j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

MRH: A Large-Scale Text Dataset for Web Content Mining

Mohammed Ali Mohammed [*1], Hasan Aqeel Abbood [2], Raad Mahmood Mohammed[3]

[1&2&3] *College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq.*

mohammed.ali@uoitc.edu.iq

Abstract The amount of information, specifically the information related to website environment, is increasing and becoming larger and larger day by day, thus, playing an important role in the discovery of diverse knowledge on the web. In this paper, our goal is the creation of a new dataset for web content mining that is used in the testing and evaluation of any new system before the production phase. Key characteristics for our dataset are semi-structured, size=4.05 MB, type: text, No. of rows: 298, No. of columns: 6, file type: .csv (comma separated value), domains: Computer, Mathematical, Physics, Chemistry Sciences. Python code will be used to read set of links from set of websites, then read and save the web page content as text of these links. Our dataset discussed based on (Dataset Overview and Scope, Data Quality and Robustness, Utility and Applications), and evaluated and showed the with its robust structure—comprising domain, website, and webpage data—it supports a variety of web content mining applications.

10.36371/port.2025.4.2

Keywords: *web content mining, MRH Dataset, Text Mining Dataset, Data Collection.*

1. INTRODUCTION

Web Mining is the practice for sifting through the expansive amount of data in the system that is available on the Web, including web documents, interconnections between documents, and website usage patterns, to find and extract valuable and pertinent information. The primary objective is to reveal patterns within the vast expanse of the web [1][2].

Web content mining is one of the three different types of techniques in web mining that involves the extraction of pertinent information from multiple and diverse sources throughout the web. This process aims at the discovery and analysis of data that is contained within web pages and other web-based resources [3][4].

Database techniques used in web services are focused on the problems of managing and querying the data. There are three categories of tasks related to handling those problems: modeling and querying the web, information extraction and integration, and website construction and restructuring [5].

Web content mining uses two types of approaches, the agent-based approach and the database approach. For the latter, it requires a new database in terms of data and content, specializes in text-type data and in various differences, practical specializations. The need for this data can be summarized into: Classifying web documents into categories, identifying topics of those documents, and finding similar web pages across different web servers and applications related to the relevance of this data. Our proposal is to try and view the dataset in order to infer the structure of the website or to transform said website to a database where better information management and querying becomes possible.

The remaining sections of the paper are structured as follows: Section 2 presents a how to create our dataset. Section 3 describe out dataset from many parameters. Section 4 showed the analysis of our dataset. Section 5 discussion our dataset. Finally, Section 6 concludes the paper.

2. Dataset Creation Methodology

In this section, we're going to describe how to create a new dataset, what is the source of this data, how this data is going to be collected, processing and storage, and what is the structure of this dataset.

Data Sources: we start here by selecting the websites for web content mining instead of going through the whole internet for specific resources. These websites are selected from various science domains (Computer Science, Physics, Chemistry and Math).

Data Collection: here, we're going to utilize the use of Python programming language. Data collection will be completed in two phases. First phase: the code automatically read the links from the main website table and assigns the URL to "DURL" variable. Second phase: Python code uses each link from "DURL" variable, opens it, and reads all the links into "ALL_URLS" variable. Finally, for each link in "All_URLS" variable, the code will open the link, read the content as text, then save all the information in our dataset.

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

3. Dataset Description

This section describes the dataset details after it was created in the previous section. The dataset is semi-structured type with size 4.05 MB. Dataset will be stored in CSV format (.csv) which can be opened in MS Excel. Dataset structure (shown in figure 1) contains many sheets (as a table in a database). The structure is designed as a relationship between the sheets in order to apply normalization, reduce the size, thus, faster processing

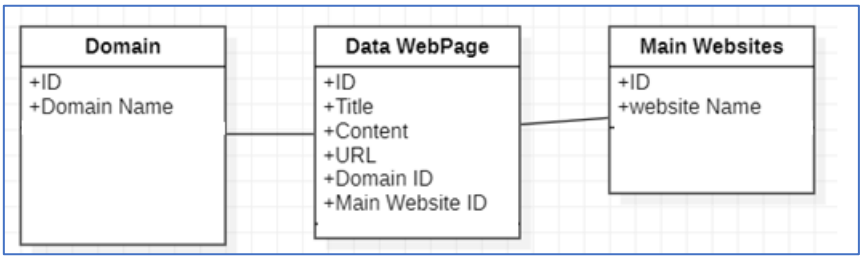

Figure 1: *Structure of Our Dataset*

The total number of webpages (no. of rows) in the dataset is 298 broken into 144 in Computer Science domain, 30 in Chemistry Science domain, 38 in Physics Science domain, and 86 in Mathematics Science domain.

