Using Text Mining Techniques for Extracting Information from Research Articles

Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan

Abstract Nowadays, research in text mining has become one of the widespread fields in analyzing natural language documents. The present study demonstrates a comprehensive overview about text mining and its current research status. As indicated in the literature, there is a limitation in addressing Information Extraction from research articles using Data Mining techniques. The synergy between them helps to discover different interesting text patterns in the retrieved articles. In our study, we collected, and textually analyzed through various text mining techniques, three hundred refereed journal articles in the field of mobile learning from six scientific databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge. The selection of the collected articles was based on the criteria that all these articles should incorporate mobile learning as the main component in the higher educational context. Experimental results indicated that Springer database represents the main source for research articles in the field of mobile education for the medical domain. Moreover, results where the similarity among topics could not be detected were due to either their interrelations or ambiguity in their meaning. Furthermore, findings showed that there was a booming increase in the number of

S.A. Salloum (✉) · K. Shaalan
Faculty of Engineering & IT, The British University in Dubai, Dubai, UAE
e-mail: ssalloum@uof.ac.ae

K. Shaalan
e-mail: Khaled.shaalan@buid.ac.ae

S.A. Salloum
University of Fujairah, Fujairah, UAE

M. Al-Emran
Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Gambang, Malaysia
e-mail: malemran@buc.edu.om

M. Al-Emran
Al Buraimi University College, Buraimi, Oman

A.A. Monem
Faculty of Computer & Information Sciences, Ain Shams University, Cairo, Egypt
e-mail: azza_monem@hotmail.com

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_18

published articles during the years 2015 through 2016. In addition, other implications and future perspectives are presented in the study.

Keywords Text mining · Information extraction · Topic identification Scientific databases · Mobile learning · Higher education

1 Introduction

Nowadays, almost all of the existing information in different institutions (e.g. government, business, industry, and others) is preserved in electronic documents in which it contains semi-structured data. In these documents, the "abstract" is an example of unstructured text component. Whereas, examples of structured fields in a document are: author's name, publication date, title, and category [1]. A study by [2] stated that text mining has become one of the trendy fields that has been incorporated in several research fields such as computational linguistics, Information Retrieval (IR) and data mining. Text mining is different from data mining [3]. Data mining is focused on discovering interesting patterns from large databases rather than textual information [4]. Information recovery methodologies like text indexing techniques have been developed for handling unstructured documents. In conventional researches, it is assumed that a user mostly searches for known terms, which have been previously used or written by someone else. The main problem is that the search results are not relevant to the user's requirements. One solution is to use text mining in order to find out relevant information, which is not indicated explicitly nor written down so far. The procedure of text mining begins with gathering documents through different resources. A particular document would be recovered through text mining instrument and by checking its format and character sets; it will be pre-processed by this instrument. The document would then pass through a text analysis stage. Text analysis includes semantic analysis intended to obtain high-quality information through text. Different text analysis methods are available. Different methods can be used based on the organization's objective. In some cases, text analysis methods are repeated until information is extracted. The outcomes can be stored in a management information system that provides a large amount of significant information for the user of that system.

Text mining intends to detect the information that was not recognized before through extracting it automatically from various text-based sources. Structured data can be handled through data mining tools while unstructured or semi-structured datasets like full-text documents, emails, and HTML files can be handled through text mining. Typically, the information will be kept in a natural form known as text. Text mining is not similar to web mining. When something is explored on the web by the user, it means that it is previously known and it was written by someone else [5]. For example, in E-commerce, a major issue with web mining is buying all the materials which are not relevant to the user's search and it will not show unknown

(hidden or implicit) information, while the major objective of text mining is to find out the unknown information [6]; something that is not recognized by anyone.

Data is the basic kind of information, which is required to be organized and mined for the knowledge generation. Discovering patterns and trends from huge data is a significant challenge. Finding out the unknown trends and patterns from databases properly is a major objective of data mining. It is a method where data pre-processing is necessary before applying any other method. Many approaches like clustering, classification, and decision trees are involved in data mining. All the textual based information is stored by electronic means, either on a client's personal computers or on a web server. Due to the increasing growth in hardware storage devices, any computer or laptop has the ability to store an enormous amount of data. Creating new information can be simple while finding out relevant information from a huge amount of data is challenging. In order to extract the relevant information, knowledge, or patterns from various sources that are in unstructured form, text mining technique can be employed. The common structure of text mining involves two consecutive stages: text refining and knowledge distillation. In text refining, free-form text documents are converted into an intermediate form, whereas in knowledge distillation, patterns or knowledge are derived from intermediate form. Intermediate form (IF) can be either semi-structured like the theoretical graph illustration or structured like the relational data illustration. IF can be either a document-based where every entity symbolizes document, it can be a concept-based where every unit symbolizes an object or a concept of interest in a particular area.

Various research areas, techniques, and models are involved in different research domains. The hottest topics of the research domains are the primary focus of many research papers. The research results of a particular domain may influence other research domains since some research domains may have similar topics. These research topics always discuss such a promising research area that is worth studying. Therefore, the trend of cross domain is determined in this research. The longitudinal trends of academic articles in Mobile Leaning (ML) were explored in this research with the help of text mining methods. We recovered and examined (300) refereed journal articles and conference proceedings from various authentic databases.

The primary goals of this research are (1) Using text mining techniques for identifying the topics of a scientific text related to ML research and developing a hierarchical and evolutionary connection among these topics. (2) Using visualization tools for presenting both the topics and the association among them as a convenient way to help users to determine relevant topics.

This paper is categorized as follows: Sect. 2 provides an inclusive background concerning in the text mining field. Other related studies are addressed by Sect. 3. Research methodology is presented in Sect. 4. The results are demonstrated in Sect. 5. Conclusion and future perspectives are presented in Sect. 6.

2 Background on Text Mining and Information Extraction

2.1 Text Mining

The development in the fields of web, digital libraries, technical documentation, medical data has made it easier to access a larger amount of a textual documents, which come together to develop useful data resources [7]. Therefore, it makes text mining (TM) or the knowledge discovery from textual databases a challenging task owing to meet the standards of the depth of natural language which is employed by most of the available documents. The available textual information in the form of databases and online sources [7–9] raises a question about who is responsible for keeping a check on the data and analyzing it? Keeping in view the pertaining condition, it is not possible to analyze and effectively extract the useful information manually. There is a need to employ software solutions which may employ automatic tools for analyzing a considerable amount of textual material, extract relevant data, analyze relevant data, and organize relevant information. Owing to the increasing demands to obtain knowledge from a large number of textual documents accessible on the web, text mining is gaining a significant importance in research [10, 11]. Generally, text mining and data mining are considered similar to one another, with a perception that same techniques may be employed in both concepts to mine text [4, 12, 13], and [3]. However, both are different in a sense that data mining involves structured data, while text deals with certain features and is relatively unstructured and usually require preprocessing. Furthermore, text mining is an interrelated field with Natural Language Processing (NLP). NLP is one of the hot topics that is concerned with the interrelation among the huge amount of unstructured available text [14], besides the analysis and interpretation of human-being languages [15, 16].

2.2 Information Extraction

An initiation point for computers to evaluate unstructured manuscripts is to use Information Extraction (IE). IE software recognizes key phrases and relationships included in the manuscript. This is performed through finding the predefined arrangements in a text; this technique is called pattern matching. Regular language text documents consist of information that cannot be utilized for mining. IE agrees with the documentation, choosing appropriate articles, and the association among them to make them more available for added guidance [17, 18]. Contrary to Information Retrieval, which deals with how to recognize relevant documents from a document collection, IE yields structured information prepared for post-processing, which is essential to various applications of Web mining and searching instruments [19]. IE deals with discovering and extracting important

information from natural language texts [18]. It consists of separating appropriate text parts, extracting the offered data in such parts, and transforming the data into the functional form. Fractional extraction from domain-particular texts is currently possible; though complete IE from the random text is still a continuing study target [20].

