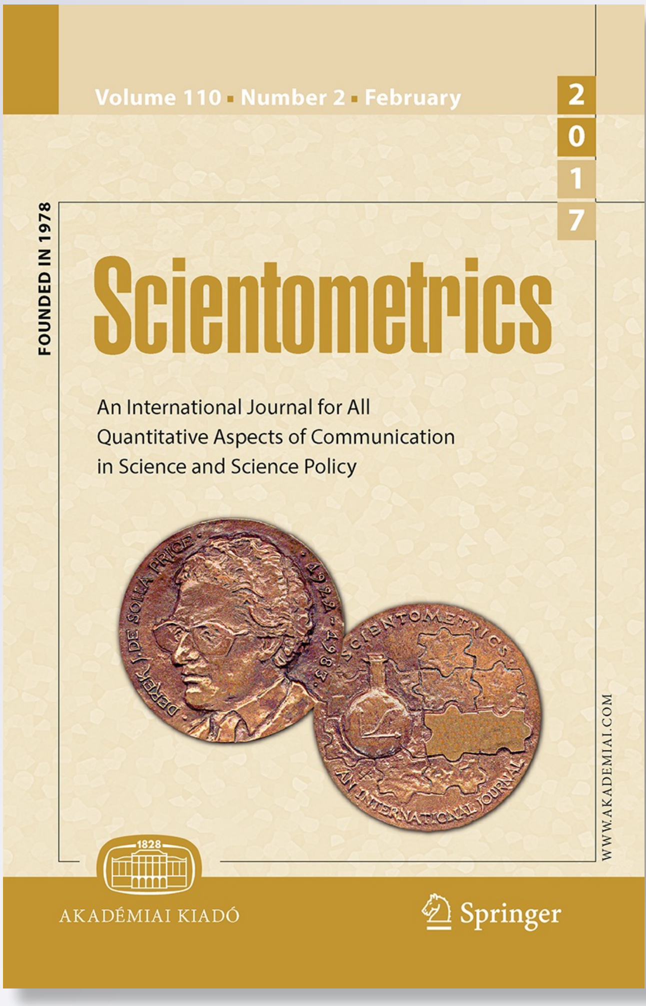
Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran

Mohammad Rabiei, Seyyed-Mahdi Hosseini-Motlagh & Abdorrahman Haeri

Scientometrics
An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

ISSN 0138-9130
Volume 110
Number 2

Scientometrics (2017) 110:815-842
DOI 10.1007/s11192-016-2195-8



Springer

Scientometrics (2017) 110:815–842
DOI 10.1007/s11192-016-2195-8

Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran

Mohammad Rabiei[1,2] · Seyyed-Mahdi Hosseini-Motlagh[2] · Abdorrahman Haeri[2]

Received: 7 June 2016 / Published online: 3 December 2016

Abstract This study aims to observe the researchers' behavior in Iranian scientific databases to determine the research gaps and priorities in their field of research. Text mining and natural language processing techniques were used to identify what researchers are looking for and to analyze existing research works. In this paper, the information about the behavior of researchers who work in the field of environmental science and existing research works in the Iranian scientific database are processed. The search trends in all areas are evaluated by analyzing the users' search data. The trend analysis indicates that in the period of February 2013 to July 2015, the growth of the researchers' requests in some domains of the environment such as Industry, Training, Assessment, Material, Water and Pollution was 1.5 up to 2 times more than the overall requests. A Combination of the trend analysis and clustering of queries led to shaping four priority zones. Then, the research priorities for each environmental research area were determined. The results show that Training, Pollution, Rangeland, Management and Law are those domains in the environmental research which have the most research gaps in Iran, but there are enough research in Forest, Soil and Industry domains. At the end, we describe the steps for the implementation of a decision support system in environmental research management. Researchers, managers and policy makers can use this proposed "research demand and supply monitoring" system or RDSM to make appropriate decisions and allocate their resources more efficiently.

✉ Seyyed-Mahdi Hosseini-Motlagh
motlagh@iust.ac.ir

Mohammad Rabiei
m.rabiei@irandoc.ac.ir; mo.rabiei@ind.iust.ac.ir

Abdorrahman Haeri
ahaeri@iust.ac.ir

[1] Iranian Research Institute for Information Science and Technology (IranDoc), Tehran 131573314, Iran

[2] School of Industrial Engineering, Iran University of Science and Technology (IUST), Tehran, Iran

Keywords Research priority · Research gap · Text mining · Environment studies · Researchers' behavior analysis

Introduction

The scientific databases store, organize and publish information about research works. The volume and diversity of such information are expanding every day. Moreover, the usage of these databases as a reliable repository of the latest studies done by the researchers in various scopes of science has increased. In addition to the content of these databases, the users' interaction with them is substantial because the users are often researchers and experts. Therefore, analyzing their interactions with these databases can yield valuable information. Applying data mining and text mining for knowledge discovery from existing sources of data is growing every day. Data mining or particularly text mining is considered as excellent tools for pattern recognition within vast repositories of information. One of the most important scientific databases in Iran is IRANDOC (Which is held by the Iranian Research Institute for Information Science and Technology).[1] This database whose users are mainly Iranian academic researchers and Ph.D. or master's students is a place where data about various scientific resources in different fields are collected, organized and published. It contains more than 850,000 scientific resources of which about 400,000 are theses or dissertations. It has more than 9000 online visitors per day and over 50,000 real users' search queries take place on this website every day. (IranDocIranian Scientific Repository 2016a, b, c).

Environmental science is a distinguished domain which joins engineering, experimental and human sciences together. This domain has multidisciplinary characteristics and is associated with economic, social, cultural, political, and geographical issues. Moreover, this field is of high importance because of its close relationship with human lives. Priority setting is recognized as an essential step in the environmental management process by a wide range of experts including government agencies, research institutes and professional management associations (Gunton 2002). Hence, it is necessary to organize a "research demand and supply monitoring" system or RDSM to continually monitor research studies in environmental sciences in order to direct future research in an effective and efficient manner so that the research needs and priorities in the country are converged. In addition, it is necessary for managers and policy makers to determine research priorities so as to allocate their limited resources in an efficient way.

Recently, many research works have been carried out to study the research trends in many disciplines of science (Kim and Chen 2015; Kademani et al. 2013; Gu and Blackmore 2016; Mesdaghinia et al. 2015). The scientometric analysis has commonly been used in these studies. These methods are based on analysis of citation, co-author relationship, author affiliations, keywords networks, and other bibliographic information of publications. These studies just focus on the supply side of the research works and the demand side is neglected. Because of the special characteristics of the environment in Iran, a lot of research studies are carried out each year. Monitoring and identifying research gaps and priorities in this scope can help managers and policy makers to effectively decide for the future. Moreover, the trend of researcher needs in the field of environment has not been

[1] http://ganj.irandoc.ac.ir.

studied yet in Iran. This study focuses on the demand side of the research works and aims to compare the researchers' needs and the trends with publications in environmental science to determine the research gaps and priorities.

This paper tries to analyze the environmental researchers' interactions with this database to explore various patterns of their behavior and identify the research priorities in this scope. Moreover, we introduce a novel method for managers and policy makers as RDSM to monitor research demand and supply and determine research priorities to invest in appropriate areas of a specific knowledge domain by understanding the researchers' needs and comparing them with existing research in that domain.

Research questions and objectives

In the text mining process, various techniques such as natural language processing (NLP), statistical analysis and data mining are used. Although many studies in the field of environmental science use data mining tools and techniques, no considerable work has been done in research management for the environmental science. Specific problems that arise frequently in research management conduction, which can potentially be addressed by text mining, include the questions such as: What areas of science or which techniques are subjected globally in the research studies? What areas are not considered? What are the promising directions for new research? Who is working on them? Where is it taking place? What are the relationships between major areas and supporting areas? What are the innovations and discoveries? (Losiewicz et al. 2000). Although answering all of these questions are not in the scope of this study, we apply the researchers' behavior in the scientific repository and focus on the first three above questions to find the research trends, gaps and priorities.

The efforts in developing countries to achieve new industries and technologies, increasing population and a shortage of critical resources, as well as climate changes are some reasons emphasizing the importance of science-based environmental management in Iran. On the other hand, managers and policy makers in the environmental scope need to know the researchers' demands and research priorities to allocate their limited resources to environmental areas appropriately. This study aims to demonstrate the research status in the environmental science in Iran and determine the research trends, gaps and priorities in various domains and subdomains of this domain of science. To address above objectives, this study tries to answer the following questions:

RQ1 What is the status of the research demands in the environmental science in Iran in comparison with the status of all research demands?
RQ2 What are the research gaps in the environmental science based on the researchers' perspectives?
RQ3 What are the research priorities in the environmental science in Iran?

Related work

Text mining and its application in knowledge management

Data mining tries to discover the information hidden in a data repository, which is not easily accessible by simple statistical techniques (Han et al. 2011). The text mining or

Textual Data Mining (TDM) is a significant subset of data mining that aims to extract knowledge from unstructured or semi-structured textual data and has widespread applications in analyzing and processing textual documents (Choudhary et al. 2009).

The text mining can be applied in the following areas: text summarization (Bijalwan et al. 2014), machine translation (Shachi and Jaiswal 2014), keywords selection or text tagging (Noh et al. 2015), similar text recognition and plagiarism detection (Oberreuter and Velásquez 2013) and its changes by using system feedbacks (Akçapınar 2015), text sentiment and feeling analysis (Erdmann et al. 2014), users' expectation discovery (Lu and Eldin 2014), and automatic hypertext construction (Hsin-Chang and Lee 2005).