The author's own dataset will be created for the purpose of web content mining application, which should contain a set of webpage contents with titles. We are going to use the following URLs: "geeksforgeeks.org" [6], "mathworld.wolfram.com" [7], "www.chemguide.co.uk" [8], "www.physicsclassroom.com" [9], and "socratic.org" [10] to create this dataset and save it as a CSV file. Table 1, Table 2, and Table 3 are samples of the content of these datasets. The new dataset is created with the following headers:

- Id: unique number
- Title: Title of web page
- Content: Content of web page
- URL: URL of the web page.
- Domain ID: contain domain id from domain sheet.
- Main Website ID: contain main website id from main website sheet.

Table 1: *Sample of data webpage sheet*

Id	Title	Content	URL	Main Website ID	Domain ID
1	Acids, bases and salts	All text in page	https://www.chemguide.co.uk/14to16/acid.html	5	2
2	C Programming Language Tutorial	All text in page	https://www.geeksforgeeks.org/c-programming-language/	4	1
3	Hardy's Rule	All text in page	https://mathworld.wolfram.com/HardysRule.html	3	4
4	Light Waves and Color	All text in page	https//www.physicsclassroom.com/Physics-Interactives/Light-and-Color	7	3

Table 2*: Sample of data domain sheet*

Id	Domain Name
1	Computer Science
2	Chemistry Science
3	Physics Science
4	Mathematics Science

M. A. Mohammed, et al. 2025, A Comprehensive Data Enhancement Method for the Pima Dataset to Improve Diabetes Prediction Performance. *Journal port Science Research*, 8(4), pp.321-326. https://doi.org/10.36371/port.2025.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

Table 3: *Sample of main website sheet*

Id	Website name
1	https://mathworld.wolfram.com/topics/NumericalIntegration.html
2	https://www.geeksforgeeks.org/
3	https://www.chemguide.co.uk/14to16menu.html
4	https://www.physicsclassroom.com/Physics-Interactives/

4. Dataset Analysis

In this section, the dataset will undergo a comprehensive analysis and evaluation based on a variety of parameters. The results will be presented in the form of a figure to provide a clear and simplified view, given the large size of the dataset and the need for easy interpretation.

Figure 2 shows the most frequent words with all titles and contents after applying the pre-processing step. Words such as 'tutorial,' 'data,' 'mission,' 'python,' and others appear most frequently in the dataset.

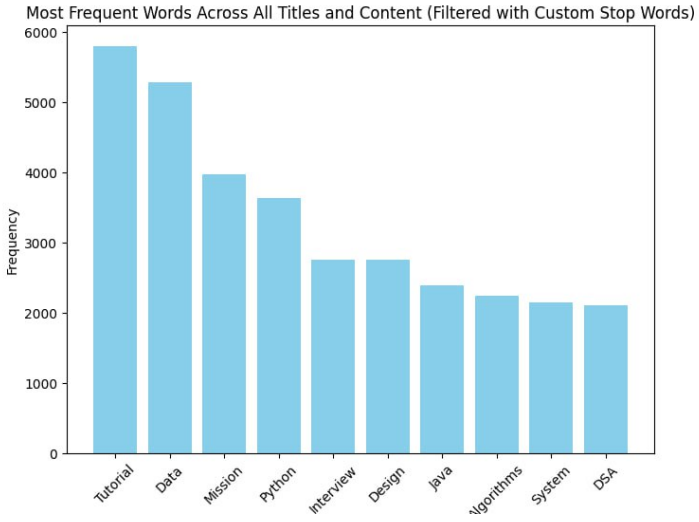

Figure 2: *most frequency word in title and content*

Figure 3 shows the distribution of IDs for the main website table. The chart reveals the frequency of occurrences for each ID, this figure helping identify potential patterns or imbalances in the dataset.

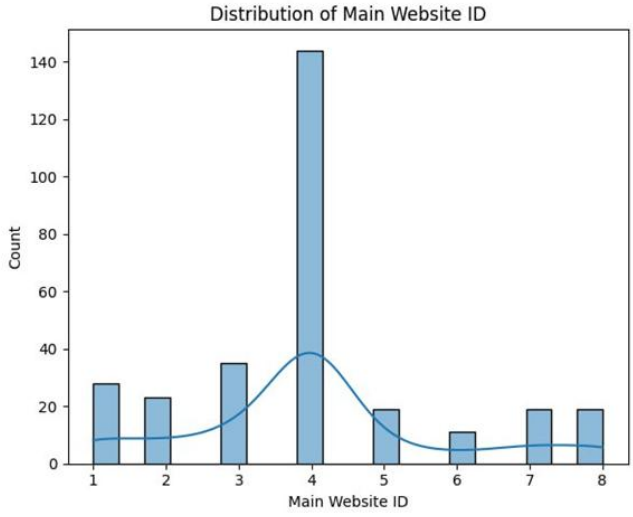

Figure 3*: Distribution of main website ID*

Figure 4 shows the distribution of IDs for the domain table. The chart reveals the frequency of occurrences for each ID to show the areas of higher activity of the website.