2.3 Extracting Knowledge from Text

Under most of the conditions, only specific data is obtained from the information extracted from unstructured text instead of abstract knowledge. In such a case, it is required to employ a text mining task along with additional techniques to mine knowledge from the data in hand [21, 22]. DiscoTEX (Discovery from Text EXtraction) is one of the major approaches employed for text mining. It involves using IE first to gather structured data from unstructured text, followed by employing traditional Knowledge Discovery from Database (KDD) tools to discover knowledge from this data. This framework for text mining was presented by [21]. In this method, the learned IE system is used to convert unstructured text into more structured data. This data is then subjected to mining to develop meaningful relationships. In a case that the information extracted from a corpus of documents is in the form of abstract knowledge instead of concrete data, IE tends to serve as the "discovering knowledge" from text. Discovery of knowledge by extracting information, such as key-phrases or keywords extraction from the text may be used for other text mining tasks, i.e. classification, clustering, summarization, and topic detection [23].

2.4 Text Mining Methods and Techniques

Text mining is usually employed to obtain quick results [24]; it has been subjected research under a number of application areas. On the basis of respective areas of application, text mining can be categorized as text categorization, text clustering, association rule extraction, and text visualization. They are discussed in the following sub-sections.

Text Clustering

Text clustering is based on the Cluster hypothesis which proposes that relevant documents must have more similarities with one another than the non-relevant ones [25]. The Clustering technique is a trust-worthy technique that is generally employed for analyzing larger amounts of data like data mining. It has been proven that text clustering is one of the most effective tools used for text theme analysis [26]. Moreover, it facilitates the method of topic analysis in which named entities having concurrent occurrence are grouped together, followed by subjecting them to

the clustering process in such a way that frequent item are placed in sets by applying the hyper graph-based method [27]. Each set of named entities is represented by a cluster that is related to one of the ongoing topics in the corpus. The process of topic tracking within dynamic text data has gained the interest from the researchers who are working on the subject of text clustering in the digital field. Various methods and algorithms based on unsupervised document management are included in the process of document clustering. In the clustering process, the numbers, properties, and associations of the grouped sets are initially unknown. The grouping of documents is performed by categorizing them into a particular category such as medical, financial, and/or legal [28].

Association Rule Extraction

A study by [29] argued that the method of association rule mining (ARM) is employed to identify relationships within a larger group of variables in a dataset. The ARM identifies the variable-value combinations which tend to occur frequently. The method of ARM in data mining also known as knowledge discovery in databases; that is similar to the correlation analysis that finds out the relationships between two variables. Wong et al. [30] provided that the Association Rules for Text Mining are majorly concerned to explore the relationships between various topics or factual notions employed for characterizing a corpus. They intend to discover key association rules relative to a corpus in such a way that the occurrence of certain topics in an article may correspond to the occurrence of another topic as well.

K-Means Algorithms

The *k*-mean approach divides the data set into *k* clusters, where every cluster is subjected to be represented by the mean of points; called the centroid. A two-step repetitive process is employed for the application of the algorithm: (1) Assigning every point to the nearest centroid. (2) Evaluating the centroids for a recently developed group. The process is ended when the cluster centroid comes to a constant value. The *k*-mean algorithm has an extensive application owing to its direct parallelization. Furthermore, the order of respective data does not affect the *k*-mean algorithm which attributes the numerical characteristics to it. It is required to mention the maximum value of *k* at the beginning of the process. The representation of the cluster is made by the *k*-medoid algorithm that chooses the object adjoining the center of the cluster. Though, the selection of the *k* objects is done randomly in the algorithm. The selected objects help to determine the distance. A cluster is formed on the basis of the nearest object to *k*, whereas the other objects acquire the position of *k* recursively till the required quality of the cluster is achieved [28].

Information Visualization

Information visualization puts great textual bases in a visual hierarchy or plan and offers browsing abilities as well as general searching. This technique offers improved and quicker comprehensive knowledge, which assists us to mine enormous accumulation documents. The operators can distinguish the colors, associations, and gaps. The assortment of documents can be demonstrated as a structured layout utilizing indexing or vector space model.

Word Cloud

Jayashankar and Sridaran [31] defined word clouds or tag clouds as the visual representation of words for a certain written content structured as per its frequency. Word cloud is among the most frequently used method to present text data in a graphical manner; making it helpful for analyzing various forms of text data such as essays and short answers or written opinions to a survey or questionnaire [32]. Word cloud tends to serve as a preliminary stage for in-depth analysis of certain text material [33, 34]. For example, word cloud assists in finding the relevancy between given text and the required information. Nonetheless, the method has certain drawbacks as well. One of the major drawbacks that is it does not consider the linguistic knowledge about the words and their respective link to the given subject while providing a purely statistical summary to the segregated words. As a result, in most systems, the word clouds are often employed in a statistical manner for summarizing text, providing very little or no means for correlating the data. It is perceived that this could be one of the most influencing paradigms of visualization for most of the analysis conditions. Thus, in this paper, we have employed the use of word clouds as the central method to text analysis.

3 Related Work

Many research works contributed to the field of IE through the use of various techniques. The primary focus of these researches was to determine how different text mining procedures can be utilized as the structured data sets exist in the text document format. This part begins with defining the topic of the research, evaluating previous researches, and then major techniques are applied using information extraction and text mining. In order to determine the topic of each research area and to develop an evolutionary and hierarchical connection between these topics, [35] used the method of text mining. Topics are presented through visualization tools. Moreover, these tools are used in order to show the connection between these topics and to offer interactive functions so that users can effectively find the cross-domain topics and know the trends of cross-domain research.

Moloshnikov et al. [36] developed an algorithm for finding documents on a particular topic depending on a selected reference collection of documents. In addition, the context-semantic graph for visualization themes in search results was also developed. The algorithm depends on the incorporation of a group of entropic, probabilistic and semantic developers for mining of weighted keywords and set of words that explain the specified topic. Results indicated that the average precision is 99% and the recall is 84%. A unique technique was also created for making graphs on the basis of the algorithm, can remove key phrases with weights. It offers the opportunity to show an arrangement of sub-topics in huge sets of documents in compact graph form.

In order to offer a reference for additional researches of other researchers, [37] discussed the research status of text mining technology when it was used in the

biomedical field that covers 10 years. Biomedical text mining literature incorporated in SCI from 2004 to 2013 were recovered, filtered and then examined from the viewpoint of research institutions, yearly changes, research areas, local distribution, journals sources, and keywords. A prominent increase in the amount of worldwide biomedical text mining literature is observed. Among this global literature, a huge percentage is taken up by literature related to named entity recognition, entity relation extraction, text categorization, text clustering, abbreviations extraction, and co-occurrence analysis. Studies carried out in USA and UK are considered to be present in the primary position.

In order to extract inter-language clusters through multilingual documents depending on Closed Concepts Mining and vector model, a new statistical approach was suggested by [38]. Formal Concept Analysis methods are used for mining Closed Concepts from similar corpora and later these Closed Concepts and vector models are utilized in the clustering and arrangement of multilingual documents. An experimental assessment is carried out over a set of French-English bilingual documents of CLEF's 2003. With a notable comparability score and in order to remove the bilingual classes of documents, results revealed that the interaction between vector model and Formal Concept Analysis is very useful.

Santosh [39] suggested the graph mining-based document content (i.e. text fields) exploitation. That is, the query generated the graph depending on the users' requirements. This is an easy and effective graph mining method to extract similar patterns through the documents and changed the query graph into model graphs which are utilized when the users are not present. An intelligent solution for document information exploitation has been created. This is characterized by simplicity, ease of use, accuracy, ease of development, and flexibility. In order to understand graph models, it does not need a huge collection of document images. Moreover, since model learning consumes less than 10 s for an input pattern per class on average, changes, amendments, and replacements can be done in the input patterns. Information exploitation average performance is shown to have 86.64% as Precision, and 90.80% as a Recall. However, the suggested technique failed to offer inclusive and accurate solutions for the patterns that have a huge collection of fields in a zigzags arrangement due to the query graph intricacy.

