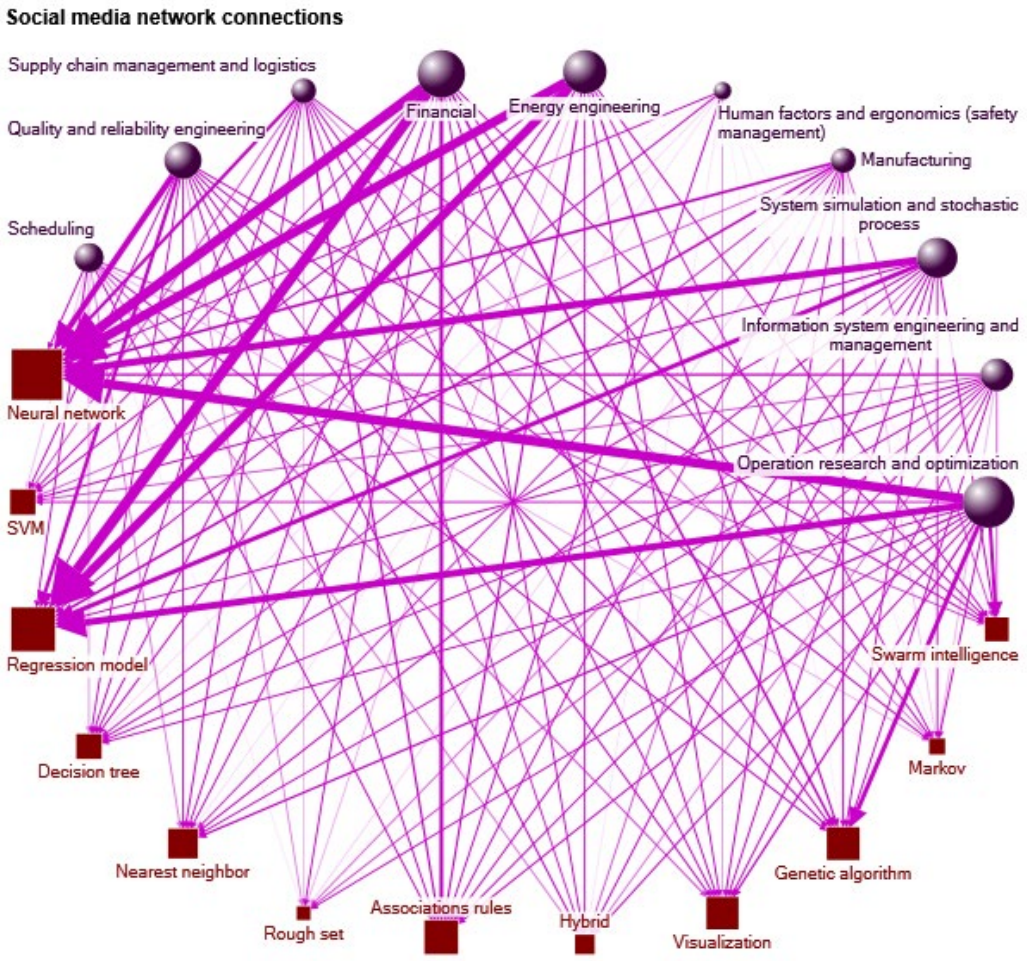


Figure 17. Relationship between IE sub-areas and DM techniques

5. Conclusion

In conclusion, this research provides a comprehensive evaluation of data mining and text mining models within the domain of industrial and systems engineering. The study employed a combination of bibliometric analysis and content analysis techniques, utilizing a dataset of over 20,000 publications from top journals in the field. The results highlight the significant growth and impact of data and text mining across various aspects of industrial engineering. The analysis reveals the leading journals, countries, and institutions contributing to the advancements in data mining applications in industrial engineering. Notably, the journal "Expert Systems with Applications" emerged as the most productive, with a concentration of influential publications in the United Kingdom. China stood out as a powerhouse in terms of both publication quantity and citation impact, with a remarkable focus on data mining research in industrial engineering. The study delves into specific sub-areas of industrial engineering, uncovering patterns of application for data mining functions and techniques. "Operation research and optimization" emerged as a dominant cluster, reflecting the growing importance of data-driven decision-making in problem-solving. The research also sheds light on areas that have received less attention, such as "human factors and ergonomics" and "information system engineering and management," indicating potential avenues for future exploration. Furthermore, the analysis of data mining functions and techniques reveals intriguing trends. Functions like classification, prediction, clustering, and regression were widely applied, with classification standing out as a powerful and pervasive tool. Among techniques, neural networks and regression models proved to be the most popular, while some techniques like Markov and hybrid approaches showed lower popularity. The study contributes valuable insights for researchers, practitioners, and policymakers in industrial engineering, highlighting the evolving landscape shaped by data and text mining methodologies. As data-driven decision-making becomes increasingly integral to industrial processes, this research underscores the transformative potential of data mining in shaping the future of industrial engineering professions.

References

Agrawal, R., Technologies for handling big data. *Handbook of Research on Big Data Clustering and Machine Learning*, IGI Global**:** 34-49, 2020.