M. A. Mohammed, et al. 2025, A Comprehensive Data Enhancement Method for the Pima Dataset to Improve Diabetes Prediction Performance. *Journal port Science Research*, 8(4), pp.321-326. https://doi.org/10.36371/port.2025.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

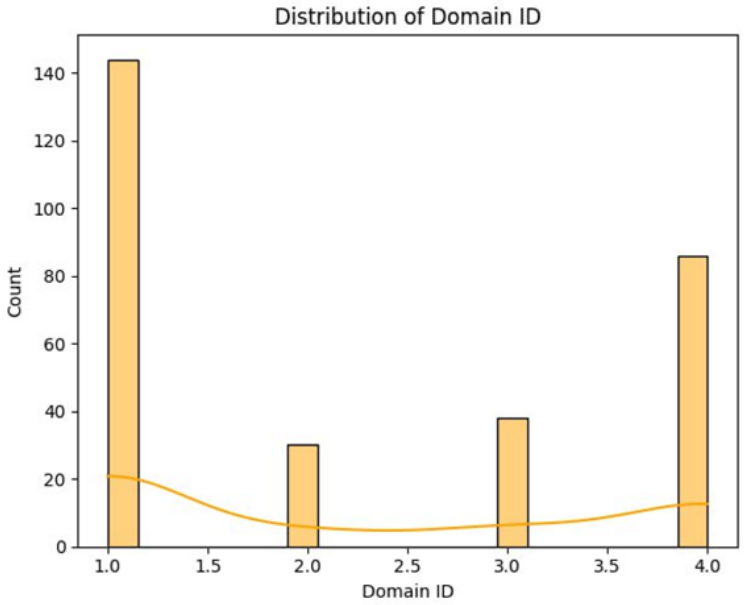

Figure 4: *Distribution of Domain ID*

Figure 5 shows the evolution of word count in the title and content columns of the webpage table. Y-axis represents the word count; the X-axis shows the index of rows in the table. This figure shows highlights the variations in word count.

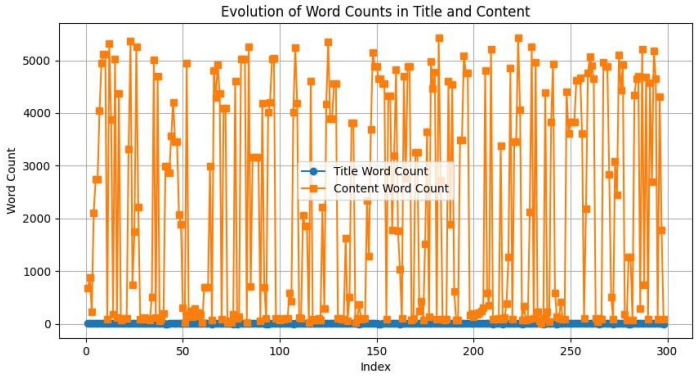

Figure 5: *Evolution of word count in title and content*

Figure 6 shows the word cloud, the word cloud defined as a useful tool for identifying dominant themes in the database, allowing for quick insights into the primary areas of focus. The figure shows the balance of frequent topics. Figure 6 represent the frequency of words found within the "Title" and "Content" columns of the database. Each words in larger font sizes indicate that this word is appearing more frequently from other words, and verse vice.

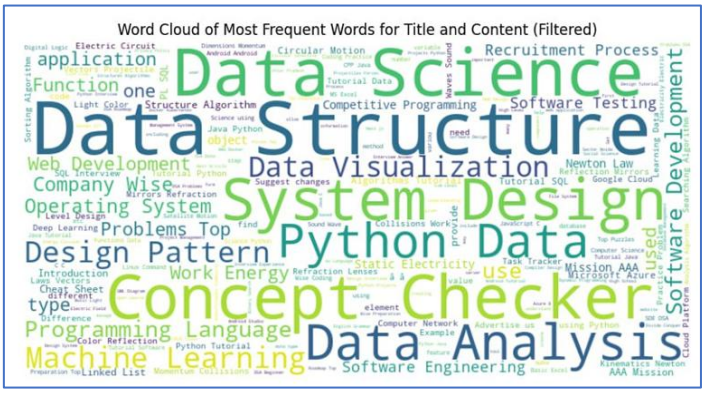

Figure 6: *word cloud of Title and Content of dataset*

M. A. Mohammed, et al. 2025, A Comprehensive Data Enhancement Method for the Pima Dataset to Improve Diabetes Prediction Performance. *Journal port Science Research*, 8(4), pp.321-326. https://doi.org/10.36371/port.2025.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

