Automatic verification of correspondence between teaching resources and executive regulations in the field of design and implementation of concrete buildings: a text mining approach

Fatemeh Hafezi Moghaddas[a], Mojtaba Maghrebi[b*]

[a] Master's student, Faculty of Engineering, Department of Civil Engineering and Construction Management, Ferdowsi University, Mashhad, Iran.

[b] Associate Professor, Faculty of Engineering, Department of Civil Engineering and Construction Management, Ferdowsi University, Mashhad, Iran.

ABSTRACT

One of the challenges of higher education in editing university educational texts is to achieve the maximum compatibility of the content of educational resources with the instructions. therefore, to reach an efficient educational system and in line with industry needs, the appropriateness of the content of the educational resources with the regulations should be evaluated and revised if necessary. The need to review educational resources in the field of engineering and technology is important because these disciplines are needed in the application of industry and services in the country, and in fact, the training of expert and technical forces that can meet the technical needs of the country at different levels is the most important task of curricula in universities. This issue doubles the importance of paying attention to the teaching resources of these disciplines. this research seeks to extract keywords of "Iranian Concrete Regulations" and "reinforced concrete structures" which using three statistical approaches, linguistic knowledge and graph-based approaches proposes a hybrid method. then, the keywords of each document are visualized in a clustered network and analyzed. Comparing the results shows that the contents of these two documents are not completely similar. In fact, it can be said that the Regulations is an instructional document in which all the details have been addressed and all the issues surrounding concrete structures have been discussed. But the book is not satisfied with the design instructions and has examined the concepts related to the design of concrete structures in detail.

KEYWORDS

Text mining, keyword extraction, network analysis, centrality index, Iran's Concrete Regulations (ABA)

[*] mojtabamaghrebi@ferdowsi.um.ac.ir

1. Introduction

Given the rapidly increasing volume of electronic texts, articles, large digital archives, and expanding social networks, the use of data mining techniques is becoming increasingly essential. According to a report by EMC, the volume of information in 2020 exceeded 40 zettabytes (40 * 10^21 bytes) [1], with 80% of this information being textual. Text mining or knowledge discovery from text was first introduced by Feldman and Dagan [2] in 1995 as a subfield of data mining [3].

Since manual processing of a large number of textual documents is time-consuming and error-prone, text mining focuses on automatic extraction of information and identification of valuable hidden patterns from unstructured textual data [4]. Text mining encompasses seven different areas: text classification, text clustering, information retrieval, information extraction, web mining, natural language processing, and concept extraction. Many text mining techniques, including natural language processing, information extraction, and concept extraction, require keywords [5]. Keywords are sets of one or more words that provide a condensed representation of a text's content. In fact, keywords are like the tip of an iceberg, indicating the content of each document [6].

There are two general approaches to finding keywords in a text: Keyword assignment and Keyword extraction [7].

Keywords can be assigned either manually or automatically but the former approach is very time-consuming and expensive. Thus there is a need for automated process that extracts keywords from documents. Keyword extraction approaches include the following [7]:

- Linguistic Knowledge-Based Approaches: These methods are generally based on linguistic knowledge rules and utilize elements such as grammar rules, lexical analysis, discourse analysis, and syntactic analysis.
- Statistical Approaches: These approaches use statistical features of the text corpus. Their advantages include speed, ease of use, and language independence.
- Machine Learning Approaches: These methods are typically supervised learning techniques. Initially, they use labeled datasets to train the model, and then employ the trained model to extract keywords from other documents.
- Network-Based Approaches: These methods are usually unsupervised and use a network constructed from words or phrases in the text. They rank vertices by applying metrics that quantify the network structure [6].

Although keyword extraction from Persian texts has gained attention in recent years, researchers in this field face challenges due to the nature of linguistic and writing elements in Persian. This research aims to extract keywords from long and unstructured Persian texts using three approaches: statistical (TF-IDF component), linguistic knowledge (grammatical role of words), and network construction (closeness centrality measure).

By developing a model based on these three components, the study attempts to examine words from various aspects to achieve the best results. The TF-IDF value only considers the importance of words from a statistical perspective and frequency of occurrence. The closeness centrality measure is calculated based on the co-occurrence network of words and does not take their nature into account. Therefore, this research simultaneously uses the three mentioned components for keyword extraction.

The proposed algorithm is used to examine the correspondence between educational resources and executive regulations. The Iran's Concrete Regulations (ABA) and the book "Reinforced Concrete Structures" (authored by Dr. Mostofinejad) have been selected as examples for examination. In this study, with the help of text mining techniques, important and key words from both documents are extracted, and a technical comparison between the two is presented. The method proposed in this study can be used as a tool to help higher education policymakers in the country review educational resources with the aim of adapting educational content to the needs of the industry.