There are numerous research studies done in research management scope by using data mining or text mining techniques that have followed their specific goals. Some of these research works tried to illustrate the future research trends via database content analysis (Delen and Crossland 2008) or compared the research trends in several scopes (Sunikka and Bragge 2012). Some others focused on research areas to find their interrelationships and to track their growth or decline and tried to classify and cluster them (Tsai 2011). For instance, the titles of theses done in the field of industrial engineering in Iran were analyzed and important words were extracted from them in the text mining phase. After that, based on the results of the first step the words network was constructed in the social network analysis phase and the knowledge discovered via network features was presented (Jalalimanesh 2012).

The Research Management is another considerable application of data mining. Data mining can be used to analyze collaborations between academic institutions and interdisciplinary research and the relationship between faculty members and industry executives (Kirkland 2010). In addition, research management for these projects in different universities and determination of associated rules between the subjects of the research projects have been investigated using text mining (Chen et al. 2012). The combination of data mining and information retrieval can be exploited to help us discover the relationship between academic publications in various fields. The relationship between conference papers and journal papers was explained by this approach and it was specified that conference papers affected the journal papers which would publish over three years (Tu and Seng 2009).

The log files as a valuable data warehouse

The log is defined as a footprint of every activity performed by any users or system. Logs are stored in a database file format and often contain an enormous amount of information. Log files are considered as precious data repositories because of their widespread applications in knowledge discovery in various scopes. For example, chat room analysis identifies user interests and attitudes, key users and important groups and determines the relationships among them (Anwar and Abulaish 2014). Moreover, it is possible for the search engines to present personalized search results to users by analyzing the search history of users, clustering their information, and constructing decision trees (Gadkari et al. 2015). Furthermore, a novel solution has been offered for search engines to provide more relevant search results to information seekers. This approach determines the relevance of information by allocating weighs to the terms stored in log files (Saneifar et al. 2014).

The clustering of clickstream data (traces of users' click events) has become one of the most efficient ways to users' behavior analysis (Wang et al. 2016). A map of science is constructed to visualize the relationships between scientific journals based on user clickstreams log data. The log data contains approximately 1 billion user interactions in scientific web portals. This map clarifies the connection among three main domains (humanities, natural science, and social science) of the journal. An analysis of 749 articles, which have been published and indexed under "Information needs" and "Information uses" in library literature and information science, shows that about 10% of them employed the transaction log analysis as a method (Julien et al. 2011). Log analysis and questionnaire data for 757 users of ScienceDirect are combined to gain a better understanding of users' behavior. This combination showed that there are large differences between users' interactions with electronic journal systems in different subjects (Nicholas et al. 2008).

There are a few numbers of studies which focused on log analysis in scientific repositories in Iran. However, all of them concentrated on usage characteristics such as time, location, session duration, number of users, hit rate, and query length (Khosravi et al. 2014; Nadjla and Sahar 2014). In some research works (Khosravi et al. 2014) query subjects are also explained, but there is no considerable study on the content analysis for queries in scientific databases in Iran.

Analysis of log files to predict software failure (Fronza et al. 2013), monitoring the behavior of students in online classes to identify their interests (Beth and Deyrup 2015), and using logs of various systems to find a collaborative process among those systems (Claes and Poels 2014) are just some examples for wide applications of logs for extracting knowledge from them. There are various types of logs, including error logs, server logs, system logs, database logs, and so on.

Data mining application in the environmental science

In spite of the adoption of Information Technologies in environmental management, there is a significant demand for environmental information. In this regard, some decision support systems (DSS) are developed for this area (Kouziokas 2016) and data mining mainly has its own technical applications in the environmental issues. Forecasting air quality parameters (Kolehmainen et al. 2011), prediction of water quality (Kusiak et al. 2013), prediction of groundwater level fluctuations (Lin et al. 2012), assessment of environmental changes as a result of natural disasters (Souza 2014), monitoring the changes in forest vegetation (Liu et al. 2008) and drought forecasting (Choubin et al. 2014) are just some examples of data mining applications in environmental science.

Methodology

In this research, CRISP DM is used to implement the text mining process. This approach, illustrated in Fig. 1, includes six steps of data mining projects: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment (Shearer 2000). In what follows, these steps are explained.

The CRISP DM model is modified for the text mining in the search log file to find the research gaps and priorities in this study. Figure 2 presents this conceptual framework in details and input and output of each step are clarified.

As Fig. 2 demonstrates, in this study the researchers' behavior in scientific data base is applied to answer the research questions. The researchers' interactions in the scientific repository are collected in a database as log records which contains information about users' interactions. These records include information about user characteristics (UserID,

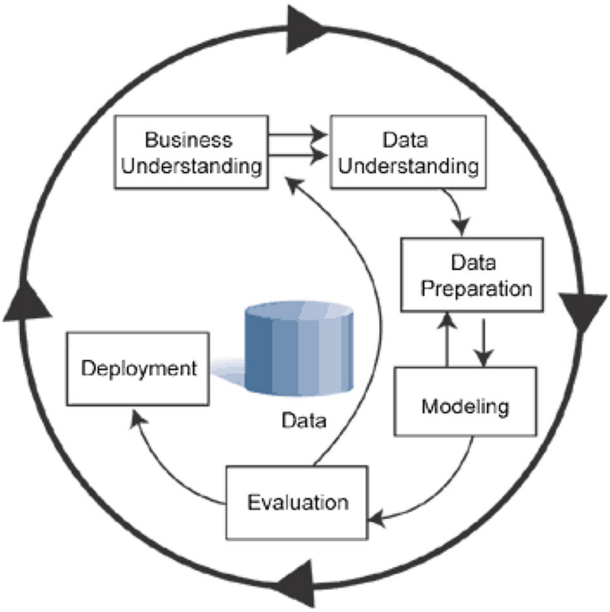


Fig. 1 CRISP DM model (Shearer 2000)

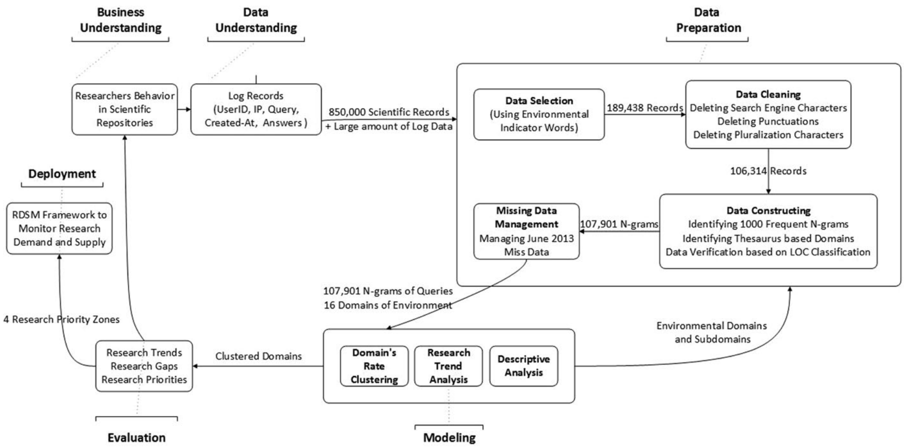


Fig. 2 The modified CRISP DM framework for researchers' behavior analysis

IP…) and query attributes (Query, Created-at, Answers…). To select an appropriate data set for the environmental scope, the relevant indicator words are determined based on related thesauri. The selected data set in this step contains 189,438 records. To purify this data set, some special characters which were inserted in the query by search engine to manage the search process are omitted. Moreover, the punctuations and pluralization characters are removed. After data cleaning phase, 106,314 records are remained. The field Query is tokenized in these records and frequent terms are identified. The terms are categorized based on the number of consequent words in each terms and single word terms

(Uni-grams), two words terms (Bi-grams), … are constructed. The domain of each term is identified by considering frequent n-grams and thesaurus based indicators. Then 16 Identified domains are compared with Library of Congress classification domains of environment. After managing the missing data, the constructed data set that contains 107,901 terms and their related domain, searching date and number of answers are fed as input in the modeling step.

In the modeling step, descriptive analysis explains the research status in the environmental science and trend analysis illustrates the research trends. Other text mining techniques are applied in this step to determine the research gaps and a hybrid approach is used to identify the research priorities. The research trends, gaps and priorities is evaluated by previous related studies and four priority zones are shaped. Finally, a decision support system is suggested based on this framework. This DSS provide the RDSM that could help the policy makers to monitor research demand and supply and identify the research priorities to allocate their limited resources effectively. The following sections describe the above steps based on proposed conceptual framework.

Business understanding

Environmental science is considered as one of the most important and challenging fields in Iran. In educational scope, various academic majors and minors related to environmental science are established in universities and research institutions. Therefore, a large amount of research is done in this area every year and many researchers desire to investigate and study these research works. This paper, aims to observe the users' interests in this scope and to identify researchers' demands. To this end, NLP and text mining techniques were used to recognize the hidden patterns in researchers' query streams. The research status in different domains and sub-domains of environmental science as well as research gaps and priorities based on researchers' demand can be obtained by analyzing these patterns. To collect the information about research works in the environmental scope, the IRANDOC database was used. This database contains more than 850,000 documents such as theses and dissertations, research proposals, research project reports, journal and conference papers, etc. (IranDoc, Iranian Scientific Repository 2016a, b, c). The majority of documents are in Persian. Also, the users' queries are often in Persian. IRANDOC is the only official reference in Iran for registering, organizing and publishing theses and dissertations. Theses and research proposals are registered in this database after they are approved in the universities where they are conducted. For these reasons, researchers often use this database to find latest research works in Iran. After searching, users will be given access to documents based on IRANDOC access policy (IranDoc, IRANDOC Information Reposiory 2016a, b, c).