Sirsat et al. [23] proposed two techniques for mining text through online sources. The first technique dealt with the knowledge that is required to be shown directly in the documents that need to be mined. Text mining and IE are considered as the only effective tools for performing that technique. The second one concerned with the documents that hold an actual data in unstructured format instead of nonfigurative knowledge. IE can help to change the unstructured data presented in the document corpus into structured one. In order to discover nonfigurative patterns in the extracted data, data mining algorithms and techniques can be used.

Song and Kim [40] presented the first attempt to apply text mining approaches to a huge collection of full-text articles for discovering the knowledge structure of the area. Instead of depending on the citation data presented in Web of Science, PubMed Central full-text articles have been used for bibliometric examination. Above all, this assisted the creation of text mining routines in order to develop a

custom-made citation database following the full-text mining. Findings showed that most of the documents that were published in bioinformatics area were not cited by others. Additionally, a constant and linear rise has been observed in the amount of publications across publication years. Results revealed that the majority of the retrieved studies were inspired by USA-based institutes followed by European institutes. Results reported that the major primary focus of the important topics was on biological factors. However, according to PageRank, the top 10 articles were highly concerned with the computational factors.

In order to facilitate the accurate extraction of text from PDF files of research articles that can be utilized in text mining applications, a "Layout-Aware PDF Text Extraction" (LA-PDF Text) system was presented by [41]. Text blocks are mined from PDF-formatted fill-text research articles under this system and then the system categorizes them into logical units depending on rules that typify particular sections. Only the textual content of the research articles is focused in the LA-PDF Text system. This system serves as a basis for new experiments into more developed extraction methods dealing with multi-modal content like images and graphs. The system goes through three phases: (1) Identifying contiguous text blocks with the help of spatial layout processing in order to discover blocks of contiguous text, (2) Categorization of text blocks into metaphorical categories with the help of a rule-based method, and (3) Joining categorized text blocks together by arranging them accurately which results in the extraction of text from section-wise grouped blocks. An evaluation of the accuracy of the block discovery algorithm used in step 2 was performed. It was also shown that the system can identify and classify them into metaphorical categories with Recall = 0.89%, Precision = 0.96%, and F = 0.91%. Moreover, the accuracy of the text mined with the help of LA-PDF Text is compared to the text from an Open Access subset of PubMed Central. This accuracy is then compared with the text that was mined using the PDF2Text system. These are the two frequently used techniques to extract text from PDF.

Mooney and Bunescu [42] described two techniques for using the natural language information extraction for text mining. First, general knowledge can be mined directly from the text. A project where a knowledge base of 6580 human protein interactions was extracted by mining around 750,000 Medline abstracts in which reconsidered as an example of this technique. Second, structured data can be mined through text documents or web pages. In order to find out patterns in the mined data, traditional KDD methods can be applied. The performed work on the DiscoTEX system and its application to Amazon book descriptions, computer science job postings, and resumes were considered as an example of this technique. In order to discover units and relations in text, research in IE keeps on creating more efficient algorithms. Valuable and significant knowledge can be mined effectively from the constantly developing body of electronic documents and web pages by using modern approaches in human language technology and computational linguistics, and linking them with the modern techniques used in machine learning and conventional data mining techniques. IE deals with determining a particular set of relevant items through natural language documents.

In order to discover topics that recur in articles of text corpus, another method TopCat (Topic Categories) was proposed by [22]. IE was used by this technique in order to discover named entities in individual articles and to characterize them as a collection of items of an article. Therefore, through recognition of frequent item sets which commonly occurred with named entities, the issues in data mining or database context were studied. Association rule data mining technique is used by TopCat to discover these frequent item sets. By using a hypergraph splitting technique, TopCat further clusters the named entities which discovers a collection of frequent item sets with significant overlie. In order to discover documents regarding the topic, IR technique was used. Different technologies like IE for named entity extraction, association rule data mining, clustering of association rules, IR techniques were used in this method. TopCat discovers topics that have a logical accuracy with reasonable identifiers. Callan and Mitamura [43] presented a new technique for named entity detection, called KENE. In order to understand the extraction rules, the knowledge-based technique is used in it. Generate-and-test approach is used for named entity extraction from structured documents.

We can observe from the surveyed literature that there is a limitation in addressing the issue of IE from research articles using data mining techniques. The synergy between these approaches (i.e. IE with data mining techniques) helps to discover different interesting text patterns in the retrieved articles. This approach could be applied to a variety of research topics, where in each topic can generate a wide range of knowledge patterns. Mobile learning (M-learning) has become one of the trendy fields in the higher education [44–50]. In accordance to the existing literature, we can perceive that IE and data mining techniques were never applied to the M-learning field. This creates a need for collecting several research articles in the field of M-learning from different scientific databases and applies the synergic approach on them. Additionally, we are trying to respond to the following research questions:

RQ1: What are the most frequent keywords in the collected articles?
RQ2: What are the most frequent terms among the collected articles?
RQ3: What are the most common topics among the collected articles?
RQ4: How are the articles interrelated to each other?
RQ5: How are the articles distributed in terms of publication year?

4 Research Methodology

4.1 Text Mining Processing Framework

We have developed our customized framework which is inspired by the designed framework proposed by [51], see Fig. 1. Three steps are included in text mining: text pre-processing, text mining operations, and post processing. Text pre-processing involves the following tasks: data selection, classification, feature extraction and text

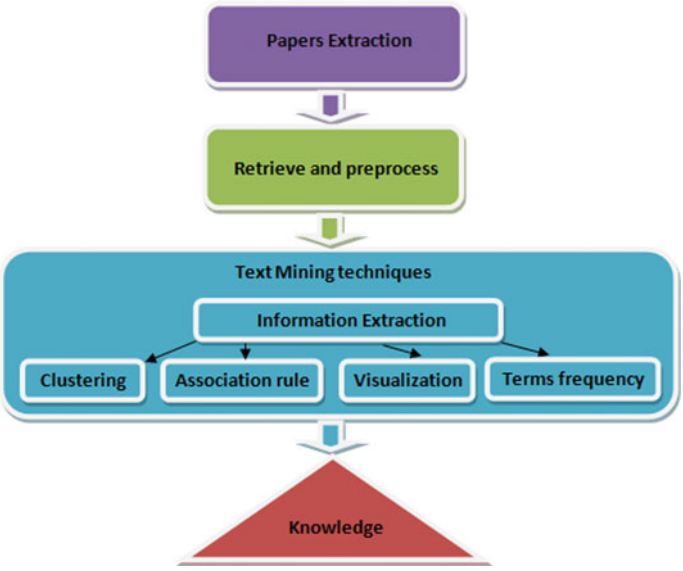


Fig. 1 Text mining processing framework

normalization, i.e. transforming the documents into an intermediate form for ensuring compatibility for various mining tools. The second step deals with different text mining techniques like clustering, association rule detection, visualization, and terms frequency. During the third step, alterations and changes are made on the data (i.e. research articles) through text mining functions like evaluation and choice of knowledge, analysis and visualization of knowledge. The main aim of this study is to extract interesting information from the collected articles using the text mining techniques.

4.2 Data Collection and Pre-processing

The research articles were collected from six scientific databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge. The search term used for data collection is simply "Mobile Learning in higher education". Based on that, 300 research articles in the field of mobile learning were collected. These articles are categorized into six folders, where each folder represents the database where these articles were retrieved.

The presence of the linguistic noise is a common problem in the content of the extracted articles and we have dealt with. Then, the cleaned data are uploaded into RapidMiner tool while the misplaced and unnecessary data have been removed from the dataset. In order to improve the performance and data quality, all the irrelevant characteristics are debarred while the data is being uploaded into

RapidMiner tool. The major steps involve the separation of the document into tokens; this task is called Tokenization [52]. The next step is concerned with the transformation process of all the characters where each document title is created in a lower case. Stop words filtering is involved in the third step, where English is filtered through this operator. A single English word is required to be signified by each token. All tokens that were similar to stop words were eradicated from the provided document by an operator. The document must have only one stop word per line. The last step is concerned with the text processing phase that involves filtering the tokens according to the length. The minimum number of the characters that the token should have is 4, while the maximum number is 25 characters.

