Survey: Work implemented in Data Mining and Clustering Algorithm

Prof. Dushyant Chawda[1] Prof. Pratik Modi[2] Monil Khamar[3]
[1,2]Assistant Professor [3]Student
[1,2,3]Department of Information & Technology
[1,2,3]LDRP-ITR, KSV, Gandhinagar, India

Abstract— Increasing progress in numerous research fields and information technologies, led to an increase in the publication of research papers. Therefore, researchers take a lot of time to find interesting research papers that are close to their field of specialization. Consequently, in this paper we have proposed a documents classification approach that can cluster the text documents of research papers into the meaningful categories in which contain a similar scientific field. Our presented approach is based on essential focus and scopes of the target categories, where each of these categories includes many topics. Accordingly, we extract word tokens from these topics that relate to a specific category, separately. The frequency of word tokens in documents impacts the weight of the document calculated by using a numerical statistic of term frequency-inverse document frequency (TF-IDF). The proposed approach uses title, abstract, and keywords of the paper, in addition to the categories topics to perform the classification process. Subsequently, documents are classified and clustered into the primary categories based on the highest measure of cosine similarity between category weight and documents weights.

Key words: Cosine Similarity Document Clustering TF-IDF Topics Web Data Mining

I. INTRODUCTION

Within the realm of online information, web document clustering emerges as a potent approach for the identification of documents sharing common content. This method is instrumental in discerning documents with analogous information on the internet [1-3]. Document clustering, by consolidating related papers in a single repository, provides a valuable and efficient approach for the discovery and examination of documents[4]. Web documents can be classified into distinct categories, each defined by a specific set of subjects. These subjects are typically centered around word tokens that play a crucial role in document analysis. Tokens, in the context of text analysis, pertain to frequently recurring terms within textual content. These terms, which are extracted from textual data, play a crucial role in facilitating document classification[5]. Tokens, in the context of text analysis, pertain to frequently recurring terms within textual content. These terms, which are extracted from textual data, play a crucial role in facilitating document classification [6]

II. RESEARCH METHOD

In this research initially, the understanding and analysis of the problem domain in literature is completed, this is later followed by the review and pre-processing of the data, Subsequently, a conceptual framework and model is developed following the literature review and analysis approaches, model testing is performed with supervised and unsupervised versions of machine learning approaches. Finally, the results of the planned analysis are evaluated[46,47,48,49,50,51,61,69]. In this paper a quantitative research paradigm is used. The paradigms of quantitative research, aim to explore the phenomena in the application of deductive techniques. They have a positivist and objectivist orientation of epistemology and ontology. Quantitative data is a type of structured knowledge that can be collected using a variety of approaches in the form of primary data sources and secondary data sources. Conventional and unconventional forms of data collection techniques can be applied via sensor measurements as input, as end user data entry as input, via web services, internet of Things, paper-administered surveys, online surveys, etc. [51,52,53,54,55,56].

In the data mining process Athena data mining model of Özerk has been employed as seen in Figure 1. In the analysis conducted Multilayer Perceptron, Bayesian Networks, Hoeffding Tree, Random Tree, Kmeans, Make density based clustering, Hierarchical Clusterer, Filtered Clusterer, Farthest First, Expectation Maximization, Cobweb, Canopy, J48, JRip, Part, OneR, ZeroR, M5Rules, Decision Table, Decision Stump, Random Forest, Random Tree Methods have been applied for the machine learning. Among these algorithms unsupervised machine learning algorithms here assesses the instance values and assigns these independent values to the respective segment clusters whereas supervised machine learning algorithms mainly focuses on mapping the multivariate variables in input layers to class labels in output layers with transformation and mapping functions. Additionally class based metrics are evaluated and associated rules are generated in an reinforced fashion some applying forward feeding and backpropagation approaches based on the algorithmic designs and architectures [51,52,53,54,55,56].