Data understanding

When users enter their queries on the website, the information about their search queries is stored in the database as a log record. The log record consists of user ID, date and time of the search, IP of the machine and the number of search results provided for the user. The search logs stored since February 2013 are available in the database. Before October 2014, users had to register and login to the website to be able to use the database, but after that IRANDOC changed its policy and all users has been able to use the website's search engine without registration ever since then. Therefore, search engine robots can search the contents of the database just like real users. The database is in MYSQL form and Persian texts are stored in the Unicode format.

Data preparation

The biggest difference between the areas of knowledge discovery is a data pre-processing step in the CRISP-DM methodology (Houškova and Houška 2011), Data pre-processing is the most time-consuming phase in the whole process of knowledge discovery (Munková et al. 2013) and it becomes more difficult when data is in text format and lingual characteristics have to be considered.

Data selection

The IRANDOC users are diverse in various domains. Therefore, it is necessary to filter the search log records to create an appropriate dataset for environmental queries. Each record in the log file contains the following fields: QUERY as a text string; USER_ID which is an identifier for registered users and is empty for unknown users; CREATED_AT that is the date and time at which the query is searched; ANSWER which is the number of answers returned to the user; and some technical fields for software controlling. Selecting the appropriate dataset from environmental information seekers' with a large number of records stored in this general log file is difficult. Moreover, just QUERY is a distinctive field. For this purpose, Persian scientific thesauri[2] were applied. "Life Science Thesaurus" and "Health Promotion Thesaurus" were used to find the relevant terms of "environmental science" in Persian (IranDoc, IrandDoc Thesauri 2016a, b, c). The translation of "environment" in Persian consists of two words "MOHIT" + "ZIST"; "MOHIT" is the meaning of "ambience" and the ZIST is the meaning of "life" and "ALOUDEGI" which refers to "pollution" is the most frequent word related in these thesauri. Based on this approach, the terms "MOHIT", "ZIST" and "ALOUDEGI" were selected as indicators of environmental scope in the Persian language. Then, users' queries that contained any combination of these three words were investigated. The first limitation of using this method to select the appropriate dataset is that some queries which contain just a subdomain of the environment science and have not any combination of the above three words in the query string are neglected. This limitation may cause us to ignore some scarce subdomains of the environmental science. Although, this method may ignore some relevant records, expanding selected indicators to more keywords leads to collect irrelevant search queries. In other words, this approach decreases the 'recall' measure,[3] but it could guarantee that the retrieved records are almost relevant to the environmental science. Whereas the selected queries are used to extract the environmental phrases and these phrases are applied in the data analysis, the 'precision' measure[4] for queries is more important than recall. The second limitation of this approach is that using previous studies to determine future priorities have been limited to retrospective analyses. But applying todays behavior of researchers could balance this limitation. After this step, 189,438 records from February 2013 to July 2015, were collected.

[2] Available in: http://thesauri.irandoc.ac.ir/.

[3] The fraction of the records that are relevant to the query that are successfully retrieved.

[4] The fraction of retrieved records that are relevant to the query.

Data cleaning

In the first step of this phase, some words which were just orthographically similar to the selected keywords and essentially different from "environment" were eliminated (e.g. "Transistor" or "Existentialism" in the Persian text format contain term "Zist", but are fundamentally different from "environment").

IRANDOC policies allow search engine robots to index metadata. Thus, a significant portion of the dataset represents the activities of these robots (e.g. Google, Bing and others). Consequently, the records related to search engines' were omitted and only the real users' searches (106,401 Records) were considered.

IRANDOC uses Sphinx[5] as its main core. Users can choose between three different search options including basic, advanced or professional search in the website. In the advanced search option, Sphinx adds some characters automatically to the search query (Sphinx 2001). In the professional search option, users can use Boolean operators.[6] Therefore, the data cleaning phase is necessary for data preparation. For this purpose, the combination of NLP and data cleaning techniques were applied to the users' queries as bellow:

- Character "+" was replaced with a blank space in the advanced search option.
- Characters "&", "/", "\" and '’|’ were eliminated in the advanced search option.
- Records in which character "!" meant "Not" were deleted.
- Characters like ",", "(", ")", "-", "_", "{", "}", "[", "]" < "?", "$", ";", "." were eliminated.
- Persian pluralization characters (same as "s", "es" in English) were eliminated.

After applying these changes, the records in the data set were reduced to 106,314 records. Table 1 shows the number of remained records after each step of the data cleaning.

Data construction

The next step of data preparation in this study is tokenizing query strings to construct phrases and find their frequency. After tokenizing each query to single words, the adjacent sequence of words (n-gram) is considered to categorize the frequent phrases. An n-gram is a contiguous sequence of n items from a text or speech. These items can be characters, syllables, words or other text parts. Using n-grams is prevalent in NLP and it has various applications such as text summarization (Günes and Radev 2004), machine translation (Marino et al. 2006), question answering (Davide et al. 2010) and so on. The N-gram frequency method provides an inexpensive and highly effective way of classifying documents (Cavnar and Trenkle 1994). While the terms with the more number of words denote to more specific topics and have more information value, N-gram is an effective and fundamental approach for text categorization.(e.g. "Water" is an Uni-gram term, whereas "Water Crisis" is a Bi-gram with more information value and "Water Crisis Management" is a Tri-gram that indicates a specific concept) In this study, N-gram is used as N continuous words in the users' queries and it applied just as a measure to categorize the queries.

[5] An open source engine for full text search.

[6] Boolean Operators are used to connect and define the relationship between search terms (e.g. +,−, &, not, etc.).

Table 1 Number of records in each step of data cleaning

Steps of data cleaning	Deleted records	Remained records
After data selection phase	–	189,438
Deleting irrelevant keywords (orthographically similar)	311	189,127
Deleting search engines' queries	82,726	106,401
Managing Sphinx characters	87	106,314

After tokenizing uni-grams, the frequency of their appearance was measured separately per each month. In the next step, Persian stop-words[7] list was extracted and these words were omitted from uni-grams. Then bi-grams, tri-grams, and quad-grams were constructed and their frequencies were measured. Table 2 presents top ten frequent n-grams in the environmental scope dataset. The original list contains Persian n-grams, but they were translated to be globally understandable. For this reason, some n-grams in this list are longer or shorter than their respective names.(for example, "Environment" in uni-grams table refers to "MOHIT" in the Persian language, but "Environmental" in the second row of Bi-grams table refers to "ZIST MOHITI" in Persian). The original Persian list of these terms is presented in "Appendix".

For the next step, the goal was to select and determine the main areas of environmental science. Investigation of the university minors in this field and monitoring top 1000 frequent n-grams yielded 16 main areas. Moreover, Life Science, Health Promotion, and Earth Science Thesauri were used to allocate indicator words of each area (IranDoc, IrandDoc Thesauri 2016a, b, c). Consequently, the n-grams were labeled by domains based on indicator words. By this approach, some phrases are labeled by more than one domain. Therefore, one of them was chosen as the main domain that has a minimum number of phrases labeled by it. Finally, 107,901 n-grams which are labeled by domains were prepared for the text mining process. Table 3 illustrates the areas of environmental science and their important indicator words extracted from thesauri. The Library of Congress (LOC) classification system classifies environmental sciences in subclass GE (general and interdisciplinary literature on the environment; i.e. communication in environmental sciences, environmental ethics, environmental education, environmental quality, global environmental change, the environmental movement, environmental management, and ecological engineering) and related publications in portions of other subclasses: subclass GF (human ecology), subclass QE (environmental geology), subclass QH (interrelationships of organisms and their environment), subclass QC (atmospheric pollutants) and subclass TD (industrial and factory environmental control) (Library of Congress Collections Policy Statements 2016). As Table 3 shows, extracted domains are almost consistent with the LOC classification and illustrates that selected data covers all basic subdomains of the environmental science.

Descriptive analysis of the users' queries

The purpose of using the partitional clustering technique is to focus on identifying independent groups of similar records. Clustering separates data into groups in a way that the records of the same group have the greatest degree of similarity. And at the same time, the degree of similarity between the records of separate groups is at the minimum level. In this

[7] Popular but invaluable words (same as "the", "in", "as"… in English).