5 Experimental Results

The application of various text mining techniques on the collected articles presents different results and suggestions. In the present study, we are trying to apply almost all of the text mining techniques that were mentioned in the literature on the collected articles. Nevertheless, these techniques have not been applied to the research articles concerning mobile learning in higher education; the reason that makes this study is unique and adds a value to the research community.

Q1: What are the most frequent keywords in the collected articles?

As per the study of [31], we used the cloud technique in order to answer the above research question. As shown in Fig. 2, we can notice that "Learning" is the most keyword that was mentioned across all the collected articles. The second highest frequent words are "Patients" and "Students" respectively. The increasing number of the words (learning and students) could be attributed to the fact that learning and students form the core of the higher educational processes. In addition, the appearance of the word (Patients) in many articles shows that most of these articles are focusing on mobile learning in medical education. Table 1 shows the distribution of the top 5 most frequent terms that were mentioned in the collected articles across each database.

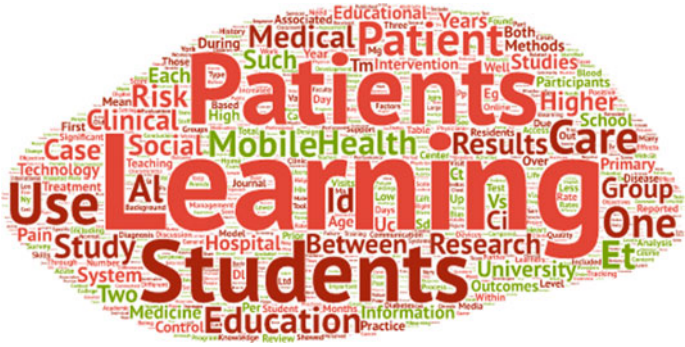


Fig. 2 Words cloud across all databases

Table 1 Words cloud terms distribution across all databases

Scientific database	Term	Frequency
Cambridge	Learning	1165
	Education	854
	University	847
	Students	831
	Higher	490
IEEE	Education	793
	Students	777
	Learning	579
	Engineering	417
	Higher	256
Science Direct	Learn	2043
	Use	837
	Mobile	701
	Education	584
	Student	551
Springer	Patients	9046
	Care	6458
	Learning	5423
	Medical	5180
	Health	4086
SAGE	Learning	2033
	Students	1719
	Mobile	1611
	Education	1112
	University	921
Wiley	Learning	7556
	Patients	6392
	Students	3266
	Care	3212
	Mobile	3193

Moreover, we have applied the word frequency technique on the text in the collected articles. As per (Fig. 3), we can notice that the most frequent linked words among all the articles are: "Learning" followed by "Patients", "Students", "Education", "Care", "Mobile", "Study", "University", "Medical", and "Clinical" respectively. These results indicate that the most frequent linked words are focused on studies targeting mobile learning in medical education. These results match the above mentioned results in terms of the word cloud. Springer database represents the most source that contains these words followed by Wiley and Science Direct respectively. The results reveal that the words (patients, care, medical, and clinical) were highly mentioned in Springer database. That is, researchers who are specialized in mobile medical education should benefit from these results as it shows

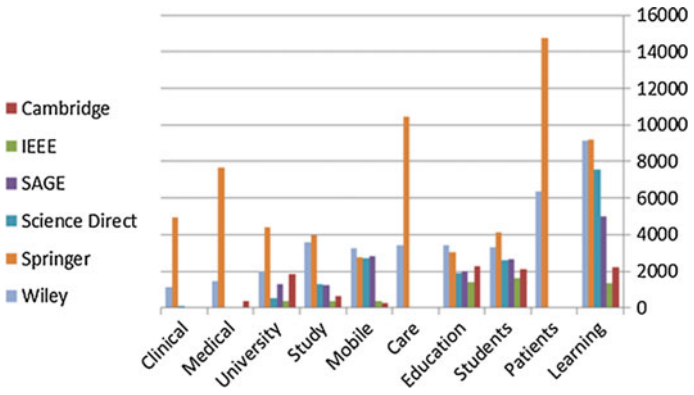


Fig. 3 Word frequency distribution across all databases

them that Springer is the top among other databases for collecting research articles in that field.

For further investigation, the above most frequent words were analyzed and distributed among all the Scientific Databases in order to represent each word separately. As per (Fig. 4), the word "Learning" was frequently mentioned by Springer database followed by Wiley, Science Direct, SAGE, IEEE, and Cambridge, respectively.

According to (Fig. 5), the word "Patients" was frequently used by Springer database followed by Wiley. This indicates that almost all of the mobile medical education articles are published under springer and Wiley databases.

As per (Fig. 6), the word "Students" was frequently utilized by Springer database followed by Wiley, SAGE, Science Direct, Cambridge, and IEEE, respectively.

According to (Fig. 7), the word "Education" was frequently reported by Wiley followed by Springer, Cambridge, SAGE, Science Direct, and IEEE, respectively.