2. Methodology

this research proposed an algorithm for keywords from Persian text. First, a code is developed in Python and receives the text file of the document under consideration. In the following stage, the input text undergoes preprocessing in 7 steps. Then, the words are scored with three criteria: TF-IDF (statistical), grammatical role of words (linguistic), and closeness centrality (network analysis). The final score is obtained by multiplying these three values, and the keywords are selected. The process of extracting keywords is summarized in Figure 1.

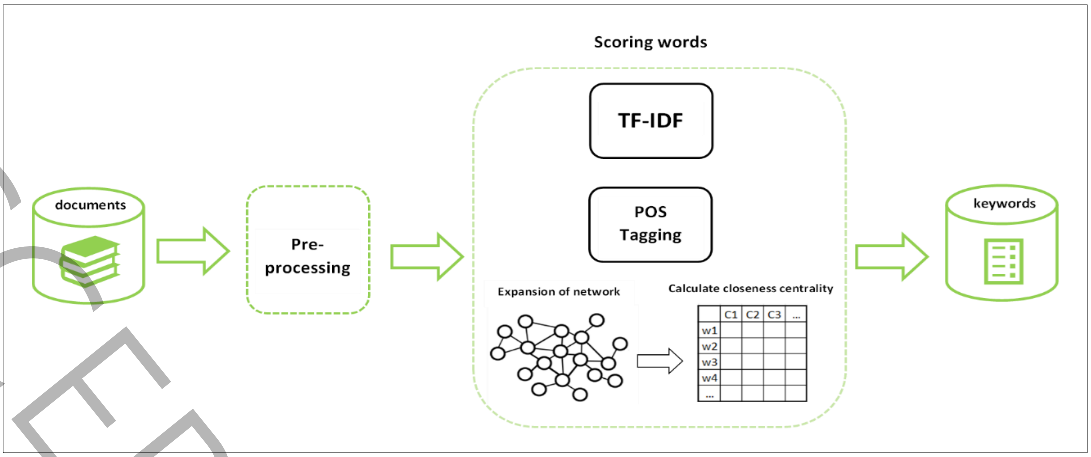

Figure 1: Stages of Keyword Extraction

3. Results and Discussion

Comparing the topics and key terms presented in the two examined sources, it can be concluded that the Iran's Concrete Regulations (ABA) comprehensively covers all aspects and issues related to the design, analysis, implementation, and specifications of materials used in concrete structures. This code serves as an appropriate benchmark for determining right and wrong practices in the field of concrete construction.

ABA ensures that concrete structures meet minimum global standards and possess adequate safety. It also guarantees the stability, strength, and durability of structures under various conditions.

However, a significant portion of the "Reinforced Concrete Structures" book focuses on examining various design methods for concrete members under different load conditions. It presents design-related topics with extensive executive details that are not found in the code. In essence, the code can be described as a comprehensive instructional document that addresses all details and discusses all issues surrounding concrete structures.

4. Conclusion

In this research, an automated model for extracting keywords from long Persian texts has been proposed. The suggested model summarizes the content of a lengthy textual document into a set of words in a short period, which can serve as a suitable tool for examining the content of Persian texts. Additionally, experts and evaluation workgroups can use the proposed method of this research to prepare and draft regulations.

However, due to existing limitations, implementing this process on Persian texts is more challenging compared to English. The proposed model has been implemented on the Iran's Concrete Regulations and the book "Reinforced Concrete Structures" authored by Dr. Mostofinejad (as one of the reference books in universities across the country).

Comparison of the results shows that the content of these two documents is not entirely similar, and the topics emphasized in each differ. The reason for the lack of complete alignment in the content of these two documents can be attributed to their different writing purposes.

5. References

[1] J. Gantz, D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, IDC iView: IDC Analyze the future, 2007(2012) (2012) 1-16.

[2] R. Feldman, I. Dagan, Knowledge Discovery in Textual Databases (KDT), in: KDD, 1995, pp. 112-117.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, K. Kochut, A brief survey of text mining: Classification, clustering and extraction techniques, arXiv preprint arXiv:1707.02919, (2017).

[4] H. Alrasheed, Word synonym relationships for text analysis: A graph-based approach, Plos one, 16(7) (2021) e0255127.

[5] G. Miner, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[6] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, An overview of graph-based keyword extraction methods and approaches, Journal of information and organizational sciences, 39(1) (2015) 1-20.

[7] S. Siddiqi, A. Sharan, Keyword and keyphrase extraction techniques: a literature review, International Journal of Computer Applications, 109(2) (2015).