Table 2 Top ten frequent n-grams in the users' queries

No.	Translated phrase	Frequency	No.	Translated phrase	Frequency
Uni-grams[a]			Bi-grams		
1	Life	61,162	1	Environment	31,630
2	Environment	33,055	2	Environmental	18,276
3	Pollution	26,213	3	Air Pollution	5600
4	Environmental	18,602	4	Biology	2454
5	Biological	13,229	5	Water Pollution	2190
6	Air	5809	6	Biodiversity	1690
7	Evaluation	5031	7	Environmental Law	1620
8	Water	3952	8	Noise Pollution	1616
9	Law	3555	9	Environment Protection	1573
10	Management	2603	10	Oil Pollution	1554
No.	**Translated phrase**	**Frequency**	**No.**	**Translated phrase**	**Frequency**
Tri-grams			Quad-grams		
1	Environmental Impact	2986	1	Environmental Impact Assessment	1386
2	Environmental Law	1614	2	Hegmataneh Environmental Assessors	1090
3	Environmental Protection	1571	3	Environmental Assessors Community	1090
4	Environmental Education	1510	4	International Environmental Law	546
5	Environmental Impact Assessment	1392	5	Department of Environmental Construction	471
6	Environmental Pollution	1267	6	Rangeland Watershed School of Fisheries	352
7	Environmental Assessors	1090	7	Watershed Fisheries Environmental	352
8	Environmental Management	852	8	Rangeland Watershed Environment Fisheries	352
9	Environmental Assessment	849	9	Koomesh Environmental Community	326
10	Environmental Management	646	10	Environmental Risk Assessment	317

[a] The original lists are in Persian and contents of this table are translated to English

Table 3 The domains and important indicator words for the environmental science

LOC classification	Extracted domain	Important indicator words[a]
QC	Pollution	Pollution, Air, Destruction, Sonic, Sound, Noise, Atmosphere, etc.
GE, GF	Sewage	Sewage, Waste, Disposal, Garbage, Recycling, Sludge, etc.
GE, QC	Water	Water Resources, Watershed, River, Sea, Lagoon, Barrier, Bay, Gulf, Fishing, Fish, Amphibians, Drinking, Rain, Snow, Ice, Drop, Fountain, etc.
GE, QH	Forest	Forest, Jungle, Wood, Tree
GE, QH	Biodiversity	Biodiversity, Habitats, Biodiversity, Living organism, Bacteria, Animals, Plants, Species, Ecology, etc.
GE, QH	Rangeland	Rangeland, Grassland, Pasture, Graze, Grassland, etc.
GE	Assessment	Assessment,Evaluation, Treatment, Effect, Synthesis, etc.
GE	Training	Promotion, Training, University, Research, Dissertation, Education, Awareness, etc.
GE	Management	Management, Planning, Protection, Water Supply, Reclamation, Embankment, Storage, Control, Design, etc.
GE, GF	Economy	Economy, Cost, Capital, Business, Accounting, Export, Import, Market Efficiency, Demand, Contract, Price, Tax, Subsidies, Loans, etc.
GE, GF	Law	Law, Regulation, Rule, Responsibility, Civil, Criminal, Criminal, Damage, Legal, Crime, Security, etc.
GE, QH, QE	Soil	Soil, Erosion, etc.
TD	Material	Material, Oil, Metals, Fertilizer, Chemicals, Fuel, Waste, Gas, Organic materials, Polymers, Synthetic, Glass, Metal, Particles of plastic, Toxic, etc.
QE	Geography	Geography, Iran, Tehran, State, Hegmataneh, City, E-zone, etc.
TD, GF	Technology	Technology, Nanotechnology, Sensor, GIS, Modeling, Information, Genetic, Power, etc.
TD, GF	Industry	Industry, Energy, Mining, Transportation, Factory, Refineries, Cement, Construction, Roads, etc.

[a] The original lists are in Persian and content of this table is translated to English

study, n-grams were clustered based on *All-Freq* and *MaxCountAnswer*. Therefore, K-Means algorithm, one of the most popular and applicable methods in the partitional data clustering, was used. In this technique, K is the number of output clusters (Han et al. 2011).

Before using the clustering algorithm, the variables Coverage and Rate, which were defined in Table 4, are explained more. The Coverage index indicates the portion of all n-grams which are included in one area and the Rate index illustrates the lack of research in one area compared with that of other domains and the necessity of assigning a higher priority to research in that area. It should be mentioned that focusing on a number of answers is not enough to find research priorities, because the number of searches, which is a reliable evidence for research demand, is different from one domain to another. In other words, if Ans_i shows the number of answers in the domain i and if $Preiority_i$ is the research priority in domain i, then domains A and B have the following relation:

$$Ans_A > Ans_B \nrightarrow Priority_A < Priority_B \tag{1}$$

This is because the number of search queries in domain A can be much more than that of domain B but the number of search results may not be as much. Hence, priority is

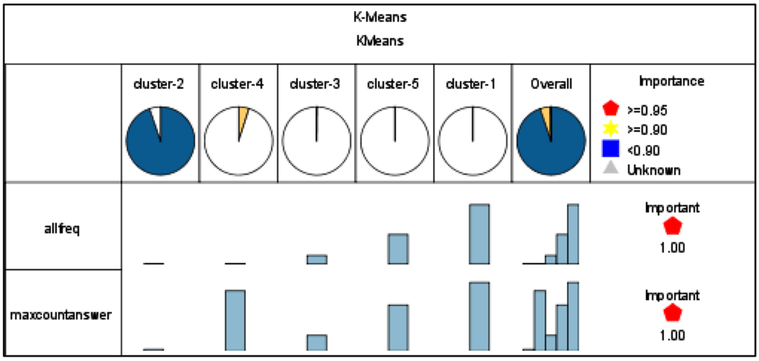


Fig. 3 N-grams clustering, k = 5, All-Freq and MaxCountAnswer are input variables

related directly to the number of search queries (Freq) and has a reverse relation with the number of answers (*Ans*). So it yields:

$$\begin{cases} Rate_A = \dfrac{Ans_A}{Freq_A} \\ Priority_A \propto \dfrac{1}{Rate_A} \end{cases} \tag{2}$$

In the first step, clustering is applied to all n-grams which were labeled by domains (107,901 records). The variables All-Freq and MaxCountAnswer were the input values and the measure of K was chosen as K = 5. Using these parameters to cluster each category of N-grams (uni-gram, bi-gram …) led to the same results and there was not a remarkable difference between each category, therefore the whole phrases were clustered as N-gram. Figure 3 represents the clustering results.

Figure 3, shows that larger clusters contain phrases, which have a small frequency of search and a small number of returned answers. Moreover, as the variable *All-Freq* grows, the number of answers increases too. In other words, popular topics of environmental science, that are searched frequently, have lots of answers (search results). But, the cluster 4 is different than the others. Table 4 describes this in detail.

As specified in Table 4, domains with low Rates such as Rangeland and Law create a small portion of the cluster 4. Also, the Rates of all domains in this cluster are several times as high as the Rates of other clusters. For example, the Rate of Law is 10 times and Management is 7 times as high as their total average Rate. To find the cause of the heterogeneous behavior in these domains, the n-grams of Management and Law in cluster-4 were analyzed. It was found that some phrases in Law such as "Damage", "Fiqh",[8] "Legislation", "Iranian Law" and "Civil liability" which have a high Rate are in cluster-4. In contrast, some low Rate phrases such as "Crime", "Environmental Crime", "Criminal Law" and "Environmental Laws" are in cluster-1 and cluster-3. In the same way, in Management domain, the high Rate n-grams such as "Safety Management", "Management System", "Management Solution" and "Resource Management" are in cluster-4. But the low Rate terms like "Strategic Management", "Risk Management" and "Sustainable Designing" are in the other clusters.

[8] The Islamic theory or philosophy of law.

Table 4 Rate and coverage in all domains of 5 clusters of all n-grams

Clusters	Domains							
	Water	Pollution	Assessment	Economy	Biodiversity	Geography	Forest	Law
1								
Rate	28	22	19	55	58	69	–	9
Coverage (%)	3	18	8	5	8	11	0	8
2								
Rate	132	99	136	153	113	213	114	71
Coverage (%)	10	14	3	6	4	11	1	12
3								
Rate	22	23	13	20	16	30	–	23
Coverage (%)	17	11	2	8	2	2	0	18
4								
Rate	3826	2355	4378	1833	2715	3337	5081	2403
Coverage (%)	12	11	4	7	4	14	1	5
5								
Rate	117	54	15	21	–	3	–	25
Coverage (%)	7	23	5	7	0	11	0	16
Total clusters								
Rate	739	431	674	507	663	812	852	240
Coverage (%)	11	14	3	6	4	11	1	11

Clusters	Domains								
	Soil	Training	Industry	Sewage	Technology	Management	Rangeland	Material	Total Domains
1									
Rate	46	11	–	–	–	57	–	58	36
Coverage (%)	5	11	0	0	0	13	0	11	0.6

Table 4 continued

Clusters	Domains								
	Soil	Training	Industry	Sewage	Technology	Management	Rangeland	Material	Total Domains
2									
Rate	108	202	122	157	159	159	19	170	138
Coverage (%)	2	5	8	3	4	6	0	10	92
3									
Rate	–	9	15	65	6	54	1	17	18
Coverage (%)	0	7	10	1	6	4	5	9	0.2
4									
Rate	3680	2359	2801	2541	4934	3728	–	2823	2985
Coverage (%)	3	4	11	2	6	6	0	11	6
5									
Rate	–	190	116	12	203	110	–	67	56
Coverage (%)	0	2	2	65	2	9	0	11	0.2
Total clusters									
Rate	604	494	826	528	1147	560	3	715	602
Coverage (%)	2	5	8	3	4	6	0.1	10	100

In the next step, to define a specified threshold for "Low" or "High" Rate, K-means technique was applied to cluster the records based on their Rate. In this case, we considered K = 3 to divide n-grams into three clusters (High Rate, Medium Rate and Low Rate). Figure 4 shows the result of this clustering.