Alkahtani, M., A. Choudhary, A. De and J. A. Harding, "A decision support system based on ontology and data mining to improve design using warranty data." *Computers & Industrial Engineering* **128**: 1027-1039, 2019.

Cancino, C. A., K. Amirbagheri, J. M. Merigó and Y. Dessouky, "A bibliometric analysis of supply chain analytical techniques published in Computers & Industrial Engineering." *Computers & Industrial Engineering* **137**: 106015, 2019.

Cattaneo, L., L. Fumagalli, M. Macchi and E. Negri, "Clarifying data analytics concepts for industrial engineering." *IFAC-PapersOnLine* **51**(11): 820-825, 2018.

Chen, T., "Estimating job cycle time in a wafer fabrication factory: A novel and effective approach based on post-classification." *Applied Soft Computing* **40**: 558-568, 2016.

Chen, T. and R. Romanowski, "Precise and accurate job cycle time forecasting in a wafer fabrication factory with a fuzzy data mining approach." *Mathematical Problems in Engineering* **2013, 2013**.

Choudhary, A. K., J. A. Harding and M. K. Tiwari, "Data mining in manufacturing: a review based on the kind of knowledge." *Journal of Intelligent Manufacturing* **20**(5): 501, 2009

Cordeiro, R. L. F., C. Faloutsos and C. T. Junior, *Data mining in large sets of complex data*, Springer Science & Business Media, 2013.

Demetgul, M., K. Yildiz, S. Taskin, I. Tansel and O. Yazicioglu, "Fault diagnosis on material handling system using feature selection and data mining techniques." *Measurement* **55**: 15-24, 2014.

Dolati Neghabadi, P., K. Evrard Samuel and M.-L. Espinouse, "Systematic literature review on city logistics: overview, classification and analysis." *International Journal of Production Research* **57**(3): 865-887, 2019.

Espadinha-Cruz, P., R. Godina and E. M. Rodrigues, "A review of data mining applications in semiconductor manufacturing." *Processes* **9**(2): 305, 2021.

Fahimnia, B., J. Sarkis and H. Davarzani, "Green supply chain management: A review and bibliometric analysis." *International Journal of Production Economics* **162**: 101-114, 2015.

Fernandes, E. C., B. Fitzgerald, L. Brown and M. Borsato, "Machine Learning and Process Mining applied to Process Optimization: Bibliometric and Systemic Analysis." *Procedia Manufacturing* **38**: 84-91, 2019.

Grabot, B., "Rule mining in maintenance: Analysing large knowledge bases." *Computers & Industrial Engineering* **139**: 105501, 2020.

Han, J., J. Pei and M. Kamber, *Data mining: concepts and techniques,* Elsevier, 2011.

He, B.-h. and G.-f. Song, *Knowledge management and data mining for supply chain risk management*. 2009 International Conference on Management and Service Science, IEEE, 2009.

Hirsch, V., P. Reimann, O. Kirn and B. Mitschang, "Analytical approach to support fault diagnosis and quality control in End-Of-Line testing." *Procedia CIRP* **72**: 1333-1338, 2018.

Hosseini, S., D. Ivanov and A. Dolgui, "Review of quantitative methods for supply chain resilience analysis." *Transportation Research Part E: Logistics and Transportation Review* **125**: 285-307, 2019.

Hu, J. and Y. Zhang, "Research patterns and trends of Recommendation System in China using co-word analysis." *Information processing & management* **51**(4): 329-339, 2015.

Hu, W., J. Dong, B.-g. Hwang, R. Ren and Z. Chen, "A scientometrics review on city logistics literature: Research trends, advanced theory and practice." *Sustainability* **11**(10): 2724, 2019.

Huang, Y., "Advances in artificial neural networks–methodological development and application." *Algorithms* **2**(3): 973-1007, 2009.

Jin, J. B., C. S. Leem and C. H. Lee, "Research issues and trends in industrial productivity over 44 years." *International journal of production research* **54**(5): 1273-1284, 2016.

Kao, C., "The authorship and internationality of Industrial Engineering journals." *Scientometrics* **81**(1): 123-136, 2009.

Kara, M. E., S. Ü. O. Fırat and A. Ghadge, "A data mining-based framework for supply chain risk management." *Computers & Industrial Engineering* **139**: 105570, 2020.

Knoll, D., G. Reinhart and M. Prüglmeier, "Enabling value stream mapping for internal logistics using multidimensional process mining." *Expert Systems with Applications* **124**: 130-142, 2019.

Köksal, G., İ. Batmaz and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry." *Expert systems with Applications* **38**(10): 13448-13467, 2011.

