International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Using Text Mining Techniques for Extracting Information

D V S NARAYANA CHOWDARY , V DHANUSH , A VISWA PRAVEEN

Presidency University

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.

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

Introduction:-

Nowadays, almost all of the existing information in different institutions 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. A study by 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 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.

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.

Background on Text Mining and Information Extraction.

1. Text Mining.
2. Information Extraction.
3. Extracting Knowledge from Text.
4. Text Mining Methods and Techniques.

Text Mining Processing Framework.

We have developed our customized framework which is inspired by the designed framework proposed by 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.

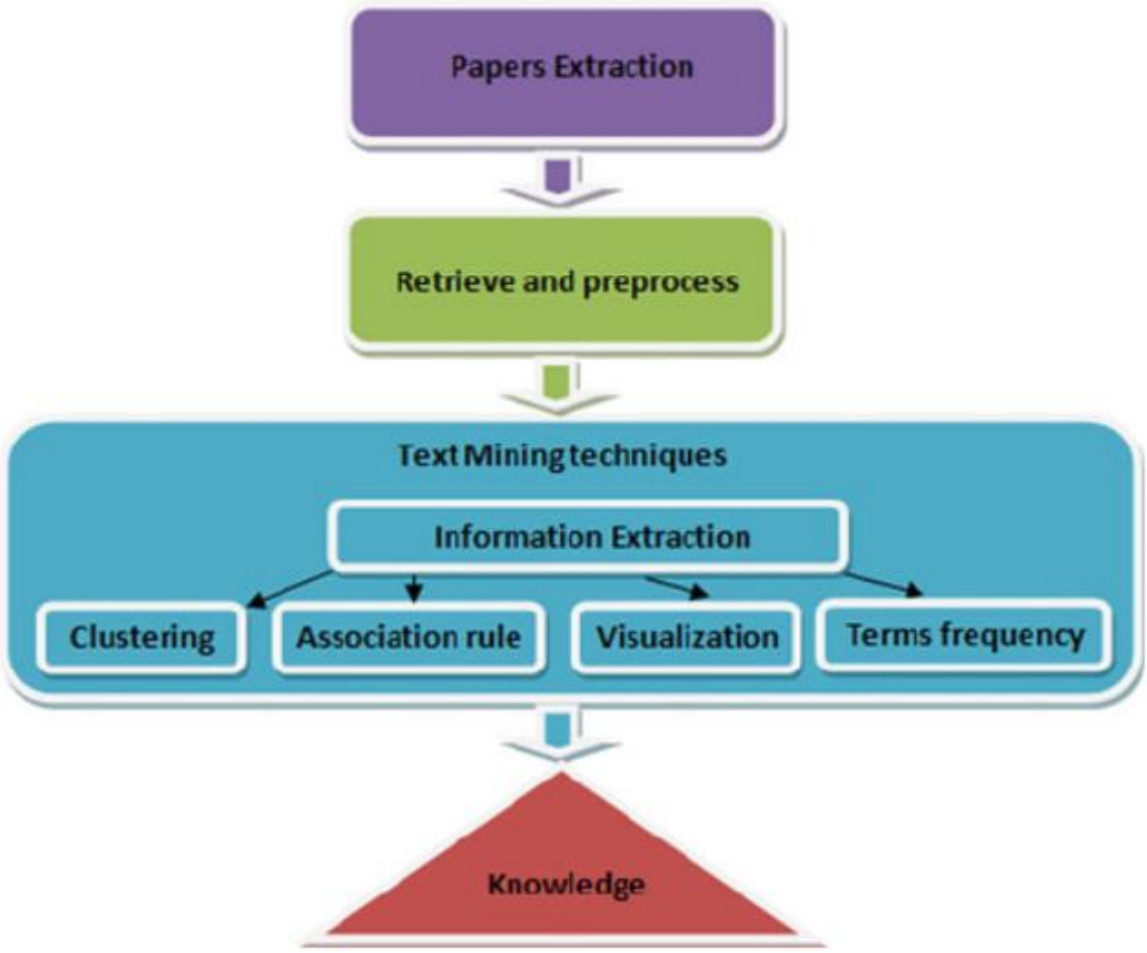


Fig. 1 Text mining processing framework.