The K-means algorithm clusters the n-grams based on the similarity of their Rates. As it is shown in Fig. 4, phrases are divided into three clusters. Cluster 1 consists of n-grams that $0 \leq Rate < 4644$ which is supposed as low Rate. As mentioned in Fig. 4, the phrases with low Rates have high priorities. In other words, these phrases are located in research gap area. The research gap refers to areas with a large amount of search queries or areas in which there is little research done compared to other areas. In cluster-3 there are topics which have a medium Rate. The clustering algorithm determined $4644 \leq Rate < 17248$ as medium measures for Rate. The n-grams of this cluster are at the second level of priority. Finally, $Rate \geq 17248$ indicates a high Rate for the phrases of cluster-2. These phrases have low research priorities, because there are low amounts of requests for them or a lot of research works have already been done in these scopes. Table 5 shows various domains of environmental science and most frequent phrases in each domain ordered by their average Rates. Also, the Rate clustering yielded some subdomains and highlighted the subdomains which have high levels of research needs and priorities.

Looking at Table 5, it can be concluded that the problem of withering Lakes in Iran in recent years, especially the Urmia Lake has brought this issue to the attention of the research communities. Therefore, its related terms are located in the High Rate column in the areas of Water and Geography. In the environmental management scope, researchers in operational management have conducted mainly midterm and long-term management studies. Due to the lower number of search results in comparison with users' search queries for issues such as strategic management or risk management, more research on these topics of the environmental science is required. The previous research about the investigation of environmental performance shows that although this subdomain is a hot topic for managers, long-term maintenance of their performance is neglected (Abedinzadeh et al. 2013). In the Training domain, research studies related to academic education are conducted more than studies about public education and public awareness.

In the field of industrial and material sciences, research in areas like processed materials has fulfilled the researchers' needs more than research in raw materials and minerals. In these areas, the need for research on topics such as oil, heavy metals, and refinery impacts

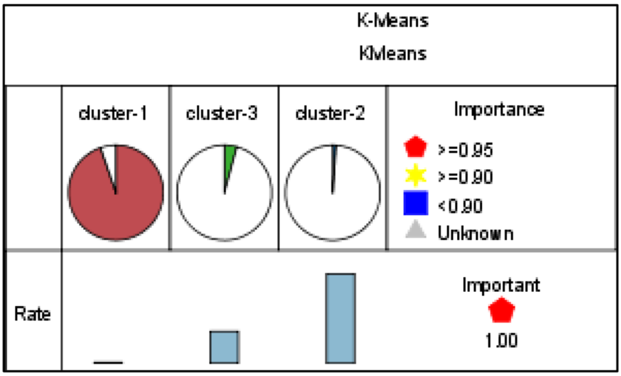


Fig. 4 N-grams clustering, k = 3, Rate is input variable

Table 5 Important phrases and subdomains of each domain in various rate level

Domains	High rate	Medium rate	Low rate
Technology	GEO-sensing, GIS Information	Bio-nanoparticles, nanostructured absorber	Nanotechnology, genetic engineering, bio-sensors, Pollution modeling
Forest	Forest biomass	Forest case study	Forest road, forest biodiversity, forest park
Industry	Environmental pollutants of steel industry, Textile factory, Extraction system	Steel industry, mulch construction, Electrical industry, Cement)	Landfill, factory, refinery, transportation
Geography	Tehran city, Urmia lake, Sirjan city	Bushehr city, Hamedan, Kurdistan province, Laar	Iran, Tehran, Industrial Zone, Air pollution in Tehran
Water	The drying of the Lake, Water emulsion, Irrigation regime	The deposition of water, the Zayandehroud river, The Bahmanshir river, Water network	Water pollution, river, watershed management, fisheries
Material	Processed products: (Iranian fuel materials, The new fuel, Polymer)	Plumbum, The oxide particles, Biological monitoring of metals	Raw materials and minerals: (oil pollution, heavy metals, organic fertilizer, oil)
Assessment	Project environmental risk assessment	Mines environmental pollution Assessment, Synthesis of compounds, Environment assessment	Environmental impact assessment, environmental assessment, economic evaluation, performance evaluation
Biodiversity	Vegetation ecological system, The pattern of biodiversity	Barley performance, Habitat zoning, Habitat mapping, Planet seed	Habitat suitability, Plant biodiversity, Plant Sciences
Soil	–	Erosion control, soil pollution, Arsenic soil	Soil pollution, Heavy metals soil, purification of soil, contaminated soil
Management	Management system effectiveness, Control management, Approaches, Environmental management, Safety management	Management planning, Management system, design, productivity, Midterm management: (metropolitan air pollution management, performance management plan), Protection	Risk management, Human resource management, Management pattern, Crisis management, Long-term management: (Strategic Planning, Development management)
Sewage	Sewer District 3	Industrial wastewater, Farms' sewage	Wastewater, garbage, Treatment plants, landfill
Economy	Macroeconomics: (Iran economic, export, economic distribution)	The stock market, Technical economy: (cost function, macro-economic variables, resource accounting, Cost criteria)	Investment, tax, Macroeconomic: (economic growth, trade, international trade) Microeconomics: (environmental accounting, environmental economics, demand, cost)

Table 5 continued

Domains	High rate	Medium rate	Low rate
Training	Formal education: (students, educational animation, educational priorities, year of study, teachers, high school students, university, research Institute)	Safety training, Formal education: (Environment department, Research's' results, Islamic Azad University)	Environmental education, education program, Training pattern, cognitive learning, Public education: (environmental awareness, knowledge-synergy, knowledge, awareness, education and participation, attitude, Life culture)
Pollution	Pollution reduction impacts, salt pollution, check emissions	Noise Measurement, levels of air pollution, Sources of pollution	Air pollution, noise, environmental pollution, Destruction
Law	–	Legal status, Legal resources: (Fiqh and law, Reviews the basics of Fiqh, International law, human rights, environmental laws)	International law, the laws of Iran, Legal analysis Administrative law: (Crime, damages, civil liability, environmental law)
Rangeland	–	–	Rangeland department, Rangeland ecology, Rangeland plants

on the environment is more than that of subjects like steel, polymers and new fuels. Obviously, it is clear that there are research gaps in the environmental law. And existing research is not sufficient for research society's demands. The demand for research in administrative issues, such as environmental crimes or environmental damages is considerable.

Managing missing data

As mentioned above, the data set contains users' search logs from February 2013 to July 2015. Unfortunately logs of June 2013 were not saved for unknown reasons. To compensate for this gap, in the cases that analyzing trends or comparing the growth or decline of searches were needed, the average count of May and July 2013 was used for June 2013. In the next section, we will use text mining models to specify the status of each area from various perspectives comprehensively. Table 6 describes the variables used in the following section.

Modeling, evaluation and results

Manual tagging of the data is an expensive and time-consuming task. Therefore, clustering, which is considered as an important tool for data analysis, can be used for this purpose. In this phase partitional and hierarchical clustering techniques were applied to analyze the

Table 6 Variables description

Variable	Description
Query	Text that searched by user
N-gram	Contiguous sequence of N words in the queries (In this paper n = 1, 2, 3, 4.)
All Freq, All-Freq, Freq	The incidence of one n-gram in queries
MaxCountAnswer, Ans	The maximum number of search results for a particular n-gram
Domain	An area of environmental science that n-grams have been assigned to.
Coverage	Ratio of the number of n-grams in one domain to the number of all n-grams
Rate	Ratio of All-Freq to MaxCountAnswer for each n-gram

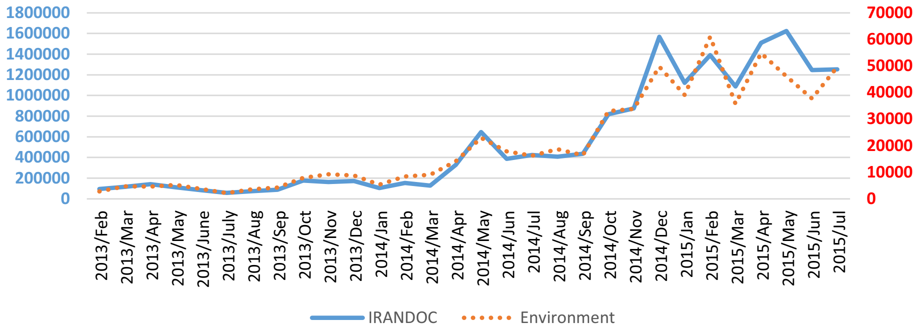


Fig. 5 Number of search queries in IRANDOC repository and Environmental science from Feb 2013 to Jul 2015

users' queries in the environment scope. These techniques are described in the next section. In this study, Clementine 12.0 and R packages were used as text mining tools.

Research trend analysis in environmental science

To assess and forecast the research trends in environmental sciences, the search queries in all domains have to be compared with the total search queries in the database. Figure 5 represents the total IRANDOC and environmental search queries from February 2013 to July 2015.

The searching pattern of IRANDOC users in each year varies with the events of the respective academic year. It is at its peak in May and September, the months which include students' deadlines for submitting their research proposals and theses. Conversely, in the summer, search queries are reduced as it is the students' holiday season. As Fig. 5 demonstrates, IRANDOC's total searches and environmental search queries have the same parallel changes. So it can be concluded that the environmental search pattern follows the overall search changes. The number of searches in the database has become tenfold in a period of 30 months. The main reason for this considerable rise is the change in IRANDOC's access policies in November 2014. According to the new policies, users no longer had to register and login to the website to be able to search the database. Moreover, the increase in the number of graduate students in Iran as well as software improvements are the other reasons for this significant growth.