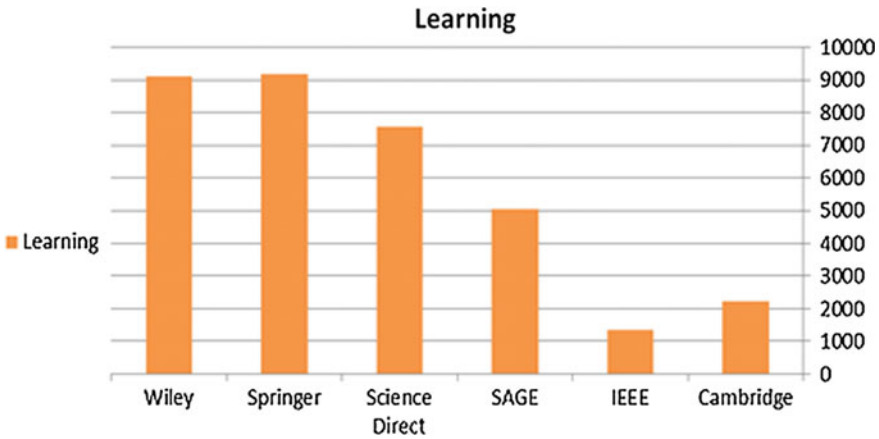


Fig. 4 The distribution of the word "Learning" among all sources

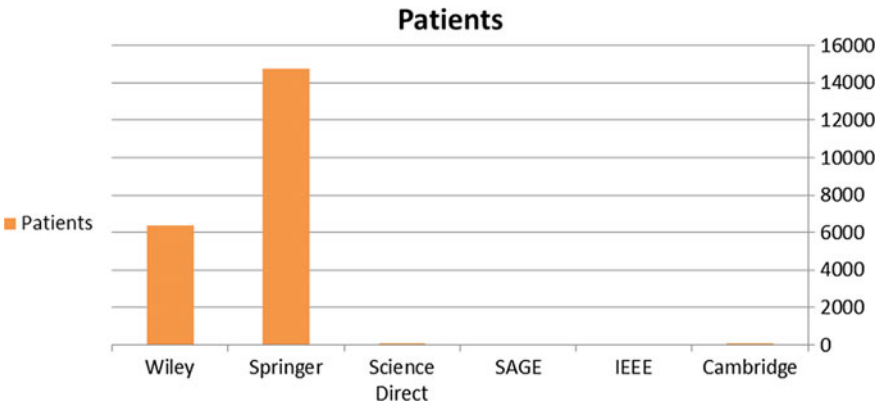


Fig. 5 The distribution of the word "Patients" among all sources

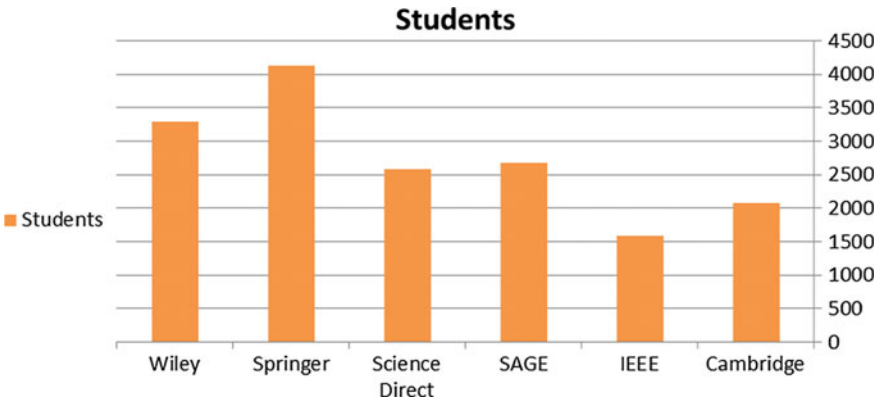


Fig. 6 The distribution of the word "Students" among all sources

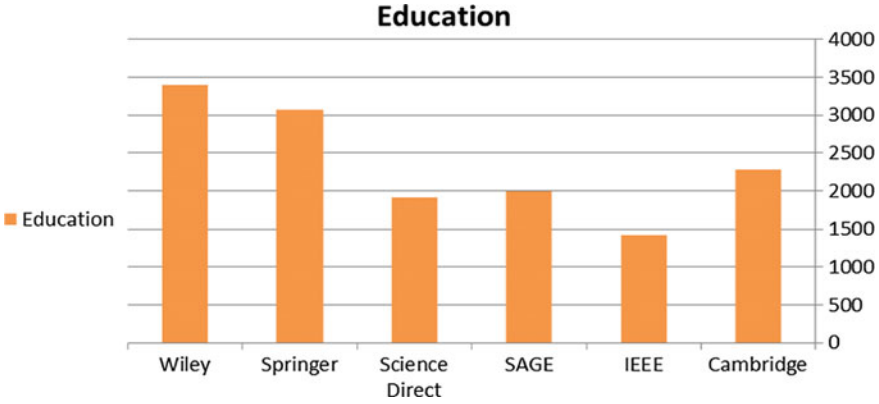


Fig. 7 The distribution of the word "Education" among all sources

According to (Fig. 8), the word "Care" was frequently mentioned by Springer database followed by Wiley. On the other side, other databases don't show the occurrence of this word. These results assist the researchers of mobile medical education that their field research articles are mainly available in Springer and Wiley, while other databases don't have enough journals that accommodate these articles.

As per (Fig. 9), the word "Mobile" was frequently used by Wiley followed by SAGE, Springer, Science Direct, IEEE and Cambridge, respectively.

As per (Fig. 10), the word "study" was frequently occurred in Springer database followed by Wiley, Science Direct, SAGE, Cambridge, and IEEE, respectively.

According to (Fig. 11), the word "University" was frequently mentioned in Springer database followed by Wiley, Cambridge, SAGE, Science Direct, and IEEE, respectively.

As per (Fig. 12), the word "Medical" was frequently used by Springer database followed by Wiley, Cambridge, and Science Direct respectively.