Kormann, B. and S. Altendorfer-Kaiser, *Influence of patterns and data-analytics on production logistics*. Digitalization in Supply Chain Management and Logistics: Smart and Digital Solutions for an Industry 4.0

Environment. Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 23, Berlin: epubli GmbH, 2017.
Kosky, P., G. Wise, R. Balmer and W. Keat, Exploring engineering, Elsevier, 2006.
Kretschmer, R., A. Pfouga, S. Rulhoff and J. Stjepandić, "Knowledge-based design for assembly in agile manufacturing by using Data Mining methods." *Advanced Engineering Informatics* **33**: 285-299, 2017.
Larose, D. T. and C. D. Larose, *Discovering knowledge in data: an introduction to data mining,* John Wiley & Sons, 2014.
Leung, X. Y., J. Sun and B. Bai, "Bibliometrics of social media research: A co-citation and co-word analysis." *International Journal of Hospitality Management* **66**: 35-45, 2017.
Lian, Y., G. Zhang, J. Lee and H. Huangm "Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles." *Accident Analysis & Prevention* **146**: 105711, 2020.
Liao, S.-H., P.-H. Chu and P.-Y. Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011." *Expert systems with applications* **39**(12): 11303-11311, 2012.
Merigó, J. M. and J.-B. Yang, "A bibliometric analysis of operations research and management science." *Omega* **73**: 37-48, 2017.
Nájera-Sánchez, J. J., "A systematic review of sustainable banking through a co-word analysis." *Sustainability* **12**(1): 278, 2020.
Ngai, E. W., Y. Hu, Y. H. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature." *Decision support systems* **50**(3): 559-569, 2011.
Qi, C.-c., "Big data management in the mining industry." *International Journal of Minerals, Metallurgy and Materials* **27**(2): 131-139, 2020.
Rabiei, M., S.-M. Hosseini-Motlagh, A. Haeri and B. Minaei Bidgoli, "Evolution of IT, Management and Industrial Engineering research: a topic model approach." *Scientia Iranica, 2020.*
Ramana, D. E., S. Sapthagiri and P. Srinivas, "Data Mining Approach for Quality Prediction and Improvement of Injection Molding Process through SANN, GCHAID and Association Rules." *International Journal of Mechanical Engineering and Technology* **7**(6), 2016.
Saad, H., "The application of data mining in the production processes." *arXiv preprint arXiv:2011.12348, 2020*.
Salvendy, G., *Handbook of industrial engineering: technology and operations management,* John Wiley & Sons, 2001.
Shah-Hosseini, A., System and method for supply chain data mining and analysis, Google Patents, 2013.
Shao, Y. E. and C.-D. Hou, "Fault identification in industrial processes using an integrated approach of neural network and analysis of variance." *Mathematical Problems in Engineering* **2013, 2013**.
Sharma, A. and P. K. Panigrahi, "A review of financial accounting fraud detection based on data mining techniques." *arXiv preprint arXiv:1309.3944, 2013.*
Shi, Y. and X. Liu, "Research on the literature of green building based on the Web of Science: a scientometric analysis in CiteSpace (2002–2018)." *Sustainability* **11**(13): 3716, 2019.
van Cruchten, R. R. and H. H. Weigand, *Process mining in logistics: The need for rule-based data abstraction.* 2018 12th International Conference on Research Challenges in Information Science (RCIS), IEEE, 2018.
Van Eck, N. J. and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping." *scientometrics* **84**(2): 523-538, 2010.
Van Eck, N. J. and L. Waltman, Visualizing bibliometric networks. *Measuring scholarly impact,* Springer**:** 285-320, 2014.
Van Nguyen, T., L. Zhou, A. Y. L. Chong, B. Li and X. Pu, "Predicting customer demand for remanufactured products: A data-mining approach." *European Journal of Operational Research* **281**(3): 543-558, 2020.
Vazan, P., D. Janikova, P. Tanuska, M. Kebisek and Z. Cervenanska, "Using data mining methods for manufacturing process control." *IFAC-PapersOnLine* **50**(1): 6178-6183, 2017.
Wang, J., J. Zhang and X. Wang, "A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system." *IEEE Transactions on Semiconductor Manufacturing* **31**(1): 173-182, 2018.
Wang, S., W. A. Chaovalitwongse and O. Seref, "Operations Research in Data Mining." Wiley *Encyclopedia of Operations Research and Management Science, 2010.*
Wei, Z., Y. Feng, Z. Hong, R. Qu and J. Tan, "Product quality improvement method in manufacturing process based on kernel optimisation algorithm." *International Journal of Production Research* **55**(19): 5597-5608, 2017.
Wu, M., K. Liu and H. Yang, "Supply chain production and delivery scheduling based on data mining." *Cluster Computing* **22**(4): 8541-8552, 2019.

Xu, S., X. Zhang, L. Feng and W. Yang, "Disruption risks in supply chain management: a literature review based on bibliometric analysis." *International Journal of Production Research* **58**(11): 3508-3526, 2020.

Yalcin, H., W. Shi and Z. Rahman, "A review and scientometric analysis of supply chain management (SCM)." *Operations and Supply Chain Management: An International Journal* **13**(2): 123-133, 2020.

Ying, H., L. Chen and X. Zhao, "Application of text mining in identifying the factors of supply chain financing risk management." *Industrial Management & Data Systems, 2020.*

Yue, D., X. Wu, Y. Wang, Y. Li and C.-H. Chu, *A review of data mining-based financial fraud detection research. 2007 International Conference on* Wireless Communications, Networking and Mobile Computing, Ieee, 2007.

Zandin, K. B., "Maynard's industrial engineering handbook.", 2001.