A New Framework for Research Supply and Demand Using Text Analytics

—MOHAMMAD RABIEI
Information Technology Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran 13157 73314, Iran

—SOMAYEH FATAHI
Information Technology Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran 13157 73314, Iran

(Corresponding author: Mohammad Rabiei.)

IEEE DOI 10.1109/EMR.2022.3179238

Abstract*—Numerous research studies are conducted annually, and many researchers look for them in the scientific databases. In order to manage the research ecosystem, it is necessary to have sufficient knowledge about the research supply and demand. This study proposes a novel expert-independent framework to investigate the research supply and demand among the huge number of studies. In this framework, the combination of quantitative, bibliographic, and content analysis methods is applied via text mining techniques. To evaluate the proposed framework, environmental data from the Iran scientific information database is used (Ganj)[1] as a rich database. It includes more than 513,000 supplied records and more than 78 million search queries as the research demands. To analyze the research supply and demand, the research topics are extracted via the topic modeling approach from both the supply and demand side. The results showed that the educational category of environmental studies is moved from medical science to engineering in the last decades. Moreover, the gap analysis between research supply and demand identified "Extraction" and "Tourism" as hot topics, "Education," "Management," and "Culture" as cold topics, "Soil" and "Energy" as silent topics and finally, "International Rights" and "River" as gap topics. The findings can support environmental policymakers and research managers in making decisions and identifying environmental research needs and priorities. The proposed framework for other research areas can be used to examine and balance research supply and demand.*

Key words: Environment, gap analysis, research ecosystem, research supply and demand, topic modeling

1. INTRODUCTION

THE terms "Research Gap,"[1] "Research Need," and "Research Priority" are used interchangeably in many documents dealing with the research management. In the first stage, these concepts and the differences between them must be explained.

The organizational knowledge gap is the distance between the knowledge an organization needs to perform its tasks and the knowledge that the organization has gathered through its various sources of knowledge [1]. Knowledge gaps can apply to existing and archived knowledge in an organization, society, or country and include deployable and efficient knowledge [2]. According to Robinson's definition, the research gap is an information constraint in the domain or subject that impedes the conclusion of a research question, and the research need is an information constraint that limits the policymakers to make a decision [3]. Therefore, from the user's perspective, the research gap could be defined as the difference between user's research demands and the

[1]This database is available at https://ganj.irandoc.ac.ir

research studies supplied their demands.

Many studies that have tried to identify the research needs and analyzed the queries of database users or library users have responded to the research demand and have failed to determine the research needs [1]. They have failed to understand research needs because they go beyond the research gaps. There are areas of knowledge that may be part of research needs but not part of research or information gaps because fewer people seek out those areas. There is not considerable demand for them.

Research prioritization is the selection of research projects based on a set of quantitative and qualitative criteria. This process is fundamentally difficult because the goals of the researchers and sponsorship organizations are different. Moreover, research needs and consequently research priorities change over time and give the rapid pace of knowledge production in various domains. This change occurs over a short period [4]. Now that the differences between the three concepts are clear, we can discuss the environmental research.

Environmental research has developed dramatically over the last decades [5]. A comprehensive overview of the evolution of the environmental footprint publications from 1992 to 2018 was presented, and the characterization of its publications was explained [6]. Although the necessity of maintaining the balance between environmental management and development was neglected during the modernization and the expansion of industrial activities in Iran [7], research trends show that this area has attracted the attention of researchers in the last two decades [8].The importance of analyzing the research supply and demand in the environmental domain and the large volume of research resources and complexities of this interdisciplinary field encouraged us to evaluate the proposed framework using the data of this research domain.

Research policymakers can identify the research needs after being aware of research gaps and research areas that do not have a significant demand in the research community. The environment is one of the significant global challenges, and accordingly, its department and related academic disciplines have been established in the universities and research institutes. For this reason, a considerable number of theses and dissertations are produced annually in this field [9]. Due to the necessity of presenting environmental attachments in many large projects, there are many various research needs in this field. The interdisciplinary nature of the environment [10] has made it difficult to delineate the boundaries required for information analysis and recognition of research supply and demand [5], [8]. Moreover, it is not possible to evaluate a large number of theses and dissertations related to this field and analyze the content of the considerable number of users' demand of this scope by common methods of research gap and priority identification such as using expert opinions.

In this research, a large volume of user queries for the environmental research studies will be compared with the related theses and dissertations presented in the Iran scientific information database (Ganj[2]). This study helps the environmental policymakers and research managers identify the research gaps and priorities and inform the researchers to avoid repetitive or unnecessary research studies. Besides, the proposed method can be applied to the other research domains for analyzing the research supply and demand and determining the research gaps and priorities.

The main research objectives are as follows:

1) proposing an expert-independent framework to analyze the research supply and demand;
2) determining the boundaries of a research domain (environment) without relying on expert knowledge;
3) machinery topics extraction of a research domain from both supply and demand side;
4) identifying the research gaps in the environmental domain.

To achieve these objectives, in the next section, the previous related studies in this domain are reviewed. Then, the methodology of the study is described. After that, in the Section 4.1, the bibliographic analysis of publications is presented. In Section 4.2, the steps of topic modeling are explained, and gap analysis between supply and demand of environmental theses is illustrated. Finally, the main conclusions are summarized, and some suggestions for future studies are outlined.

2. Literature Review

Recently, numerous studies have been carried out to analyze the research trends, identify the research supply and demand, and determine the research priorities [11]–[13]. Understanding the research evolution of scientific topics is recently done in many research domains via text mining techniques [5], [11], [12], [14]–[16]. These techniques are varied from word embedding [17], [18] to topic extraction approaches [19], [20], and the application of them are varied from sentiment analysis [18], [21], [22] and opinion mining [23] to behavior analysis [24], [25] and strategic management [26], [27]. As a sample, the big data analytics

[2]https://ganj.irandoc.ac.ir

adoption research published between 2013 and 2019 was analyzed via bibliometric methods. In this study, the models and frameworks applied by organizations to adopt big data analysis were extracted based on a quantitative approach that is mostly used in the bibliometric analysis [11]. Opinion mining and sentiment analysis studies conducted over a period of 16 years (2000–2015) were extracted from the Web of Science (WoS) database and sicentometric mapping. An analysis of them is presented in [28].