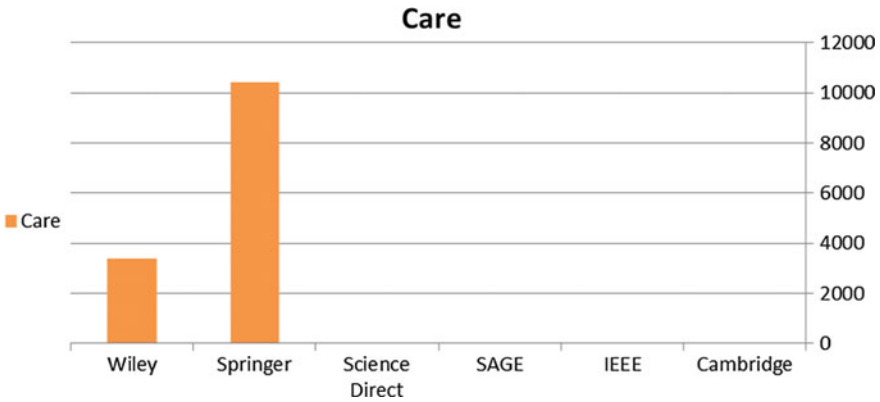


Fig. 8 The distribution of the word "Care" among all sources

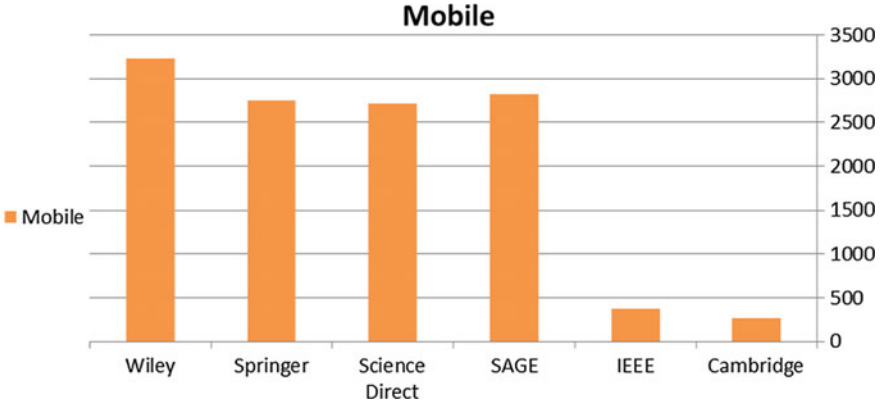


Fig. 9 The distribution of the word "Mobile" among all sources

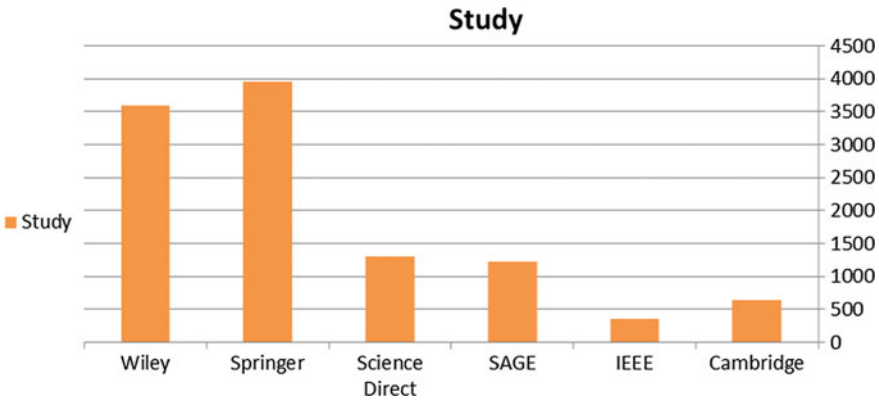


Fig. 10 The distribution of the word "Study" among all sources

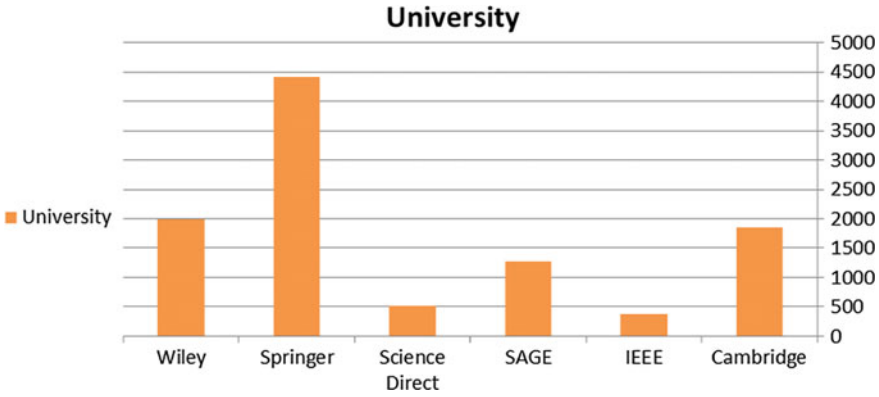


Fig. 11 The distribution of the word "University" among all sources

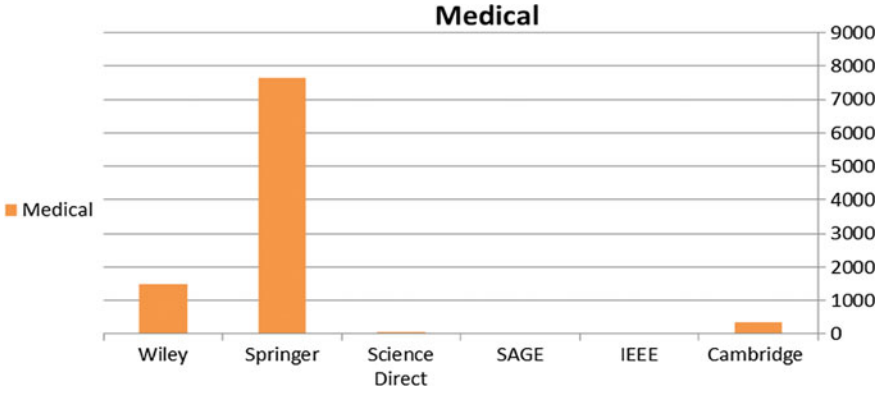


Fig. 12 The distribution of the word "Medical" among all sources

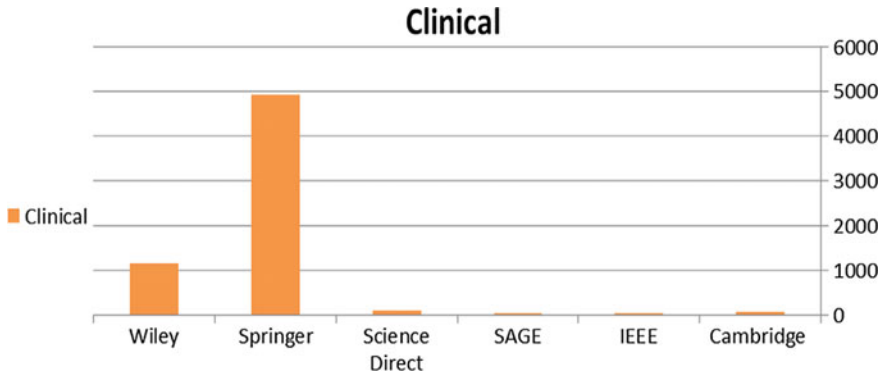


Fig. 13 The distribution of the word "Clinical" among all sources

According to (Fig. 13), the word "Clinical" was frequently occurred in Springer database followed by Wiley, Science Direct, Cambridge, IEEE, and SAGE, respectively.

Q2: What are the most frequent terms among the collected articles?

As per the study of [22], the method of the association rule is employed to identify and visualize the terms that have strong connections to each other. The most connected terms are termed as being strongly related to each other. According to (Fig. 14), the term "Education" is shown as being central to the tree structure having all the relevant words connected to it. This could be referred to the fact that the text acquired from the collected research articles is mainly concentrated on the learning field.

Q3: What are the most common topics among the collected articles?

As per the study of [25], we performed the similarity measure on the collected articles in order to identify the topics that are highly similar to each other. Figure 15 shows the similarity relationships among all the articles. As we can observe from the figure, it is very difficult to track the relationships among all the depicted topics. This could be attributed to the fact that all the collected articles are in one research field (i.e. mobile learning in higher education). To this end, the similarity operator could not detect a clear similarity among the topics since all these topics are interrelated and similar in meaning to each other.

Q4: How are the articles interrelated to each other?

According to [53] and [28], we applied the clustering technique in order to answer the above research question. We used the k-means algorithm through the use of different k values. By examining different k values, we end up with ($k = 6$) as it represents the most reasonable value for answering the above question. As per (Fig. 16), there are six clusters. Cluster 0 contains 3 items (i.e. 3 articles), cluster 1 includes 2 items, cluster 2 contains 3 items, cluster 3 includes 5 items, cluster 4

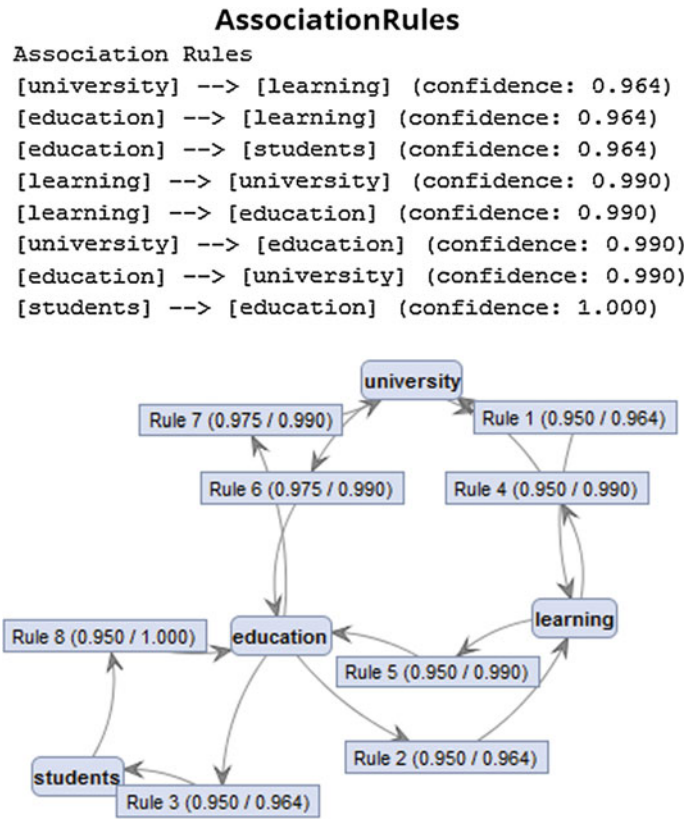


Fig. 14 Concept link diagram

includes 285 items, and cluster 5 includes 2 items. To that end, almost all of the articles (N = 285) are accumulated in cluster 4; this indicates that these articles are discussing the main studied topic (i.e. mobile learning in higher education). On the other side, by further investigating the remaining articles (N = 15) that are accumulated in the other clusters; it has been found that these articles are discussing other topics in learning and education rather than the studied topic. In addition, the reason that brought these articles in the search results when we collect them is that these articles include the word "mobile" as just a cited term in the text.

Q5: How are the articles distributed in terms of publication year?

Figure 17 shows the distribution of the collected research papers across their years of publication. The collected papers (N = 300) were published between years 1990 through 2016. It is clearly demonstrated that the mobile learning field was not common in the early 1990s. In 2009, there is a noticeable increase in the number of published articles as mobile learning becomes very popular during that period.