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.

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.

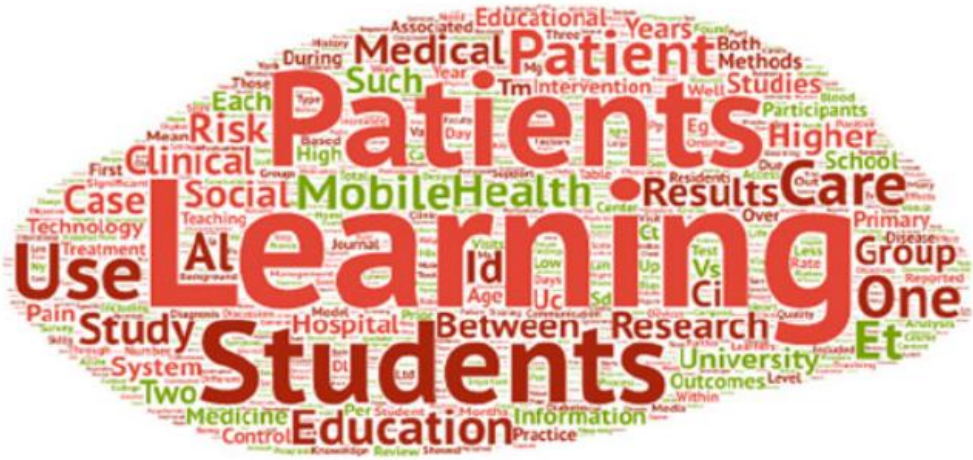
The presence of the linguistic noise is a common problem in the content of the extracted articles and we have dealt with. Then, the cleaned data are uploaded into RapidMiner tool while the misplaced and unnecessary data have been removed from the dataset. In order to improve the performance and data quality, all the irrelevant characteristics are debarred while the data is being uploaded into Fig. 1 Text mining processing framework Using Text Mining Techniques for Extracting Information … 383 RapidMiner tool. The major steps involve the separation of the document into tokens; this task is called Tokenization. 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.

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 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.

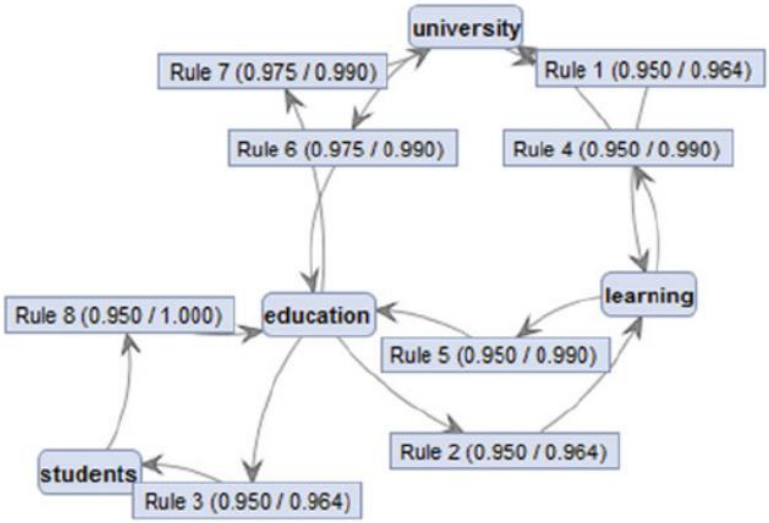


Q2: What are the most frequent terms among the collected articles?

As per the study of 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.