As with other domains, some scientometric studies addressed the characteristics of the research in the environmental areas. The challenges of integrated research on the environmental changes were described by content analysis of a questionnaire distributed among the Ph.D. students of the environmental fields [29]. A scientometric review of global Building Information Modeling (BIM) research, through coauthor analysis, co-word analysis, and co-citation analysis, was presented by Xianbo Zhao during 2005–2016. He collected 614 bibliographic records from the WoS database to explore the status and trends of global BIM research. The findings were identified as the top three most productive authors in the field and the top three most co-citations. Also, the most significant development in BIM research was determined in the USA, South Korea, and China. Besides, the results showed that BIM research had mainly focused on the subject categories of engineering, civil engineering, and construction and building technology. “Visualization” and “industry foundation classes” keywords were mainly used recently. Furthermore, the hot topics of BIM research were extracted [30]. Around 1900 articles of the green supply chain from the WoS database were collected and analyzed, and the bibliographical analysis of the data was presented [14]. Various strategies of protecting the coral reef ecosystems were investigated through the systematic literature review of articles published between 1990 and 2016 that addressed coral reef management in the context of global environmental change [31]. A novel method was proposed for managers and policymakers as research demand and supply monitoring to help them in the decision-making process of determining the research priorities for the environmental studies [9]. In this study, the research demand was captured from the queries they were searched in the database, and the supply side was obtained from the number of answers returned from the database search engine for each query. Also, the scientometric analysis was used to find out the most productive authors and introduce the core journals of environmental management research based on data collected from WoS between 1989 and 2014 [5]. A systematic review and metaanalysis of lead (Pb) concentrations in environmental media in two decades of environmental studies of the United States were conducted. More than 14 000 articles were collected from WoS and ProQuest after preprocessing [15]. International scientific cooperation of Iranian researchers in the 10-year-period (2008–2017) research of environmental health engineering studies was evaluated. The results showed that about 15% of articles were written with the participation of researchers from other countries. India, Turkey, and Malaysia had the highest level of cooperation with the Iranian researchers [32].

Regardless of the field under study, these studies relied mainly on evaluating bibliographic information related to documentation for the analysis of science trends. Previous studies used primarily qualitative methods and expert opinions. They accepted its limitations, such as the small number of experts involved in the research, bias, and the high impact of the researcher’s opinion in interviewing and summarizing the results. Also, these approaches do not apply to analyzing a considerable number of studies and are very time-consuming. On the other hand, previous studies mainly focused on existing research, and future research was neglected. Regardless of technique and method of work, identifying research trends just based on existing research can show the research studies have been done. Still, it does not specify what areas have been left unaddressed or neglected. In other words, only areas that contain a certain amount of research would be identified. Therefore, it would fail to identify new and lean areas of research. While analyzing searches of the entire research community that include new research trends may also discover new and hidden research areas. Finally, there is no considerable research on analyzing research supply and demand in the environmental domain.

3. Methodology

This study uses a combination of content analysis, supply-oriented, and demand-driven approaches. It also uses a hybrid method of quantitative, bibliographic, content analysis, and text mining techniques. Using hybrid or ensemble methods [19], [33]–[36] are common in datamining and text mining approaches to produce improved results.

This research identifies research demands based on users’ queries in the thesis and dissertation repository, and in the same way, the research supply is obtained by analyzing the existing documents in the repository. Also, the environmental studies are applied as the case study, and the proposed framework is applicable for other research disciplines. Figure 1

presents the research framework. As it illustrates, the different steps of the research proceed in two distinct and related paths, i.e., the supply side and demand side of the research. The supplied research works are theses and dissertations presented to researchers, and the demanded research works are queries searched by the researchers in the database's search engine. Each part of this framework will be explained in the following sections.

3.1. Data Collection One of the most important institutions for organizing scientific resources in Iran is the Iranian Research Institute for Information Science and Technology (IranDoc), established under the Ministry of Science, Research, and Technology in 1968. This institute is the owner of the Ganj database that is the official repository for collecting, organizing, and presenting theses and dissertations (we called both as "thesis" in this study) in Iran. Iran's scientific information database (Ganj) has more than 1.2 million scientific records, of which more than 700 000 records of them are thesis. The postgraduate students have to submit their thesis in Ganj after their university has approved it. In addition, more than 130 000 searches are done daily on this database.

There are many studies done on the log file of users' search and assessment of user satisfaction of Ganj in recent years [37], [38]. It is a valuable source among researchers in Iran. This resource was used in this study to determine supply and demand in Iranian research. It should be mentioned that the contents of all documents and queries applied as data sources in this study are in Persian.[3]

For the supply side, all 513 927 theses from all subjects were collected from 1934 to 2019. In addition, more than 78 million records of user searches between 2013 and 2019 are stored as a log file for the demand side. It should be mentioned that, the "environment" is a multidisciplinary research domain, and there is not any specific identifier to collect the theses and queries that are related to this domain. Therefore, as shown in Figure 1, a machine learning approach was utilized to extract the environmental documents and search queries. This approach needs a core dataset in its learning process. The core dataset is obtained by searching for terms including the translation of "environment" in Persian[4] among the "title," "keyword," "abstract," "subject," or "organization" fields of records. In this study, we used a core dataset that has 2244 records. This approach is explained with more details in the next section.

3.2. Identifying Environmental Contents The process of identifying the environmental-related texts has been carried out via a classification approach known as one-class classification (OCC) applied in the support vector machine (SVM) and called OSVM [39]. Using OSVM to classify the texts works more efficiently than other classification methods because of the high dimensionality of text vectors [8], [39]. For this purpose, the core dataset of this domain was created for use in the machine learning process. Natural language processing techniques were then used in preprocessing and production of the feature vector matrix, and then the test dataset was created to evaluate and improve the model. Finally, the environmental dataset was extracted by using this model from the searches as well as the research resources available in the database [8]. After applying OSVM on more than 500 000 records of research studies in the Ganj database and in the same way on the more than 70 million records of the user queries, 106 709 theses were identified as the supply side of environmental-related documents, and 1192 153 queries were extracted for environmental research demand. Tables 1 and 2 present some sample environmental theses and queries, respectively. The original terms in these tables are in Persian, but to make them more understandable, they were translated into English.

In Table 2, "Answers" is the number of records returned by the search engine for each search. Rabiei *et al.* [9] used this field as the presenter of existing resources or research supplies. In other words, it is assumed that the search engine response is reliable for the number of existing resources. As we know, this is not always a true assumption because it depends on various factors, such as search engine settings, search algorithms, index types of information, and date of search. Therefore, this field is not applicable to indicate the number of existing resources, and the supply side should be analyzed independently.

4. Results and Discussion

4.1. Bibliographic Analysis The bibliographic analysis of supply and demand research studies of the environmental domain in the Ganj database is illustrated based on extracted related documents and queries as follows.

4.1.1. Trend Analysis of the Environmental Theses:
The trend of environmental theses is depicted in Figure 2.

As shown in Figure 2, until 1990, less than 100 theses were produced per year, which are related to the environment. They are mainly in the fields of pharmacy, health, veterinary, and civil engineering. This amount

[3]In this article, all the Persian text in the original research is translated into English.

[4]The translation of "environment" in Persian is "زیست محیط"

grew six times between 1991 and 2000 and produced over 600 theses annually on average. This growth increased steeply over the years 2007–2014 and remained steady at an average of 12 000 related documents per year in the last five years. Analysis of the environmental footprint studies indicated the same trend and showed the increasing number of publications between 2007 and 2016 [6].