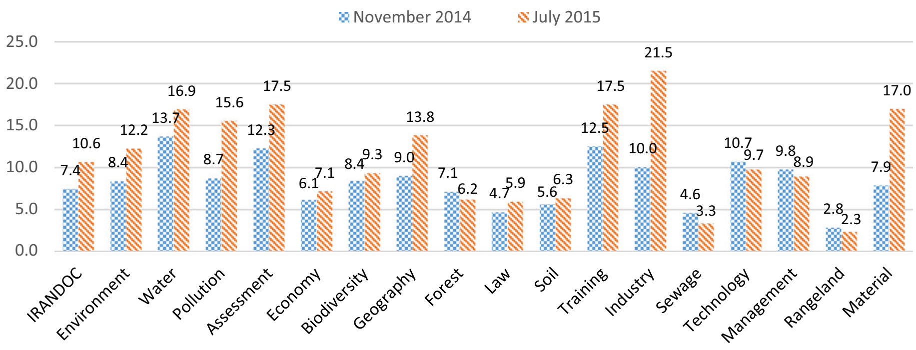


Fig. 6 The growth of the environment and its domains searches from Feb 2013 to Jul 2015

Figure 6 illustrates the changes in search trends for all scopes of environmental science and total search queries in IRANDOC from February 2013 to July 2015. These changes are compared with their respective average number in the first three months and the figure shows the magnitude of the changes. The changes are presented in two milestones. The first one is November 2014 the month in which access policies were changed and the second milestone is July 2015 which contains the last series of data that were used in this research.

As Fig. 6 shows, the total searches in the IRANDOC database has grown 7.4 times from February 2013 to November 2014. This increment was 8.4 times for the environment, and it was between 6 and 9 times for domains such as Pollution, Economy, Biodiversity, Forest and Material. Areas of Water, Assessment, Geography, Training, Industry and Technology domains were more considered by researchers. On the contrary, Law, Soil, Sewage and Rangeland were increased with lower rates and researchers had less interest in them.

In the period of Feb 2013 to Jul 2015, the amount of searches on the website has risen 10.6 times. This growth was 12.2 times the initial amount for the Environment area. And the growth of the Industry, Training, Assessment, Material, Water and Pollution was between 15 and 21 times their initial amounts. On the other hand, the search queries in Rangeland, Sewage, Soil, Forest and Economy domains increased less than 7 times. According to the observations above, it can be concluded that the environment is one of the hot topics for researchers as its search queries growth in the last month is higher than the average growth in total searches in all topics in the Iranian research society.

Domain's rate hierarchical clustering

A hierarchical clustering yields a dendrogram representing the nested grouping of objects. The clustering is performed by merging the most similar objects in the cluster set to form a bigger one (Bouguettaya et al. 2015). In the next step of data modelling and evaluation, the records were clustered by using a hierarchical clustering technique. Data can be clustered hierarchically in various groups by using this technique (Han et al. 2011). The purpose is to categorize the environmental domains in various levels based on their Rate to determine research priorities from the researchers' perspective. This method helps decision makers and research policy makers to divide environmental domains into some branches at each level based on their Rates. One of the important benefits of this method for managers is that it helps them allocate their limited resources to specific domains purposefully. The hierarchical clustering of domains leads to a dendrogram presented in Fig. 7.

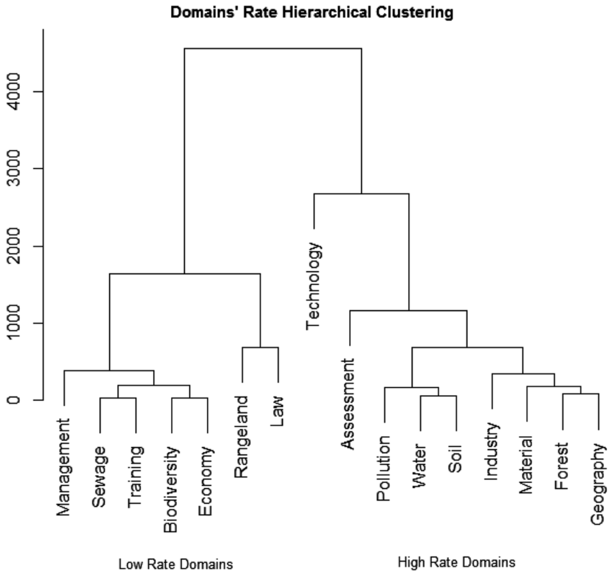


Fig. 7 The dendrogram of domains' rate hierarchical clustering

As illustrated in Fig. 7, the left side of dendrogram includes the low Rate areas of the environment. Rangeland, Law and Management have the lowest Rate. Each couple of areas are joint based on Rate similarity. (E.g. Rangeland and Law, which have the lowest Rate, are joined to each other). Domains located in this category need more research to compensate for the lack of the low number of answers per each search. In contrast, the right side of dendrogram presents the high Rate areas. Technology, Assessment and Industry have the highest Rate. This category contains queries that have considerable results. Therefore, it has a lower level of priority for more research compared with the previous category. As Fig. 7 shows, the descriptive features of this method help us know the Rate, priority, and status of a new topic just by knowing the domain to which it belongs.

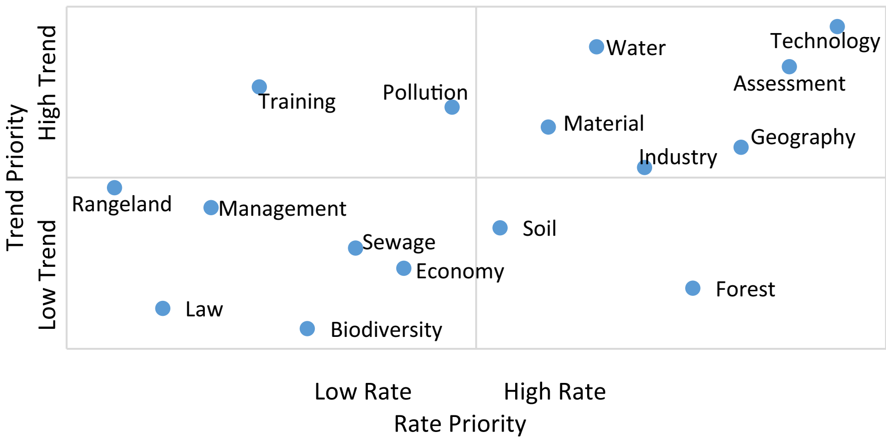


Fig. 8 Domains' research priorities based on trend and rate

The combination of the results of Rate clustering and information of search growth trends yields Fig. 8 which determines the research priority of each domain based on their Rate and Trend.

According to Fig. 8, the domains are divided into four priority zones. The right-bottom side of Fig. 8 contains Forest and Soil domains and indicate the high Rate- low Trend zone. These domains are not considered very much by users and have enough research works to provide users with sufficient search results. Therefore, the domains of this zone have the least priority for future research. In contrast, the left-upper side of the figure, shows the low Rate- high Trend zone. This zone illustrates that the Pollution and Training domains, which belong to it, have maximum priority based on users' trends and researchers' need for more results for their queries. The right-upper side presents the high Rate-high Trend research areas. The domains which belong to this part have been searched a lot by users and a large amount of research is done in these domains. Finally, the last zone is low Rate- Low Trend zone and it consists of domains which have a low Rate, which indicates the necessity of research in these areas, but they are not considered as much by researchers.

Deployment and managerial application

The last step of CRISP-DM is the deployment. In what follows, the development strategy is determined and the monitoring and control levels are explained (Shearer 2000). Although the deployment and exploring the methods and implementation of previous results are not in the scope of this paper, we propose a monitoring system as RDSM which help research managers and policy makers to apply the results of this study.

Developing a DSS based on the results of the previous steps is useful for managers and policy makers. Such DSS can help us observe the users' behavior and identify their needs and interests in specific research areas such as the environment. The RDSM can suggest the research priorities to managers and policy makers in science and technology so they can direct the researchers to focus on these research areas.

For the implementation of RDSM the following steps are recommended:

- Standardization of user input data
- Determining the domains and sub-domains
- Deploying the text mining tools as an embedded module of the database
- Controlling, monitoring and performance measurement

Following above steps and using the same framework to determine research priorities can help managers in various domains.

Managers of research repositories, who can access the information about research studies, could determine the research priorities by monitoring the information seekers behaviors and analyzing the research trends and gaps by using RDSM. Scientific centers and journals' editorial boards can choose important topics which have been neglected and label them as research gap in their special issues. This strategy motivates researchers to do more research on these topics. The resource allocation is one of the most important activities of research policy makers. They can do this by being aware of the research priorities. The RDSM also helps them to make decisions in order to fund special research projects, find and support special start-ups by tracking research trends. This monitoring system allows managers of universities and research centers to encourage students and researchers to focus on research gaps and direct them to do applied research.

Discussion

The first question of this study (RQ1) is looking for the status of research demands in the environmental science in Iran. The trend analysis shows that the amount of searches on the website has risen 10.6 times from Feb 2013 to Jul 2015. And the growth of the Industry, Training, Assessment, Material, Water and Pollution was between 15 and 21 times. This change is understandable since in the recent years, climate changes have caused frequent droughts, floods have adversely damaged water resources in Iran (Faramarzi et al. 2009; Gorjian and Ghobadian 2015) and dust and sand storms in the west of the country have affected the environment (Abrishamchi 2013). On the other hand, the search queries in Rangeland, Sewage, Soil, Forest and Economy domains were increased less than 7 times. The "Research Trend Analysis in Environmental Science" sectionillustrates these results in details.