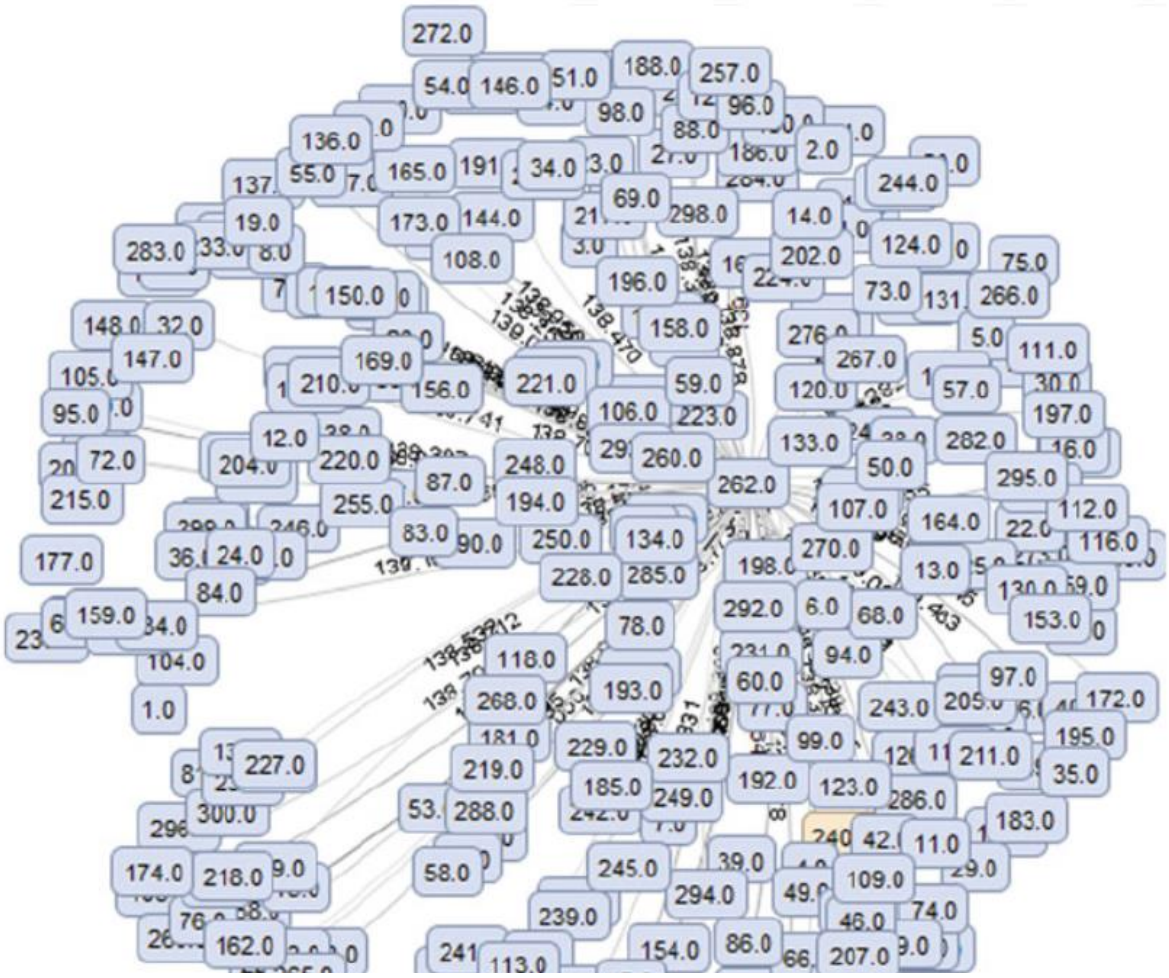
AssociationRules

```
Association Rules
[university] --> [learning] (confidence: 0.964)
[education] --> [learning] (confidence: 0.964)
[education] --> [students] (confidence: 0.964)
[learning] --> [university] (confidence: 0.990)
[learning] --> [education] (confidence: 0.990)
[university] --> [education] (confidence: 0.990)
[education] --> [university] (confidence: 0.990)
[students] --> [education] (confidence: 1.000)
```



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

As per the study of 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 mean



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 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).

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).