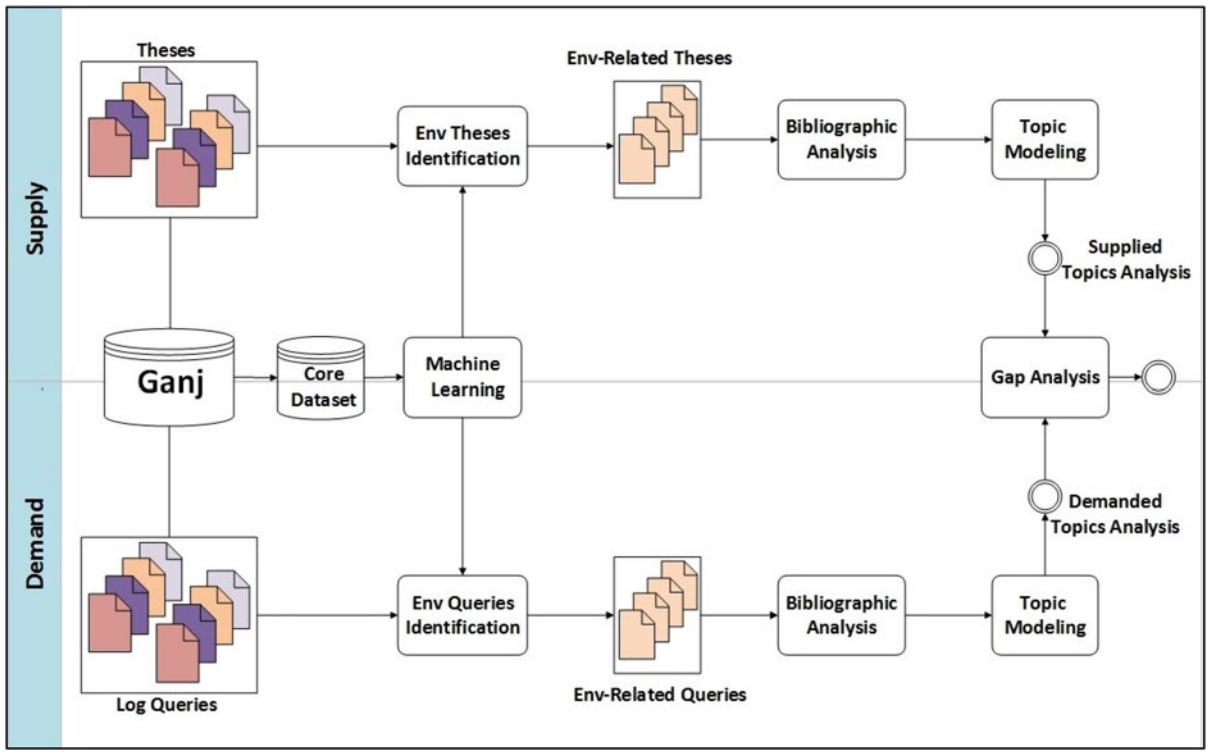

Figure 1. Research framework.

Table 1. Sample These Identified as Environmental- Related Supplied Theses.

Title	Subject area	Date	Degree	Keywords
Economic impacts of forest in Iran	Human sciences	1971	Master	Forest products, national income, forest conservation, and forest
Heavy metals contamination of Tehran groundwater	Medical Sciences	1973	Doctorate	Chrome, water pollution, copper, iron, heavy metals, tehran, groundwater, industrial pollution, and industrial wastewater
Investigation of water resources of Maragheh	Engineering	1970	Master	Ghale chai river, aqueduct, water, geology, groundwater, surface, water resources, and climatology
The effect of Tehran metropolitan development on the development of worn building	Art	2010	Master	Worn building, Tehran, metropolis, and development
Demand of fish species in Mazandaran province	Agriculture	2011	Master	Fish, mazandaran, and demand
The effect of cosmic rays on the Earth's atmosphere	Basic science	2015	Master	Observatory, detector, positrons, mesons, ozone, Earth's atmosphere, and cosmic rays

Table 2. Sample Queries Identified as Environmental-Related Demands.

Query	Answers	Date
("Desertification," "Khorasan," and "Wind erosion")	2	2013/02/16
Dust phenomenon	14	2013/02/17
Industrial wastewater	685	2013/02/17
Erosion and sedimentation in tea syrup	0	2013/02/19
Water demand for wheat production in Ramshir	0	2013/02/20
("Suspended Sediment" and "Neural Network")	27	2013/02/21

4.1.2. *Trend Analysis of the Environmental-Related Demands:*

Figure 3 shows the trend of the user's demands for the environmental-related theses in the Iran scientific information database (Ganj) between 2013 and 2019.

Except for the first year and the last year of this period, which contains the information of only a few months, the key point is the search leap in the years 2015 and 2016. This change is due to the change in the policy of IranDoc on how users access the database, so users had not needed to sign up and log in since October 2014, and a new user friendly version of Ganj was launched in 2015.

4.1.3. *Broad Subject Area of Theses:*

The postgraduation consists of six broad subject areas, which include engineering, human science, basic science, medical science, agriculture, and art. The participation of different broad subject areas in the environmental resources supplied in Iran has varied over time. Figure 4 illustrates the portion of these areas in the environmental research in Iran over a given period.

Changing the educational category of environmental theses in Iran from medical science to engineering is the most important fact that can be seen in Figure 4. The portion of the engineering subject area ranged from 4% to 7% before 1980, but in recent decades, the engineering approach in environmental theses has expanded, and nowadays, it exceeds all subject areas (27%). The disciplines of civil engineering, chemical engineering, mechanical engineering, industrial engineering, etc., have had a considerable role in this domain in recent decades. In contrast, the medical sciences, which covered nearly 40% of environmental theses before 1980, has paid less attention to this domain in recent years and currently covers less than 5% of environmental theses in Iran.

4.1.4. *Educational Degree of Studies:*

For the resources identified as environmental-related studies, 89% and 11% belong to the master's degree, the doctoral degree, respectively. The educational degree of these documents is presented separately for each broad subject area in Figure 5. More than 60% of the environmental research extracted from Ganj in medical science has been completed by Ph.D. students.

4.1.5. *Frequent Terms:*

To extract the frequent terms of the supply side and demand side of the environmental theses, a bag of words (BoW) for each record is created. The BoW is formed based on the *n*-gram concept. An *n*-gram is a set of *n*-consecutive elements in a text or speech that can be letters, syllables, words, etc.

$$n\_\text{gram}_i\,(\text{St},n) = \text{word}_i\ \text{word}_{i+1}\ldots\text{word}_{i+n-1}. \quad (1)$$

In the abovementioned equation, i is the index of the start position in a statement (St) to extract the n consecutive words. In this study, the BoW for each demand is created by extracting n-grams of the query string where $n = 1, 2, 3, 4$. Some of these n-grams are stop words and have not valuable means in most types of texts (for example, a, an, the, of, in the . . .). To avoid the extracting of these meaningless or unimportant n-grams, the acceptable n-gram is defined as follows:

$$\text{Acptble\_NG}(\text{St},n) = \begin{cases} \bigcup_{i=1}^{k} n\_\text{gram}_i(\text{St},1) \mid n\_\text{gram}_i(\text{St},1) \notin \text{StpW} \\ \bigcup_{i=1}^{l} n\_\text{gram}_i(\text{St},n) \mid \\ \{\text{word}_i(n\_\text{gram}_i(\text{St},n)), \text{word}_n(n\_\text{gram}_i(\text{St},n))\} \cap \text{StpW} = \emptyset. \end{cases}$$

In the abovementioned equation, StpW is a set of stop words. Consequently, some terms, such as "research and development," are acceptable, but "and," "research and," and "and development" are not