To address RQ2 and identify the research gaps in the environmental science, a clustering techniques is applied. This technique determines some subdomains of environmental science which include research gaps. For instance, areas of the administrative law (crime, damages, civil liability, and environmental law) have a large research gap. The previous research works about environmental awareness in Iran verify that academic education increases the environmental awareness of students. Although this awareness is acceptable in universities students, it is low in under graduated people (Salehi and Pazoki Nejad, Environment in Higher Education: the Evaluation of Environmental Awareness in the Mazandaran Students 2013; Salehi, Environmental Behavior and Education 2012). Tables 4 and 5 in "Descriptive Analysis of the Users' Queries" section present the research gaps in each domain and subdomain of the environmental studies in Iran.

Finally, the main question of this study tries to determine the environmental research priorities in Iran. To answer RQ3, the environmental research priorities are categorized in four priority zones. The high Rate- low Trend zone consists of Forest and Soil, which are not hot subjects from users' perspective and have considerable research works to respond the users' requests. In contrast, the low Rate- high Trend zone illustrates that the Training and Pollution have the maximum amount of users' trends and researchers need more results for their requests. This finding is supported by Hemmati and Shabiri's research, which have compared the situation of the Iranian environmental education with the other countries and concluded that the academic and public education in this scope should be considered more seriously (Hemmati and Shobeiri 2016). The third zone presents the high Rate-high Trend research areas. The areas of Material, Industry, Geography, Assessment, Water and Technology are in this zone. These areas have been searched a lot and a large amount of research has been done in these domains. Finally, Economy, Sewage, Management, Law, Biodiversity and Rangeland are located in low Rate- Low Trend zone. The previous research works about solid waste-related research in the last 30 years period in Iran include the same results and shows a considerable drop in the number of publications on these topics and it has predicted that a downward trend in the number of publications is likely to be continued in future years (Mesdaghinia et al. 2015). Furthermore, the citation analysis of global articles in the science and social citation index shows that environmental law research was given an increasing amount of attention over the period of 1992–2014 and developing countries such as China, Brazil and India are among the top 10 productive countries in the environmental law studies (Ma and Ho 2013). These zones are presented in Fig. 8.

Conclusion

According to the results obtained from clustering the environmental phrases, the majority of searches have a small frequency and a small number of returning answers or search results. Moreover, when the search frequency rises, the number of answers increases too. But, it is not an immutable fact. For example, the areas of Rangeland, Law, Management and Biodiversity have a low ratio in the number of answers to the number of searches. In contrast, the areas of Technology and Assessment have considerable amounts of answers in comparison with the number of searches. So the environmental research policy makers should encourage researchers to do more research on the environmental law, environmental management and economic related topics.

Although the trend analysis shows that environmental search pattern generally follows the overall search changes, the amount of searches on the website has risen 10.6 times but the growth was 12.2 times for the Environment area. According to the observations above, it can be concluded that the environment is one of the hot topics for researchers since the growth in its search queries in the last months is higher than the average growth of the overall searches in the Iranian research society.

Applying the clustering technique on extracted environmental phrases leads to determine the research gaps. For instance, areas of administrative law, environmental strategy management, and public awareness are known as research gaps. Therefore, establishing university majors or courses in environmental science for administrative environmental law, environmental long-term management, environmental public awareness, the effects of raw materials and minerals in the environment are recommended.

The combination of trend analysis and clustering of queries shapes four priority zones. The high Rate- low Trend zone consists of Forest and Soil. This zone is marked as low priority research zone. In contrast, the low Rate- high Trend zone, which could be assigned as a high priority research zone, consists the Training and Pollution. The third zone is the high Rate-high trend research areas and Material, Industry, Geography, Assessment, Water and Technology are located in this zone. Policy makers could decide to direct researchers to focus on these areas based on other research prioritization factors. The last zone is low Rate- Low Trend and Economy, Sewage, Management, Law, Biodiversity and Rangeland are located in this zone. This zone contains topics with a low Rate, which indicates the necessity of research in these areas, but unfortunately they are not considered as much by researchers. Therefore, Iranian research managers and policy makers should encourage researchers to pay more attention to these areas.

An implementation of a DSS based on text mining in various fields can help managers, to take the right decisions, especially about the resource allocation (Kouziokas 2016). The results show that establishing specific laws can help to fill the research gap. For example, in Iran the managers of large-scale projects have to deliver the environmental evaluation report as an essential attachment to their project reports.

Using the RDSM framework to monitoring research demand and supply and determine research gaps and priorities can help managers of environmental research to direct their activities to fill the gaps and allow research managers and policy makers to efficiently allocate limited resource to environmental research areas. Moreover, the same method can be applied to other research domains.

The following subjects are recommended for the future research:

- Comparing the research production trends with the research demand trends and analyzing the research demand and supply.

- Using geographical information of users and studies and analyzing the geographical research gaps in the environmental science.
- Analyzing the users' clickstream information to observe their behavior in order to understand their interests in scientific databases.
- Monitoring and determining the research gaps in the other research areas using the proposed framework and process in this study.

The first limitation of this study is that some queries which contain just a subdomain of the environment science and have not any combination of the selected keywords in the query string are neglected in the data selection phase. Although, comparing the extracted domains with the LOC environmental classification shows that selected data covers all basic subdomains of the environmental science, some scarce subdomains may have been neglected in this study. The second limitation is that using previous studies to determine future priorities have been limited to retrospective analyses. But applying todays behavior of researchers could balance this limitation.

Acknowledgements The authors gratefully acknowledge the support of the Iranian Research Institute for Information Science and Technology (IRANDOC). And specially acknowledge the help of IRANDOC's R&D department.

Appendix

Table 7 Original Persian list of top 10 frequent N-grams in users' Queries

Uni-gram Terms			Bi-gram Terms		
No.	**Persian Term**	**English Term**	**No.**	**Persian Term**	**English Term**
1	زیست	Life	1	محیط زیست	Environment
2	محیط	Environment	2	زیست محیطی	Environmental
3	آلودگی	Pollution	3	آلودگي هوا	Air Pollution
4	محیطی	Environmental	4	زیست شناسي	Biology
5	زیستی	Biological	5	آلودگي آب	Water Pollution
6	هوا	Air	6	تنوع زیستي	Biodiversity
7	ارزیابی	Evaluation	7	حقوق محیط زیست	Environmental Law
8	آب	Water	8	آلودگي صوتي	Noise Pollution
9	قانون	Law	9	حفاظت محیط	Environment Protection
10	مدیریت	Management	10	آلودگي نفتي	Oil Pollution
Tri-gram Terms			Quad-gram Terms		
No.	**Persian Term**	**English Term**	**No.**	**Persian Term**	**English Term**
1	اثرات زیست	Environmental Impact	1	ارزیابي اثرات	Environmental Impact Assessment
2	حقوق محیط	Environmental Law	2	ارزیابان زیست	Hegmataneh Environmental Assessors
3	حفاظت محیط	Environmental Protection	3	انجمن ارزیابان	Environmental Assessors Community
4	آموزش محیط	Environmental Education	4	حقوق بین الملل	International Environmental Law
5	ارزیابي	Environmental Impact Assessment	5	دانشکده عمران	Department of Environmental Construction
6	آلودگي محیط	Environmental Pollution	6	دانشکده مرتع	Rangeland Watershed School of Fisheries
7	ارزیابان	Environmental Assessors	7	آبخیزداري شیلات	Watershed Fisheries Environmental
8	مدیریت زیست	Environmental Management	8	مرتع آبخیزداري	Rangeland Watershed Environment Fisheries
9	ارزیابي	Environmental Assessment	9	انجمن محیط زیست	Koomesh Environmental Community
10	مدیریت محیط	Environmental Management	10	ارزیابي ریسك	Environmental Risk Assessment

References

Abedinzadeh, N., Jamalzade Fallah, F., Pendashteh, A., Mokrem, R., Panahandeh, M., Moghadami, S., et al. (2013). Investigation of effectiveness of EMS establishment in improvement of environmental performance in industrial units accordance with Iso 14000 Standards. Rasht: SID. Retrieved from http://fa.projects.sid.ir/ViewPaper.aspx?ID=84034#.

Abrishamchi, A. (2013). Overview of key urban air pollution problems in Iran and its capital city, Tehran. In *Section 3. Case Studies on Specific Urban Areas: Understanding the Roles of Key Economic, Geographic, and Urban Design Inputs in the Pollution Characterization or Mitigation Scenarios 87* (pp. 11–18). Irvine, California.

Akçapınar, G. (2015). How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers and Education, 87*, 123–130.

Anwar, T., & Abulaish, M. (2014). A social graph based text mining framework for chat log investigation. *Digital Investigation, 11*(4), 349–362.

Beth, B., & Deyrup, M. M. (2015). The SHU research logs: Student online search behaviors trans-scripted. *The Journal of Academic Librarianship, 41*(5), 593–601.

Bijalwan, V., Kumar, V., Kumari, P., & Jordan, P. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application, 7*(1), 61–70.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications, 42*(5), 2785–2797.

Cavnar, W., & Trenkle, J. (1994). N-gram-based text categorization. *Ann Arbor MI, 48113*(2), 161–175.