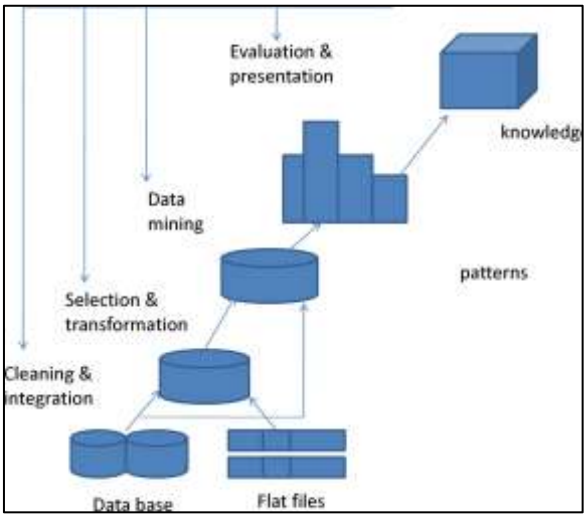

Fig. 1: Knowledge Discovery in Database (KDD)

III. Related work

In this research, a multitude of scholars emphasized the utilization of data mining technology by educational experts for the purpose of guiding and supervising technical integration initiatives within school settings. The significance of constructing KDD data mining technology using educational data is emphasized in the study, which also highlights prevalent data mining techniques including factor analysis, regression, and correlation mining. In our earlier literature reviews within the initial section, it was noted that research papers are commonly classified and retrieved using various approaches, such as user queries and semantic representation, among others. Nevertheless, when it comes to the effectiveness of clustering algorithms, it becomes imperative to define the specific attributes (e.g., words, terms, or phrases) that need to be extracted from documents in order to support the clustering algorithms performance.

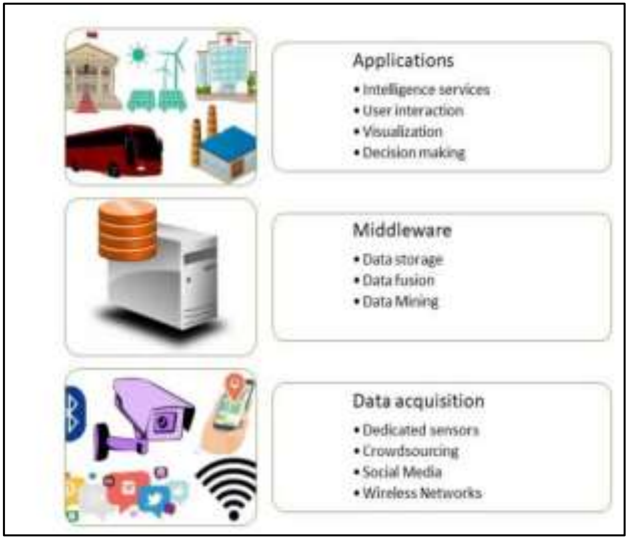

Fig. 2: The three layers of architecture.

IV. Conclusion

In this survey, Work implemented in Data Mining, Clustering Algorithm and classification approach for clustering automate the process using web data mining techniques. Some knowledge is obtained in the following interpretations of the first cluster, this cluster is typically high, the second cluster, this cluster is typically low and the third cluster, this cluster is a medium type. A healthy life, right nutrition with a lifestyle balanced with exercise constitutes and important value. Many hormones and enzymes have influence in the digestion of carbohydrates, proteins and fats in the metabolis. The selected articles were divided into four categories based on their focus: architecture/ platform, framework, applications, and security. Most of the reviewed articles were focused on architecture/platform, and the Kyungpook National University of South Korea was found to have the greatest contribution to the examined literature. The countries with the highest number of articles in this field were found to be South Korea, China, India, and the United States in that order.

REFERENCES

[1] Text documents clustering using data mining techniques Ahmed Adeeb Jalal, Basheer Husham Ali Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq

[2] Analysis Of Drug Data Mining With Clustering Technique Using K-Means Algorithm To cite this article: April Lia Hananto et al 2021 J. Phys.: Conf. Ser. 1908 012024

[3] A Classification and Clustering Approach Using Data Mining Techniques in Analyzing Gastrointestinal Tract Özerk Yavuza,1 a,1Haliç Üniversity, İstanbul 34445, Turkey ORCID ID: 0000-0002-1371-688X

[4] Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining GUIYUN FENG , MUWEI FAN , AND YU CHEN School of Management, Guizhou University, Guiyang 550025, China

[5] Review Data Mining Algorithms for Smart Cities: A Bibliometric Analysis Anestis Kousis and Christos Tjortjis

[6] A systematic survey of data mining and big data analysis in internet of things Yong Zhong1 · Liang Chen2 · Changlin Dan1 · Amin Rezaeipanah3 Accepted: 8 May 2022 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022