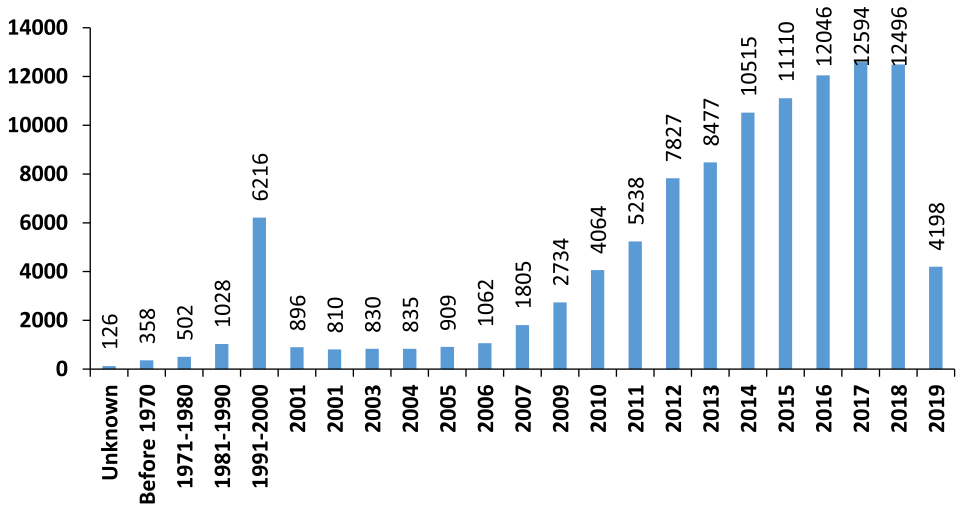

Figure 2. Trend of the environmental theses in Iran.

allowed us to be in the BoW. In the same way, the BoW is created by the sum of the keywords and extracted n-grams of the Title for each supplied document

$$\text{BoW}(\text{Doc}) = \left(\bigcup_{n=1}^{4} \text{Acptble\_NG}(\text{Title}(\text{Doc}), n)\right) \cup \text{Keywords}(\text{Doc}). \quad (3)$$

After creating the BoW for both supplied documents and demanded queries, the most frequent terms are identified for them. Tables 3 and 4 present the most 20 frequent terms for environmental theses and search queries in Iran.

4.2. Topic Modeling To demonstrate the environmental research supply and demand, it is necessary to identify and extract the topics of this domain. Therefore, we need text classification methods that have high dimensional feature space [36].Topic modeling is one of the most effective methods of text classification that aims to group a set of different texts into meaningful subcategories called the "Topic." Topic modeling does this process with minimum human dependency, and this has made it attractive and prevalent in the analysis of huge volumes of texts. In this approach, instead of using predefined titles, topics are extracted only by specifying the number of them, and the degree of relevance of each document to each topic is also determined [40]. The latent Dirichlet allocation (LDA) algorithm, which is the most common form of topic modeling, is based on the assumption that the words in a text are semantically related and the meaning contained in a document of a topic is derived from the set of words in the document [41]. Based on this assumption, if the corpus contains n documents and each document is composed of m-values to indicate the weight or importance of m different terms for that document, the LDA can then identify the k topics for the corpus as follows:

Since the number of topics k is much smaller than the number of terms m, using a smaller matrix $(k * n)$ to describe the corpus is more understandable and interpretable. In other words, whatever the meaning or place of the text is, the syntax of the words is analyzed and, the documents are considered a set of words [42].

The important parameter of the LDA is the number of topics k. There are various methods for determining the number of topics [43]. The advantage of choosing a large number of topics is that all the subject areas will be covered. On the other hand, the advantage of the limited number of topics is that the experts and policymakers will better understand the topics, and it will be easier to interpret and analyze [27]. One of the most common criteria to determine the

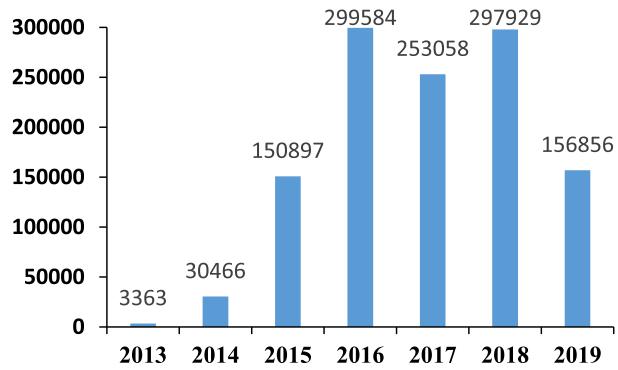

Figure 3. Trend of the user's demands for the environmental theses.

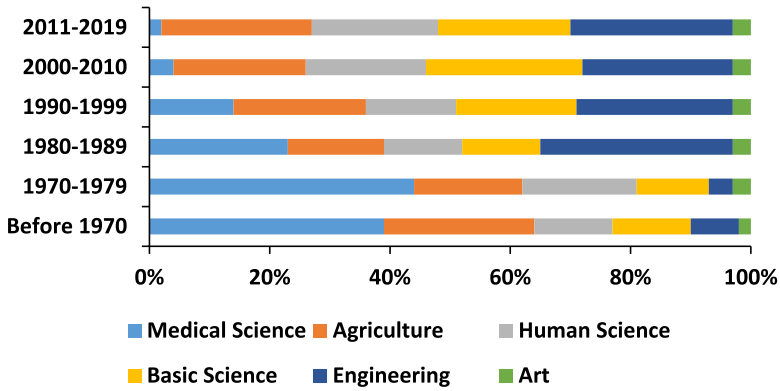

Figure 4. Contribution of educational category in environmental research in each time period.

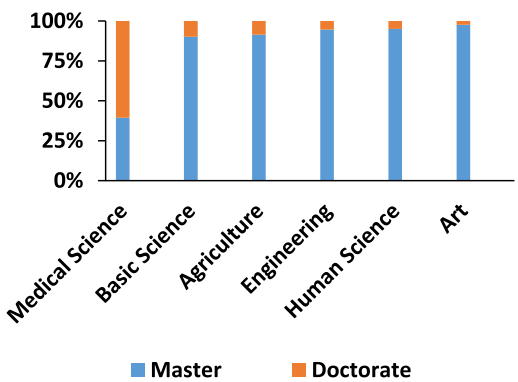

Figure 5. Educational degree of environmental studies.

$$\begin{array}{c} \\ \text{Term}_1 \\ \text{Term}_2 \\ \vdots \\ \text{Term}_m \end{array} \begin{array}{c} \begin{array}{cccc} \text{Doc}_1 & \text{Doc}_2 & \dots & \text{Doc}_n \end{array} \\ \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \dots & w_{m,n} \end{bmatrix} \end{array} \xrightarrow{\text{LDA(Terms,Docs, } k)} \begin{array}{c} \\ \text{Term}_1 \\ \text{Term}_2 \\ \vdots \\ \text{Term}_m \end{array} \begin{array}{c} \begin{array}{cccc} \text{Topic}_1 & \text{Topic}_2 & \dots & \text{Topic}_k \end{array} \\ \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,k} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{m,1} & \beta_{m,2} & \dots & \beta_{m,k} \end{bmatrix} \end{array} \begin{array}{c} \\ \text{Topic}_1 \\ \text{Topic}_2 \\ \vdots \\ \text{Topic}_k \end{array} \begin{array}{c} \begin{array}{cccc} \text{Doc}_1 & \text{Doc}_2 & \dots & \text{Doc}_n \end{array} \\ \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,n} \\ \theta_{2,1} & \theta_{2,2} & \dots & \theta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{k,1} & \theta_{k,2} & \dots & \theta_{k,n} \end{bmatrix} \end{array} \tag{4}$$

k is perplexity. It is a measurement criterion in statistical models used to check the appropriateness of a probability distribution or predict a probabilistic model of samples. The lower amount of perplexity shows that the model is better and has greater versatility [42]. Running the model with different values of k (2–200) showed that the least amount of perplexity was for $k = 20$.