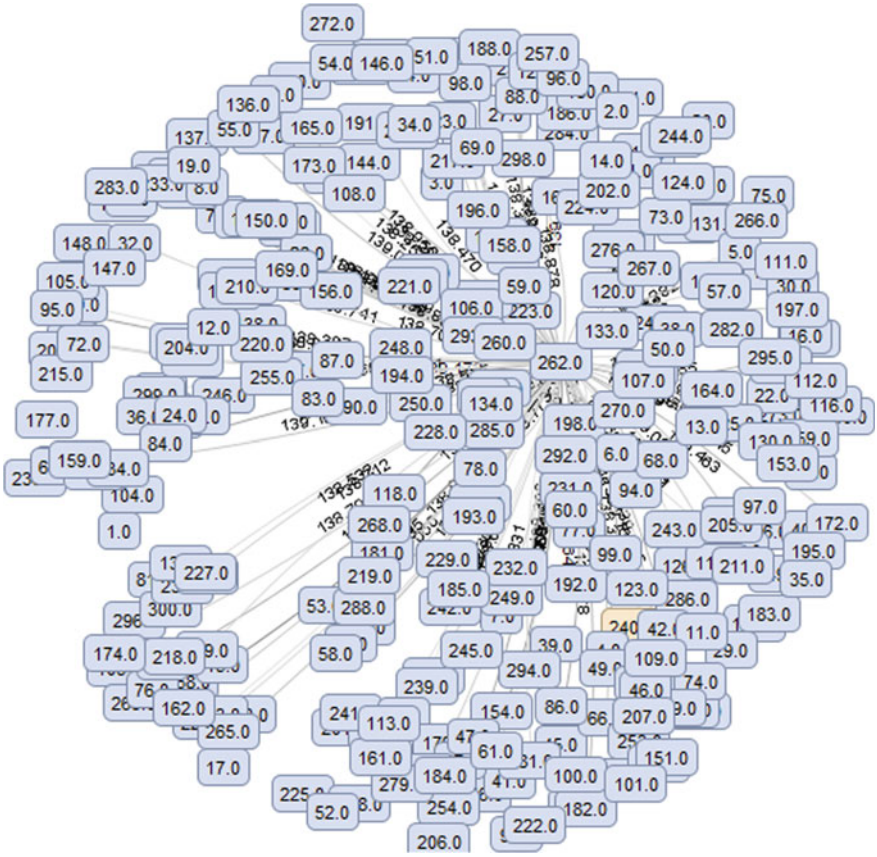
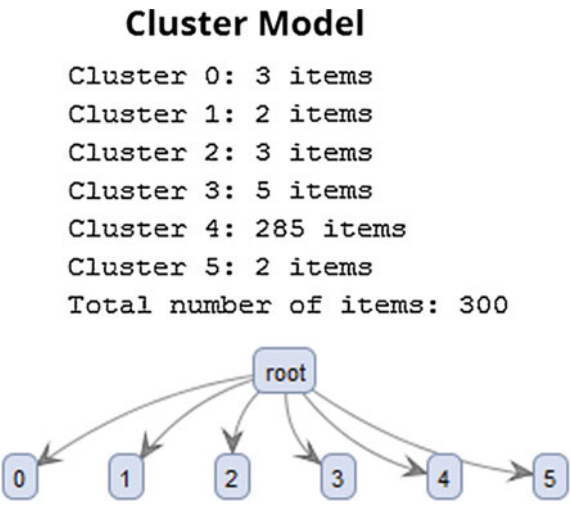


Fig. 15 Similarity diagram

Fig. 16 Cluster model



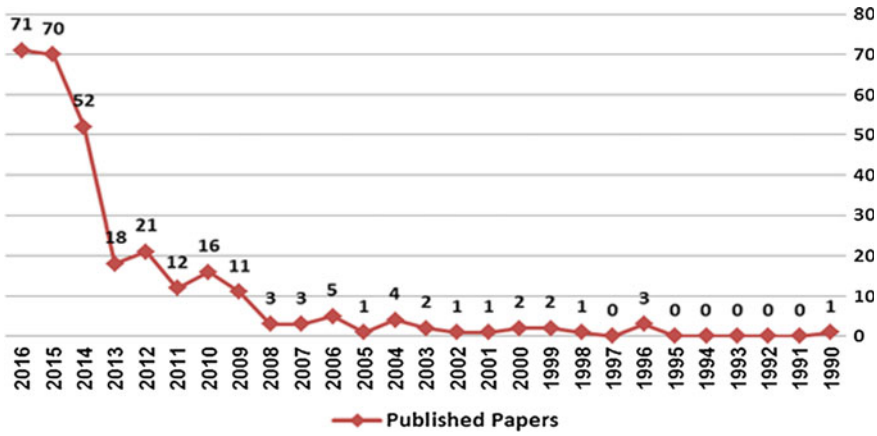


Fig. 17 Distribution of research articles in terms of publication year

Furthermore, during the years 2015 through 2016, mobile learning witnessed an enormous attraction from a lot of scholars who published many articles that contribute to the evolvement of mobile learning.

6 Conclusion

The present study demonstrates a comprehensive overview about text mining and its current research status. According to the surveyed literature, there is a limitation in discussing the issue of information extraction from research articles using data mining techniques. The synergy between information extraction and data mining techniques helps to discover different interesting text patterns in the retrieved articles. This approach could be applied to a variety of research topics, where in each topic it can generate a wide range of knowledge patterns. Mobile learning has become one of the trendy fields in the higher education. Accordingly, we can perceive that information extraction and data mining techniques were never applied to the mobile learning field. This creates a need for collecting a dataset that consists of several research articles in the field of mobile learning from different scientific databases, and applying the proposed approach on them.

Three hundred refereed journal articles from six scientific databases were collected, and textually analyzed through text mining techniques. The six databases are Science Direct, IEEE, Wiley, Cambridge, SAGE, and Springer. The selection of the collected articles was based on the criteria that all these articles should incorporate mobile learning as the main component in the higher educational context. In the present study, text clustering, association rule, word cloud, and word frequency are the main tasks used for text analysis.

By applying the word cloud and the word frequency techniques, results indicated that “Learning” is the most frequent keyword across all the collected articles;

followed by “Patients” and “Students”, respectively. The increasing number of the words: “learning” and “students” could be attributed to the fact that learning and students form the core of the higher educational processes. In addition, results revealed that the words: “patients”, “care”, “medical”, and “clinical”, were frequently mentioned in Springer database. These results indicate that the most frequent linked words are those focused on studies targeting mobile learning in medical education. Springer database represents the richest source that contains these words followed by Wiley and Science Direct, respectively. That is, researchers who are specialized in mobile medical education should benefit from these results as it shows them that Springer database is the topmost among other databases for finding research articles in this field.

By applying the association rule technique, findings showed that the term “Education” is shown as being central to the tree structure having all the relevant words connected to it. This could be referred to the fact that the text acquired from the collected research articles is mainly concentrated on the learning field. In addition, we performed the similarity measure on the collected articles in order to identify the topics that are highly similar to each other. Results revealed that the similarity operator could not detect a clear similarity among some topics the reason is that these topics are interrelated and similar in meaning to each other (i.e. all the articles are discussing the topic of mobile learning in higher education).

By applying the clustering technique, we used the *k*-means algorithm through the use of different *k* values. Results indicated that there were six clusters. Almost all of the articles (N = 285) were accumulated in one cluster; this indicates that these articles are discussing the main studied topic (i.e. mobile learning in higher education). On the other side, by further investigating the remaining articles (N = 15) that are accumulated in the other clusters; it has been found that these articles are discussing other topics in learning and education rather than the studied topic. By distributing the collected research papers across their years of publication, findings showed that there was a booming increase in the number of published articles during the years 2015 through 2016. This could be referred to the reason that mobile learning has witnessed in these years an enormous attraction from a lot of scholars who published many articles that contribute to the evolvement of mobile learning.

As a future work, we are interested in collecting articles from various research topics, i.e. not to focus on one area. This will help us to find more interesting patterns in these articles and how such articles are distributed among the targeted databases. In addition, this will allow the similarity operator to work properly and to draw a clear relationship among the articles.