Figure 7 shows the radar chart visualizes three variables: title length, domain ID, and main website ID. This figure shows the view of how each variable rank relative to the others variables. The top value is 6 in the "title length" which means this variable has the greatest significance of the dataset and so on for the rest of the values. Radar charts useful tool for identifying areas with strong variation.

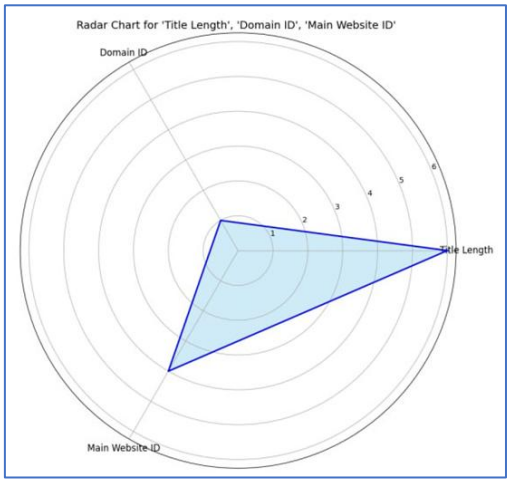

Figure 7: *Radar chart for title length, domain id, main website id*

5. Dataset Discussion

5.1. Dataset Overview and Scoe:

- dataset created includes text data from variable scientific webpages for domains: computer science, mathematics, physics, and chemistry.
- effectively organize of structure dataset: Domain, Main Website, and Data Web Page. These tables with relationship through domain IDs and website IDs.
- dataset collected from a wide of topics, making it valuable for tasks such as natural language processing (NLP), topic modeling, and etc.

5.2. Data Quality and Robustness:

- The content from each webpage has diverse and efficient which ensuring the dataset is robust in diversity and richness of text.
- Dataset contain a variety of domains which will be essential for developing more sophisticated content mining algorithms.
- Dataset inclusion of content such as titles and URLs provides valuable metadata to enhancing the usefulness of the dataset for web content mining tasks.

5.3. Utility and Applications:

The dataset has significant potential in web content mining, with a wide range of possible applications:

- Topic Modeling.
- Content Recommendation Systems.
- Text Classification and Clustering.
- Trend Analysis.

6. Conclusion

This large-scale text dataset is a valuable tool for web content mining in scientific fields like computer science, math, physics, and chemistry. Its structured format supports applications such as topic modeling, classification, and trend analysis, making it ideal for machine learning and NLP tasks. It also enables advanced methods to analyze and interpret scientific web content, aiding in the study of how scientific knowledge is shared online.

REFERENCES

[1] P. Sukumar, L. Robert, and S. Yuvaraj, “Review on modern Data Preprocessing techniques in Web usage mining (WUM),” in *2016 international conference on computation system and information technology for sustainable solutions (csitss)*, 2016, pp. 64–69.

[2] P. Ristoski and H. Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey,” *J. Web Semant.*, vol. 36, pp. 1–22, 2016.

[3] S. Yadao and A. Vinaya Babu, “Usage of Web Mining for Sales and Corporate Marketing,” in *Communication Software and Networks: Proceedings of INDIA 2019*, 2021, pp. 55–60.

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

[4] Mohammed, M. A., Hamid, R. A., & AbdulHussein, R. R. (2024). Data Collection and Preprocessing in Web Usage Mining: Implementation and Analysis. *Iraqi Journal for Computers and Informatics*, *50*(2), 54-74.

[5] Florescu, D., Levy, A., & Mendelzon, A. (1998). Database techniques for the World-Wide Web: A survey. ACM Sigmod Record, 27(3), 59-74.

[6] [geeksforgeeks], https://www.geeksforgeeks.org/ , Access on 3 March 2025.

[7] [Wolfram MathWorld], https://mathworld.wolfram.com , Access on 4 March 2025.

[8] [chemguide], https://www.chemguide.co.uk , Access on 6 March 2025.

[9] [The Physics Classroom], https://www.physicsclassroom.com , Access on 7 March 2025.

[10] [Socratic] , https://socratic.org , Access on 7 March 2025.