The output of the LDA is determining the explanation probability of a document by a topic. The other important parameter is setting an explanation threshold to accept a topic as an explanatory document. If a low value is selected for this threshold, the topic and document will be linked together in an inferior relationship (low quality of explanation). On the other hand, choosing the high threshold leads to the condition that lots of documents will not be described by any topic (low corpus coverage) [44]. Our evaluation shows that in the threshold of 0.3, the average corpus coverage and quality of explanation are maximized.

After topic modeling, 20 topics are extracted, and the most frequent terms of each topic are identified. Based on these frequent terms, a label is assigned for each topic. Table 5 lists the frequent terms and assigned labels for each topic.

Table 3. Top 20 Frequent Terms in the Environmental Theses (Supply Side).

Rank	Term	Frequency	Rank	Term	Frequency
1	Iran	915	11	Earthquake	676
2	Environment	908	12	Adsorption	605
3	Wheat	900	13	Salinity	596
4	Stability	898	14	Social capital	592
5	Soil	847	15	Temperature	591
6	Tourism	836	16	Nano composite	552
7	Quality of life	828	17	Corn	533
8	Sustainable development	812	18	Nickel	527
9	Heavy metals	781	19	Nanoparticles	511
10	Geographic Information System	779	20	Drought stress	508

Table 4. Top 20 Frequent Terms of User Environmental Queries (Demand Side).

Rank	Term	Frequency	Rank	Term	Frequency
1	Absorption	121 898	11	Variety	48 240
2	Earthquake	79 445	12	Management	37 679
3	Extraction	78 256	13	Deposition	37 658
4	River	69 632	14	Sources	34 747
5	Health	6652	15	Waste	34 485
6	Wheat	66 325	16	Wastewater	31 305
7	Water	56 569	17	State	31 050
8	Nanoparticles	55 848	18	Jungle	31 008
9	Protection	50 883	19	Mazandaran	30 362
10	Wastewater	50 180	20	Absorbent	29 579

4.2.1. Supplied Topics of the Environmental Thesis:

After topic extraction, the number of documents, which are explained by the topic with more than an explanation threshold is counted. It is clear that with the given threshold (th $= 0.3$), research can be described by more than one topic (maximum by three). Moreover, some documents may not be described by any topic. Figure 6 shows the number of documents described by each topic.

It is important to note that Figure 6 is relevant to all known sources related to this area and has not necessarily been performed by environmental researchers or students of this field. In other words, it depicts the distribution of all the research related to environmental science from all disciplines. Most research in the environment field involves soil, tourism, and extraction. Education, management, and culture are the topics with minimum documents in this domain.

4.2.2. *Demanded Topics of the Environmental Thesis:*

Twenty environmental topics were identified among users' searches as the demand side of environmental theses. Meanwhile, education, with just over 1000 related queries, was the smallest environmental topic, and "Extraction" with more than 145 000 related searches, had the highest demand. Figure 7 shows the number of queries in each topic from the environmental search.

4.2.3. *Gap Analysis between Supply and Demand of Environmental Theses:*

In order to analyze the gap between the supply and demand of the extracted topics, it is necessary to normalize the number of items in the

Table 5. Frequent Terms of the Extracted Topics.

Topic	Label	Frequent terms
T01	Soil	Soil, plumb, heavy metals, arsenic, adsorption, chromium, soil pollution, cadmium, heavy metals, iron, and nickel
T02	Extraction	Extraction, micro extraction, solid phase extraction, preconcentration, isolation, gas chromatography, nanoparticles, and experimental design
T03	Environmental health	Environmental health, development, health, recycling, environmental assessment, safety, pollution, waste, sustainable development, and population
T04	Energy	Energy, energy consumption, oil, power plant, development, solar energy, Persian gulf, economic growth, renewable energy, green supply chain, and sustainability
T05	Law	Civil liability, damages, rights, environmental rights, legal system, participation, environmental behavior, natural resources, and environmental ethics
T06	Wastewater	Treatment, sewage, effluent, wastewater, green space, nitrate, wastewater treatment, urban environment, green chemistry, pollutants, chromium, and ammonia
T07	Air pollution	Carbon dioxide, pollution, air pollution, Kuznets environmental curve, carbon dioxide emissions, pollutants, polyacrylamide, aldehydes, and environmentally friendly
T08	Tourism	Tourism, development, ecotourism, ecology, industry, rural tourism, urban development, land use, and ecosystem
T09	Sustainable development	Sustainability, environmental sustainability, development, sports, environmental factors, environmental impacts, population, Mazandaran
T10	Management	Environmental management, sustainable development, green supply chain management, protection, participation, evaluation, and energy consumption
T11	Garbage	Garbage, environmental pollution, leachate, environmental pollution, landfill, biogas, pollution, pollution, waste disposal, soil, pollution, and geochemistry
T12	Waste	Waste, pollutant, nanoparticles, social responsibility, cement, waste management, nanocomposite, degradation, compressive strength, and recycling
T13	International rights	Environmental protection, legal system, international law, Caspian sea, rural development, sustainable development, biodiversity, and south pars
T14	River	River, environmental flow, fish, water quality, water resources, Urmia lake, ecological, sediment, salinity, migration, and Karun river
T15	Natural resources	Nature, degradation, natural resources, environmental degradation, conservation, degradation model, green architecture, and tribal settlement
T16	Vegetation	Dust, geology, vegetation, sediments, pollutants, environmental geology, sediment, soil erosion, floods, desertification, plant, and Ahwaz
T17	Industry	Development, industry, environmental performance, iron, economic growth, financial development, heavy metals, energy, and commercial liberalization
T18	Culture	Environmental behavior, lifestyle, environmental culture, conservation, environmental attitude, intellectual capital, and environmental educa1tion
T19	Water resources	Water resources, groundwater, aquifer, water quality, drinking water, wetland, water resource management, and surface water
T20	Education	Environmental literacy, risk assessment, environmental awareness, environmental education, safety, tourism, ecology, and conservation