References

1. Gaikwad, S.V., Chaugule, A., Patil, P.: Text mining methods and techniques. Int. J. Comput. Appl. **85**(17) (2014)
2. Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K.: A Survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. (2017)

3. Navathe, S.B., Ramez, E.: Data warehousing and data mining. Fundam. Database Syst., 841–872 (2000)
4. Gupta, V., Lehal, G.S.: A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell. **1**(1), 60–76 (2009)
5. Gupta, S., Kaiser, G.E., Grimm, P., Chiang, M.F., Starren, J.: Automating content extraction of html documents. World Wide Web **8**(2), 179–224 (2005)
6. Hassani, H., Huang, X., Silva, E.S., Ghodsi, M.: A review of data mining applications in crime. Statistical Anal. Data Min.: ASA Data Sci. J. **9**(3), 139–154 (2016)
7. Feldman, R., Dagan, I.: Knowledge discovery in textual databases (KDT). KDD **95**, 112–117 (1995)
8. Tan, A.H.: Text mining: The state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, vol. 8, pp. 65–70 (1999)
9. Hearst, M.A.: Untangling text data mining. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 3–10. Association for Computational Linguistics (1999)
10. Rajman, M., Besançon, R.: Text mining: natural language techniques and text mining applications. In: Data Mining and Reverse Engineering, pp. 50–64. Springer, US (1998)
11. Mahgoub, H., Rösner, D., Ismail, N., Torkey, F.: A text mining technique using association rules extraction. Int. J. Computat. Intell. **4**(1), 21–28 (2008)
12. Akilan, A.: Text mining: challenges and future directions. In: 2015 2nd International Conference on Electronics and Communication Systems (ICECS), pp. 1679–1684. IEEE (2015)
13. Sukanya, M., Biruntha, S.: Techniques on text mining. In: 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 269–271. IEEE (2012)
14. Salloum, S.A., Al-Emran, M., Shaalan, K.: A Survey of lexical functional grammar in the Arabic context. Int. J. Com. Net. Tech. **4**(3) (2016)
15. Al Emran, M., Shaalan, K.: A survey of intelligent language tutoring systems. In: 2014 International Conference on Advances in Computing, Communications and Informatics ICACCI, pp. 393–399. IEEE (2014a)
16. Al-Emran, M., Zaza, S., Shaalan, K.: Parsing modern standard Arabic using Treebank resources. In: 2015 International Conference on Information and Communication Technology Research (ICTRC), pp. 80–83. IEEE (2015)
17. Pazienza, M.T. (Ed.): Information extraction: Towards scalable, adaptable systems. Springer (2003)
18. Cowie, J., Lehnert, W.: Information extraction. Commun. ACM **39**(1), 80–91 (1996)
19. Velasco-Elizondo, P., Marín-Piña, R., Vazquez-Reyes, S., Mora-Soto, A., Mejia, J.: Knowledge representation and information extraction for analysing architectural patterns. Sci. Comput. Program. **121**, 176–189 (2016)
20. Hsu, J.Y.J., Yih, W.T.: Template-based information mining from HTML documents. In: AAAI/IAAI, pp. 256–262 (1997)
21. Mooney, R.J., Nahm, U.Y.: Text mining with information extraction, multilingualism and electronic language management. In: Proceedings 4th International MIDP Colloquium, pp. 141–160 (2003)
22. Clifton, C., Cooley, R., Rennie, J.: TopCat: data mining for topic identification in a text corpus. IEEE Trans. Knowl. Data Eng. **16**(8), 949–964 (2004)
23. Sirsat, S.R., Chavan, D.V., Deshpande, D.S.P.: Mining knowledge from text repositories using information extraction: A review. Sadhana **39**(1), 53–62 (2014)
24. Madani, F.: Technology Mining bibliometrics analysis: applying network analysis and cluster analysis. Scientometrics **105**(1), 323–335 (2015)
25. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56 (2008)

26. Clifton, C., Cooley, R.: TopCat: Data mining for topic identification in a text corpus. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 174–183. Springer, Heidelberg (1999)
27. Han, E.H., Karypis, G., Kumar, V., Mobasher, B.: Clustering based on association rule hypergraphs. In: DMKD (1997)
28. Irfan, R., King, C.K., Grages, D., Ewen, S., Khan, S.U., Madani, S.A., … & Tziritas, N.: A survey on text mining in social networks. Knowl. Eng. Rev. **30**(2), 157–170 (2015)
29. Goh, D.H., Ang, R.P.: An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents. Behav. Res. Methods **39**(2), 259–266 (2007)
30. Wong, P.C., Whitney, P., Thomas, J.: Visualizing association rules for text mining. In: 1999 IEEE Symposium on Information Visualization, 1999. (Info Vis' 99) Proceedings, pp. 120–123. IEEE (1999)
31. Jayashankar, S., Sridaran, R.: Superlative model using word cloud for short answers evaluation in eLearning. Educ. Inf. Technol., 1–20 (2016)
32. DePaolo, C.A., Wilkinson, K.: Get your head into the clouds: using word clouds for analyzing qualitative assessment data. TechTrends **58**(3), 38–44 (2014)
33. Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? J. Inf. Sci. **34**(1), 15–29 (2008)
34. Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J., McKeon, M.: Manyeyes: a site for visualization at internet scale. IEEE Trans. Vis. Comput. Graphics **13**(6), 1121–1128 (2007)
35. Jiang, X., Zhang, J.: A text visualization method for cross-domain research topic mining. J. Vis., 1–16
36. Moloshnikov, I.A., Sboev, A.G., Rybka, R.B., Gydovskikh, D.V.: An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach. Proc. Comput. Sci. **66**, 297–306 (2015)
37. Zhai, X., Li, Z., Gao, K., Huang, Y., Lin, L., Wang, L.: Research status and trend analysis of global biomedical text mining studies in recent 10 years. Scientometrics **105**(1), 509–523 (2015)
38. Chebel, M., Latiri, C., Gaussier, E.: Extraction of interlingual documents clusters based on closed concepts mining. Proc. Comput. Sci. **60**, 537–546 (2015)
39. Santosh, K.C.: g-DICE: graph mining-based document information content exploitation. Int. J. Doc. Anal. Recogn. (IJDAR) **18**(4), 337–355 (2015)
40. Song, M., Kim, S.Y.: Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics **96**(1), 183–201 (2013)
41. Ramakrishnan, C., Patnia, A., Hovy, E., Burns, G.A.: Layout-aware text extraction from full-text PDF of scientific articles. Source Code Biol. Med. **7**(1), 1 (2012)
42. Mooney, R.J., Bunescu, R.: Mining knowledge from text using information extraction. ACM SIGKDD Explor. Newsl. **7**(1), 3–10 (2005)
43. Callan, J., Mitamura, T.: Knowledge-based extraction of named entities. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 532–537. ACM (2002)
44. Al-Emran, M.N.H.: Investigating Students' and Faculty members' Attitudes Towards the Use of Mobile Learning in Higher Educational Environments at the Gulf Region (2014)
45. Al Emran, M., Shaalan, K.: E-podium Technology: A medium of managing Knowledge at Al Buraimi University College via M-learning. In: BCS International IT Conference (2014)
46. Al-Emran, M., Shaalan, K.: Attitudes towards the use of mobile learning: a case study from the gulf region. Int. J. Interact. Mobile Technol. (iJIM) **9**(3), 75–78 (2015)
47. Al-Emran, M., Shaalan, K.: Learners and educators attitudes towards mobile learning in higher education: State of the art. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 907–913. IEEE (2015)
48. Al-Emran, M., Elsherif, H.M., Shaalan, K.: Investigating attitudes towards the use of mobile learning in higher education. Comput. Human Behav. **56**, 93–102 (2016)

49. Al-Emran, M., Malik, S.I.: The Impact of Google Apps at Work: Higher Educational Perspective. Int. J. Interact. Mobile Technologies (iJIM) **10**(4), 85–88 (2016)
50. Al-Emran, M., Shaalan, K.: Academics' awareness towards mobile learning in Oman. Int. J. Com. Dig. Sys. **6**(1) (2017)
51. Zhang, Y., Chen, M., Liu, L.: A review on text mining. In: 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 681–685. IEEE (2015)
52. Verma, T., Renu, R., Gaur, D.: Tokenization and Filtering Process in Rapid Miner. Int. J. Appl. Inf. Syst. **7**(2), 16–18 (2014)
53. Zaza, S., Al-Emran, M.: Mining and exploration of credit cards data in UAE. In: 2015 Fifth International Conference on e-Learning (econf), pp. 275–279. IEEE (2015)