Chen, L., Mao, K., Zheng, Y., Zhou, X., & Zhu, C. (2012). Research on mining association rules in university scientific projects management. *Communications in Computer and Information Science, 345*, 561–567.

Choubin, B., Khalighi-Sigaroodi, S., Malekian, A., Ahmad, S., & Attarod, P. (2014). Drought forecasting in a semi-arid watershed using climate signals: a neuro-fuzzy modeling approach. *Journal of Mountain Science, 11*(6), 1593–1605.

Choudhary, A., Oluikpe, P., Harding, J., & Carrillo, P. (2009). The needs and benefits of text mining applications on post-project reviews. *Computers in Industry, 60*(9), 728–740.

Claes, J., & Poels, G. (2014). Merging event logs for process mining: A rule based merging method and rule suggestion algorithm. *Expert Systems with Applications, 41*(16), 7291–7306.

Davide, B., Rosso, P., Gómez-Soriano, J., & Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems, 34*(2), 113–134.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications, 34*(3), 1707–1720.

Erdmann, M., Ikeda, K., Ishizaki, H., Hattori, G., & Takishima, Y. (2014). Feature based sentiment analysis of tweets in multiple languages. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakal, & Y. Zhang (Eds.), *Web information systems engineering—WISE 2014* (pp. 109–124). Thessaloniki, Greece: Springer.

Faramarzi, M., Abbaspour, K. C., Schulin, R., & Yang, H. (2009). Modelling blue and green water resources availability in Iran. *Hydrological Processes, 23*(3), 486.

Fronza, I., Sillitti, A., Succi, G., Terho, M., & Vlasenko, J. (2013). Failure prediction based on log files using random indexing and support vector machines. *Journal of Systems and Software, 86*(1), 2–11.

Gadkari, N., Savio Raj, S., & Raka, H. (2015). Query subtopic mining from search log data. *International Journal of Current Engineering and Technology, 5*(3), 2058–2062.

Gorjian, S., & Ghobadian, B. (2015). Solar desalination: A sustainable solution to water crisis in Iran. *Renewable and Sustainable Energy Reviews, 48*, 571–584.

Gu, X., & Blackmore, K. (2016). Recent trends in academic journal growth. *Scientometrics, 108*(2), 693–716. doi:10.1007/s11192-016-1985-3.

Günes, E., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.

Gunton, T. (2002). Establishing environmental priorities for the 21st century: Results from an expert survey methodology. *Environments, 30*(1), 71–98.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Boston: The Morgan Kaufmann Series in Data Management Systems.

Hemmati, Z., & Shobeiri, S. M. (2016). Review the status of environmental education in Iran and comparison with other countries. *Journal of Human and Environment, 14*(2), 61–81.

Houškova, B., & Houška, M. (2011). Data, information and knowledge in agricultural decision-making. *Agris on-line Papers in Economics and Informatics, 3*(2), 74–82.

Hsin-Chang, Y., & Lee, C.-H. (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems with Applications, 29*(4), 723–734.

IranDoc. (2016, 11 14). *IrandDoc Thesauri*. (IranDoc) Retrieved 11 14, 2016, from IranDoc: http://thesauri.irandoc.ac.ir/.

IranDoc. (2016, 11 14). *IRANDOC Information Reposiory*. Retrieved 11 14, 2016, from http://irandoc.ac.ir/db/databases-about.html.

IranDoc. (2016, 11 14). *Iranian Scientific Repository*. Retrieved from IRANDOC: http://ganj.irandoc.ac.ir/.

Jalalimanesh, A. (2012). Knowledge discovery in scientific databases using text mining and social network analysis. In *Control, Systems and Industrial Informatics (ICCSII)* (pp. 46–49). IEEE.

Julien, H., Pecoskie, J., & Reed, K. (2011). Trends in information behavior research, 1999–2008: A content analysis. *Library and Information Science Research, 33*(1), 19–24.

Kademani, B., Sagar, A., Surwase, G., & Bhanumurthy, K. (2013). Publication trends in materials science: A global perspective. *Scientometrics, 94*(3), 1275–1295.

Khosravi, M., & Jamali, H. R. (2014). Log analysis of the IRANDOC database and the analysis of its users' information seeking behavior. *Iranian Journal of Information Processing and Management, 29*(4), 979–1006. Retrieved from http://jipm.irandoc.ac.ir/article-1-2444-fa.html.

Kim, M., & Chen, C. (2015). A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics, 104*(1), 239–263.

Kirkland, J. (2010). The management of university research. In P. A. Peterson (Ed.), *International Encyclopedia of Education (Third Edition)* (Third Edition ed., pp. 316–321). Oxford: Elsevier. doi:10.1016/B978-0-08-044894-7.00877-0.

Kolehmainen, M., Martikainen, H., Hiltunen, T., & Ruuskanen, J. (2011). Forecasting air quality parameters using hybrid neural network modelling. In *International Conference on Urban Air Quality: Measurement, Modelling and Management. 65* (pp. 277–286). Madrid: Springer Science and Business Media.

Kouziokas, G. (2016). Technology-based management of environmental organizations using an Environmental Management Information System (EMIS): Design and development. *Environmental Technology and Innovation, 5*, 106–116.

Kusiak, A., Verma, A., & Wei, X. (2013). A data-mining approach to predict influent quality. *Environmental Monitoring and Assessment, 185*(3), 2197–2210.

Library of Congress Collections Policy Statements. (2016). Retrieved from LIBRARY OF CONGRESS: https://www.loc.gov/acq/devpol/environ.pdf.

Lin, H.-C., Hong, Y.-M., & Kan, Y.-C. (2012). The backend design of an environmental monitoring system upon real-time prediction of groundwater level fluctuation under the hillslope. *Environmental Monitoring and Assessment, 184*(1), 381–395.

Liu, K., Li, X., Shi, X., & Wang, S. (2008). Monitoring mangrove forest changes using remote sensing and GIS data with decision-tree learning. *Wetlands, 28*(2), 336–346.

Losiewicz, P., Oard, D., & Kostoff, R. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems, 15*(2), 99–119.

Lu, G., & Eldin, N. (2014). Employers' expectations: A probabilistic text mining model. *Procedia Engineering, 85*, 175–182.

Ma, R., & Ho, Y.-S. (2013). Comparison of environmental laws publications in science citation index expanded and social science index: A bibliometric analysis. *Scientometrics, 109*(1), 227–239. doi:10.1007/s11192-016-2010-6.

Marino, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., & Costa-Jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics, 32*(4), 527–549.

Mesdaghinia, A., Mahvi, A., Nasseri, S., Nodehi, R., & Hadi, M. (2015). A bibliometric analysis on the solid waste-related research from 1982 to 2013 in Iran. *International Journal of Recycling of Organic Waste in Agriculture, 4*(3), 185–195.

Munková, D., Munk, M., & Vozár, M. (2013). Data pre-processing evaluation for text mining: Transaction/sequence model. *Procedia Computer Science, 18*, 1198–1207.

Nadjla, H., & Sahar, M. (2014). Search strategies in nanotechnology databases: Log analysis. *Iranian Journal of Information Processing and Management, 29*(1), 233–252. Retrieved from http://jipm.irandoc.ac.ir/article-1-2192-fa.html.

Nicholas, D., Huntington, P., & Jamali, H. (2008). User diversity: As demonstrated by deep log analysis. *The Electronic Library, 26*(1), 21–38.

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications, 42*(9), 4348–4360.

Oberreuter, G., & Velásquez, J. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40*(9), 3756–3763.

Salehi, S. (2012). Environmental behavior and education. *Journal of Education, 18*(2), 201–226. Retrieved from http://education.scu.ac.ir/article_10133.html.

Salehi, S., & Pazoki Nejad, Z. (2013). Environment in higher education: The evaluation of environmental awareness in the mazandaran students. *Educational Planning Studies, 2*(4), 199–220. Retrieved from http://www.eps.journals.umz.ac.ir/?_action=articleInfo&article=764#.

Saneifar, H., Bonniol, S., Poncelet, P., & Roche, M. (2014). Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback. *Computers in Industry, 65*(6), 937–951.

Shachi, M., & Jaiswal, U. (2014). Resolving issues in parsing technique in machine translation from hindi language to english language. In *International Conference on Computer and Communication Technology (ICCCT)* (pp. 55–58). Allahabad: IEEE.

Shearer, C. (2000). The CRISP-DM Model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.

Souza, F. (2014). A data-based model to locate mass movements triggered by seismic events in Sichuan, China. *Environmental Monitoring and Assessment, 186*(1), 575–587.

Sphinx. (2001). *Open Source Search Server*. Retrieved 11 14, 2016, from Sphinx Search: http://sphinxsearch.com/docs/current/extended-syntax.html.

Sunikka, A., & Bragge, J. (2012). Applying text-mining to personalization and customization research literature—Who, what and where? *Expert Systems with Applications, 39*, 10049–10058.

Tsai, H.-H. (2011). Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. *Scientometrics, 87*(3), 425–450.

Tu, Y.-N., & Seng, J.-L. (2009). Research intelligence involving information retrieval—An example of conferences and journals. *Expert Systems with Applications, 36*(10), 12151–12166.

Wang, G., Zhang, X., Tang, S., Zheng, H., & Zhao, B. (2016). Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 225–236). San Jose: ACM. doi:10.1145/2858036.2858107.