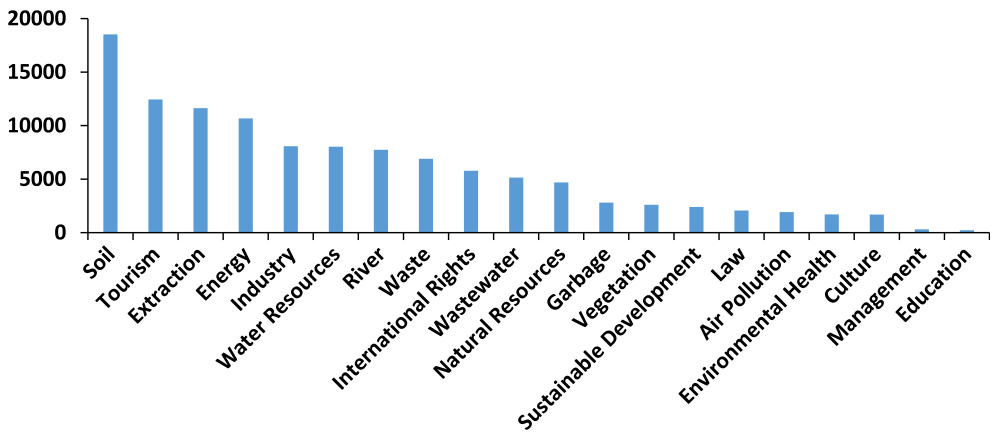

Figure 6. Number of documents in each topic of the environmental theses.

topics as the first step. To reach this goal, the number of records a topic is divided by the most considerable topic quantity on each side. As a result, all values are scaled between 0 and 1 as follows:

$$N\_\text{Supply}(\text{Topic}_i) = \frac{\text{Supply}(\text{Topic}_i)}{\text{Max}(\text{Supply}(\text{Topic}))} \quad (5)$$
$$N\_\text{Demand}(\text{Topic}_i) = \frac{\text{Demand}(\text{Topic}_i)}{\text{Max}(\text{Demand}(\text{Topic}))}$$

In the next step, topics are mapped based on their position in the two axes of supply and demand of the environmental theses. By dividing the horizontal axis of supply into the "high supply" and "low supply" segments, and by dividing the vertical axis into two "high demand" and "low demand" segments, a matrix that contains the four parts will be obtained. The value of 0.5 is defined as the boundary between high and low for both the supply and demand axes. Figure 8 depicts the location of each topic in this matrix.

As shown in Figure 8, the top right corner of this chart contains highly researched topics, and many researchers search for and explore their research needs. Therefore, the "extraction" and "tourism" topics in this part are considered as "hot topics." The hot topics are assumed as topics that are highly demanded and also highly be generated by researchers. In other words, research on these topics is still used and expanded. In the previous research in

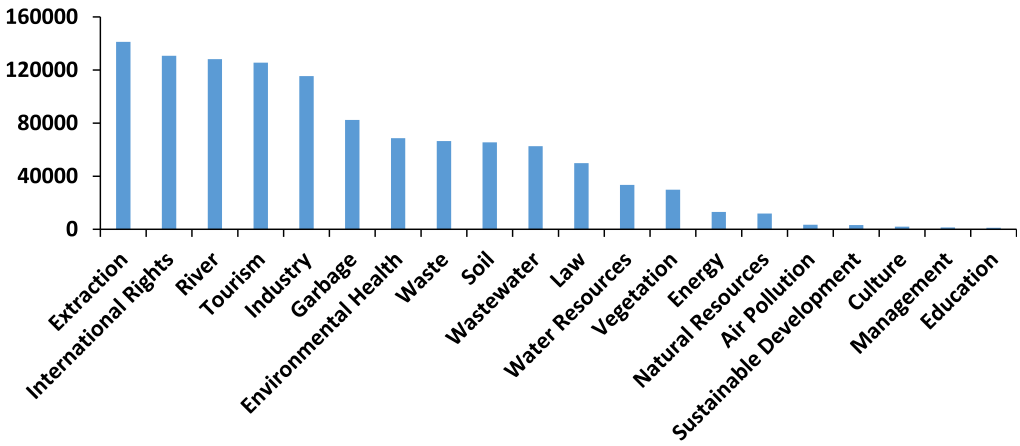

Figure 7. Number of queries in each topic of the environmental search.

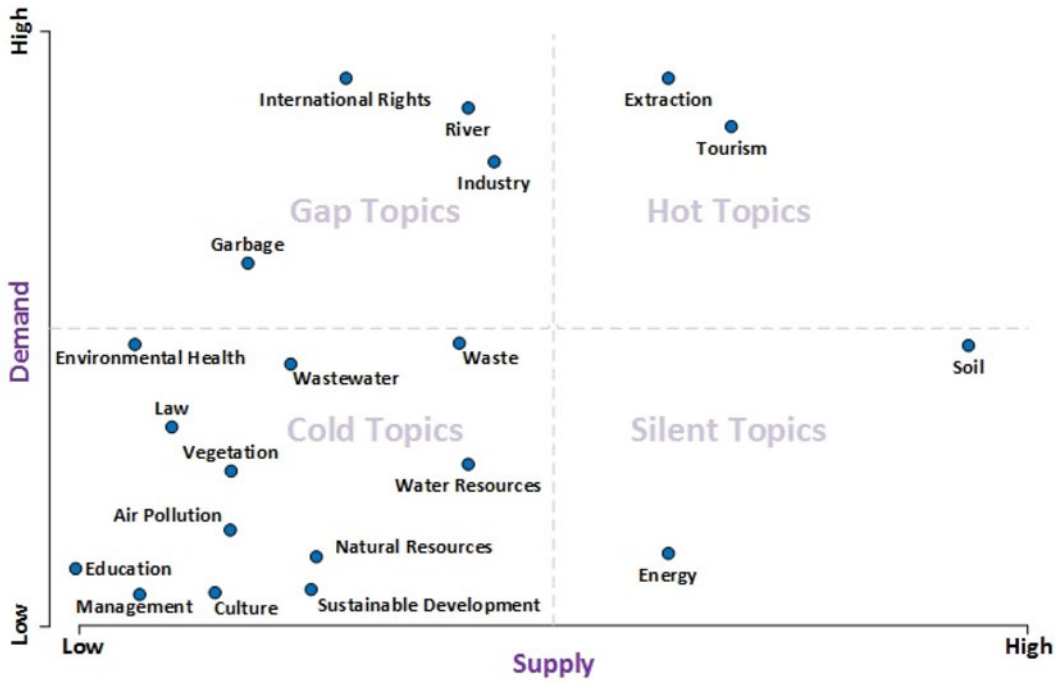

Figure 8. These supply and demand positions of environmental topics.

this field, the topic of "materials" that is very similar to the topics we called "extraction" was identified as a topic with high search rates and high response rates [9].

We called the bottom right corner "silent topics," which contains the topics with a significant number of sources, but the users are currently uninterested in using them. It seems that some of the older research works and partially saturated research areas fall under this category. The "soil" topic is very close to the hot topic zone and can be categorized as a hot topic with some ease, but the "energy" topic can be classified as a silent topic. Lots of energy-related studies are done annually in Iran under the support of the ministry of petroleum. The soil and jungle were known as the low search and high response rate topics in the previous study [9].

The lower left part contains topics that are poorly researched and are not welcomed by researchers. This area, which is precisely the opposite of the hot topics, is labeled "cold topic." Due to the small number of research studies in this area and the lack of attention of environmental researchers to the topics of this area, it can be predicted that if research policymakers do not have the necessary solutions, these topics will be neglected in the future, and they will lose their importance as independent topics. Prior research on prioritizing environmental research topics through the analysis of user search logs in Iran was identified the areas of "law," "management," and "economics" as areas that had low response rates and low search rates [9]. In this category, the "education," "culture," and "management" more than other topics require the researchers' attention. Unfortunately, environmental education is one of the most useful tools to increase the awareness of environmental concepts and is vital for improving the environmental issues in a society [45] is located in this category.

Finally, the top left corner contains topics highly seeks by users but do not have sufficient research resources. Due to the high demand for research on these topics, users do not have sufficient research resources; this part is known as "gap topics." The "river," "industry," "international rights," and "garbage" are the most important environmental research gaps in the theses. "Waste," "environmental health," and "wastewater" are also very close to this part, and they can also be considered as a gap. Previous research in this field identified "training" and "pollution" in the category of high search rate and low response rate [9].

Previous scientometric studies analyzed the supply side of a research area [11], [13]–[16], [28], [30]–[32], [44] and in the studies trying to identify the research priorities or making policy in the research management [3], [4], [6], [46], experts almost identify the demand side. Although experts' knowledge to identify the research needs and priorities is inevitable, using experts' opinions for different domains is very time-consuming and expensive, and in some cases, because of the complexity of the relationships among the research domains, it is impossible. For example, in [28], the target documents directly described, the research domain was extracted manually after the data extraction from WoS. This research applied text mining tools and techniques to propose a novel framework to analyze the actual behavior of a large number of researchers (as experts) instead of using a limited number of experts, which is regular in existing studies, such as [9], [15].

Unlike most studies that examined the supply side, Rabiei *et al.* [9] analyzed the demand side to identify the research gaps and priorities. This study analyzes the both supply and demand sides of the research ecosystem. Using OCC, the boundaries of the target domain in the search log database are identified as environmental records. This approach automatically extracts the target domain documents without any dependency on expert knowledge. This process is almost relying on expert knowledge or applying specifically limited dataset in the previous studies [5], [13]–[16], [44]. Therefore, it is useful in the macro analysis of the research, especially when the data are huge in volume and diverse (big data). After that, the bibliographic analysis of the environmental supply and demand was presented.

Undoubtedly, the research demands of users are very in line with the users' actual information needs [47]; it should be emphasized that some information needs may never be expressed as demand for various reasons. Information limitation [48], low information literacy [49], or users' disappointment of finding the right answer from the database are some of these reasons. Therefore, some parts of the research needs remain dormant need [50].

Research policymakers can encourage researchers to do research on issues that are located in the gap or cold zone (see Figure 8). On the other hand, researchers can focus on topics that are located on the left-hand side of Figure 8 rather than on topics that have been adequately researched. As Figure 8 points out, topics of human science, such as management, education, culture, and law are neglected or diminished in environmental research.

5. Conclusion

This study proposed a new framework that analyzes the research

supply and demand without depending on the domain expert's knowledge. Although environmental research is applied in this framework, this framework can be used in future studies of other research areas.

The trend analysis of the environmental theses indicated that the number of resources grew significantly between 2007 and 2014. A similar upward trend was evident in the environmental demands. The category of the environmental theses was changed from medical science before 1980 to engineering science after 1980.

The environmental topics are identified with the topic modeling approach. Then the gap analysis of supply and demand leads to determine the four categories. These categories are hot topics (high demand and high supply), silent topics (low demand and high supply), gap topics (high demand and low supply), and cold topics (low demand and low supply). Research policymakers can identify the status of the environmental topics by studying and monitoring the supply and demand of research in this field. They can also apply this information along with the information of other resources (macroperspective of national research, comprehensive scientific program, available financial resources, etc.), such as experimental or observational knowledge, to manage the research direction of the environmental topics.

This study neglects the geographical location of research demands and, on the other hand, does not care about the location from which a research is supplied. For example, there may have been a lot of research on drought in the western provinces, but due to the geography of the country, the problem of water shortage has been requested by the central provinces. In this case, it is clear that the geographical distribution of these studies is not balanced with the research demands. The geographical analysis of the research supply and demand and determination of the location-based research gap are suggested for future research. This research implicitly assumed that users' demands were in line with their information needs. Discovering the dormant needs of research can provide research policymakers with more comprehensive information on the research demand, which can be considered in future studies.

References

[1] Mai, J.-E., D. O. Case, and L. M. Given, *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*, 2016, Bingley, U.K.: Emerald Group Publishing.

[2] Lind, F. and H. G. Boomgaarden, What we do and don't know: A meta-analysis of the knowledge gap hypothesis. *Annals of the International Communication Association*, 2019. 43(3): pp. 210–224.

[3] Robinson, K.A., I. J. Saldanha, and N. A. Mckoy, Development of a framework to identify research gaps from systematic reviews. *Journal of Clinical Epidemiology*, 2011. 64(12): pp. 1325–1330.

[4] Nasser, M. and V. Welch, Prioritization of systematic reviews leads prioritization of research gaps and needs. *Journal of Clinical Epidemiology*, 2013. 66(5), p. 522.

[5] Amsaveni, N. and H. C. Krishnan, A scientometric analysis of environmental management research output during 1989 to 2014. *Library Philosophy and Practice*, 2018, p. 1.

[6] Martinez, S., *et al.,* Science mapping on the Environmental Footprint: A scientometric analysis-based review. *Ecological Indicators*, 2019. 106, pp. 1–11.

[7] Fatemi, M., *et al.,* Sustainability of environmental management in Iran: An ecological footprint analysis. *Iran Agricultural Research*, 2018. 37(2), pp. 53–68.

[8] Rabiei, M., S.-M. MahdiHosseini-Motlagh, and B. Minaei Bidgoli, Using one-class SVM for scientific documents classification case study: Iranian environmental thesis. *Iranian Journal of Information Processing and Management*, 2019. 34(3), pp. 1211–1234.

[9] Rabiei, M., S.-M. Hosseini-Motlagh, and A. Haeri, Using text mining techniques for identifying research gaps and priorities: A case study of the environmental science in Iran. *Scientometrics*, 2017. 110(2), pp. 815–842.

[10] Rodela, R. and D. Alašević, Crossing disciplinary boundaries in environmental research: Interdisciplinary engagement across the Slovene research community. *Science of the Total Environment*, 2017. 574, pp. 1492–1501.

[11] Aboelmaged, M. and S. Mouakket, Influencing models and determinants in Big Data analytics research: A bibliometric analysis. *Information Processing & Management*, 2020. 57(4), 102234.
[12] Ebadi, A., *et al.,* Application of machine learning techniques to assess the trends and alignment of the funded research output. *Journal of Informetrics*, 2020. 14(2), 101018.
[13] Hua, J. and Y. Zhang, Research patterns and trends of Recommendation system in China using co-word analysis. *Information Processing & Management*, 2015. 51(4), pp. 329–339.
[14] Amirbagheri, K., *et al.,* Research on green supply chain: A bibliometric analysis. *Clean Technologies and Environmental Policy*, 2019. 21(1), pp. 3–22.
[15] Frank, J.J., *et al.,* Systematic review and meta-analyses of lead (Pb) concentrations in environmental media (soil, dust, water, food, and air) reported in the United States from 1996 to 2016. *Science of the Total Environment*, 2019. 694, 133489.
[16] Hu, K., *et al.,* Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Information Processing & Management*, 2019. 56(4), pp. 1185–1203.
[17] Onan, A., Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, 2019. 7, pp. 145614–145633.
[18] Onan, A., Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 2021. 33(23), e5909.
[19] Onan, A., Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine*, 2018. 2018, 2497471.
[20] Onan, A. Topic-enriched word embeddings for sarcasm identification. in *Computer Science On-line Conference*, 2019, pp. 293–304.
[21] Onan, A. and S. Korukoğlu, A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, 2017. 43(1), pp. 25–38.
[22] Toçoğlu, M.A. and A. Onan. Sentiment analysis on students' evaluation of higher educational institutions. in *International Conference on Intelligent and Fuzzy Systems*, 2020, pp. 1693–1700.
[23] Onan, A., Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, 2020. 28(1), pp. 117–138.
[24] Fatahi, S. and M. Rabiei, Users clustering based on search behavior analysis using the LRFM model (case study: Iran scientific information database (Ganj)). *Iranian Journal of Information Processing and Management*, 2020. 36(2), pp. 419–442.
[25] Fatahi, S., A.H. Seddighi, and M. Rabiei, Analyze user behavior patterns on academic search engines, in *4th International Conferene on Soft Computing*, 2021, pp. 1438–1449.
[26] Onan, A., V. Bal, and B. Yanar Bayam, The use of data mining for strategic management: A case study on mining association rules in student information system. *Croatian Journal of Education: Hrvatski časopis za odgoj i obrazovanje*, 2016. 18(1), pp. 41–70.
[27] Rabiei, M., *et al.,* Evolution of IT, management and industrial engineering research: A topic model approach. *Scientia Iranica*, 2021. 28(3), pp. 1830–1852.
[28] Piryani, R., D. Madhavi, and V.K. Singh, Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, 2017. 53(1), pp. 122–150.

[29] Tress, B., G. Tress, and G. Fry, Integrative research on environmental and landscape change: PhD students' motivations and challenges. *Journal of Environmental Management*, 2009. 90(9), pp. 2921–2929.
[30] Zhao, X., A scientometric review of global BIM research: Analysis and visualization. *Automation in Construction*, 2017. 80, pp. 37–47.
[31] Comte, A. and L.H. Pendleton, Management strategies for coral reefs and people under global environmental change: 25 years of scientific research. *Journal of Environmental Management*, 2018. 209, pp. 462–474.
[32] Tirgar, A., S.A. Sajjadi, and Z. Aghalari, The status of international collaborations in compilation of Iranian scientific articles on environmental health engineering. *Globalization and Health*, 2019. 15(1), p. 17.
[33] Onan, A., Classifier and feature set ensembles for web page classification. *Journal of Information Science*, 2016. 42(2), pp. 150–165.
[34] Onan, A., Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*, 2017. 46(2), 330–348.
[35] Onan, A., An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, 2018. 44(1), pp. 28–47.
[36] Onan, A., S. Korukoğlu, and H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems With Applications*, 2016. 57, pp. 232–247.
[37] Fatahi, S. and M.J. Ershadi, Assessment of user satisfaction of research theses and theses in iranian scientific database (treasure): Based on E-Qual model. *Iranian Journal of Information Processing and Management*, 2020. 35(2), pp. 399–424.
[38] Fatahi, S. and A. Naimi-Sedigh, Analysis of researchers information seeking behavior in national search engine thesis information system in Iran. *Iranian Journal of Information Management*, 2017. 2(2), pp. 31–58.
[39] Manevitz, L.M. and M. Yousef, One-class SVMs for document classification. *Journal of Machine Learning Research*, 2001. 2, pp. 139–154.
[40] Mohr, J.W. and P. Bogdanov, Introduction—Topic models: What they are and why they matter. *Poetics*, 2013. 41(6), pp. 545–569.
[41] Arun, R., et al. On finding the natural number of topics with latent Dirichlet allocation: Some observations. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2010, pp. 391–402.
[42] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003. 3, pp. 993–1022.
[43] Blei, D.M., Probabilistic topic models. *Communications of the ACM*, 2012. 55 (4), pp. 77–84.
[44] De Battisti, F., A. Ferrara, and S. Salini, A decade of research in statistics: A topic model approach. *Scientometrics*, 2015. 103(2), pp. 413–433.
[45] Staniskis, J.K. and Z. Stasiskiene, An integrated approach to environmental education and research: A case study. *Clean Technologies and Environmental Policy*, 2006. 8(1), pp. 49–58.
[46] Gosselin, F., *et al.,* Ecological research and environmental management: We need different interfaces based on different knowledge types. *Journal of Environmental Management*, 2018. 218, pp. 388–401.
[47] Oeldorf-Hirsch, A., et al. To search or to ask: The routing of information needs between traditional search engines and social networks. in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2014, pp. 16–27.
[48] Ganji, A., M. Rabiei, and A. Rahmati, The effect of information limitations on information usage of Tehrani youth users. *Iranian Journal of Information Processing and Management*, 2018. 33(3), pp. 1301–1318.

[49] Day, R.E., Before information literacy [Or, Who Am I, as a subject-of-(information)-need?]. *Proceedings of the Association for Information Science and Technology*, 2017. 54(1), pp. 57–70.
[50] Omiunu, O.G., Conceptualizing information need: A phenomenological study. *Journal of Library and Information Sciences*, 2014. 2(2), pp. 29–54.

Mohammad Rabiei received the B.Sc. degree in computer science from the Shahid Bahonar University of Kerman, Kerman, Iran, in 2007, and the M.Sc. and Ph.D. degrees from the K.N.T University of Technology, Tehran, Iran, in 2009, graduated in 2019 from the Iran University of Science and Technology, Tehran, both in IT engineering (E-commerce). He began his scientific experience as a Lecturer in 2006 and joined IranDoc as a Faculty Member in 2010. Now, he is an Assistant Professor of E-Business Group. His efforts at IranDoc has mainly devoted to application of computer sciences in research management process and information systems analysis. His research interests include text mining, natural language processing, and user information behavior analysis.

Somayeh Fatahi received the B.Sc. and M.Sc. degrees in computer engineering from Razi University, Kermanshah, Iran, and Isfahan University, Isfahan, Iran, in 2006 and 2008, respectively, and the Ph.D. degree in computer engineering (artificial intelligence and robotics) graduated in 2016 from the University of Tehran, Tehran, Iran. From 2010 to 2011, she worked in Kermanshah University of Technology, Tehran, as a Faculty Member. From 2011 to 2014, she was a Researcher with Iran Telecommunication Research Center. From 2014 to 2016, she was a Researcher in Big Data Institute with Dalhousie University, Halifax, Canada. She began her research in 2016 as a Faculty Member in IranDoc. Now, she is an Assistant Professor of Information System Research Group. Her efforts at Irandoc have mainly devoted to analyzing researchers' seeking information behavior on GANJ to improve search engine. Her research interests include user modeling, data mining, big data analytics, pattern recognition, e-learning, and machine learning.