International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Leading the Way in Efficient Web Content Mining through Advanced Classification and Clustering Techniques

[1]Yogesha T, [2]Dr. Thimmaraju S N

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: The clustering techniques in online content mining for knowledge discovery is the main topic of the abstract for the article "Clustering Techniques in Knowledge Discovery for Web Content Mining". The application of association rule mining, sequential pattern discovery, and clustering as data mining techniques for knowledge extraction is mentioned.

When the data comes from the online, web mining—the process of obtaining information from web data—is referred to as a subset of knowledge discovery from databases (KDD). A particular kind of web mining called web use mining (WUM) seeks to identify, assess, and make use of hidden knowledge from online data sources. Data from user registration forms, server access logs, user profiles, and transactions are used in web use mining.

It is mentioned that one technique utilized in online content mining for knowledge discovery is clustering algorithms. In the context of online content mining, clustering is the process of assembling comparable data points into groups according to their shared traits or patterns. Clustering may be used to find page sets, page sequences, and page graphs.

The use of text analysis methods for knowledge discovery from unstructured materials, including feature extraction, theme indexing, clustering, and summarization, is also mentioned in the abstract. Press releases, emails, notes, contracts, government reports, and news feeds are just a few of the documents from which valuable information may be extracted thanks to these strategies.

An overview of the use of clustering algorithms in knowledge discovery for online content mining is given in the abstract overall. It highlights the use of text analysis tools to extract knowledge from unstructured documents and the clustering approach in online use mining.

Key Words: *Data Mining, Heterogeneous networks, Knowledge Discovery, Text mining, Web mining*

1. Introduction

Web content mining is the use of data mining techniques [7]to extract knowledge from web data, such as web documents, document hyperlinks, and website usage records. With the enormous quantity of material available online, the World Wide Web (WWW) has become a popular and dynamic platform for conveying information.

In online content mining, effective methods for categorization and clustering are essential for knowledge discovery.[3] By classifying and structuring online data, these strategies facilitate the discovery of important trends and insights. While clustering groups related online content together based on their intrinsic similarities, classification entails applying preset categories or labels to web content based on its features 1.

The significance of effective classification and clustering methods for knowledge discovery [4]in online content mining in this answer.

Web Mining as a Process for Knowledge Discovery

Pre-processing, analysis, generalization, and resource discovery are the four stages of web mining.

1. **Resource identification**: This stage involves determining the resources required to retrieve information.

2. **Pre-processing**: From the resources that have been discovered, pertinent data is chosen. Techniques for extracting information are intimately tied to this stage.

3. **Generalization**: Several web publications are subjected to automatic pattern recognition. This stage makes use of classification trees, clustering, and data [7] mining techniques.

4. **Analysis**: In this stage, the pattern finding is verified and examined.

During the generalization stage, effective classification and clustering algorithms are used to extract patterns and insights from online pages.

[1]*Assistant Professor, Department of CS&E, VTU-RRC, PG Centre Mysuru, Karnataka, India*
yogesh@vtu.ac.in
[2] *Professor. Department of CS&E, VTU-RRC, PG Centre Mysuru, Karnataka, India*
Thimmaraju_sn@vtu.ac.in

2. Review of Literature:

Prem Sagar Sharma, Divakar Yadav, and R. N. Thakur's Author's work on search engines return relevant web pages to users based on their queries. However, the most relevant web pages may not always be the most important for user queries. Therefore, new techniques are required that consider user queries as an additional parameter to find the relevant web pages. And increasing size of the web, search engines often delay returning a list of web pages as output to users. This delay is referred to as perceived latency. To reduce this latency, a pre-fetching mechanism needs to be developed.

This observation suggests that personalized content can improve the search results for user queries. These gaps define the challenges and new research paths for researchers in the field of web page ranking and web [13] mining techniques.

The literature review in the research paper "Knowledge Discovery [4] using Text Mining: A Programmable Implementation on Information Extraction and Categorization" discusses the various aspects of text mining, including its purpose, functions, and implementation. The authors have used a modified version of Porter's Algorithm for inflectional stemming and a domain dictionary for the 'Computer Science' field to implement Information Extraction and Categorization. The paper also discusses the limitations of the current implementation, such as its accuracy only for documents related to the Computer Science field and the limited words in the domain dictionary. The authors suggest future work, including introducing a 'Self Learning' functionality, using active learning methods, incorporating clustering, improving question answering, and giving the text mining tool a web interface.

A Web Mining Process for Knowledge Discovery[4] of Web Usage Patterns" discusses the vast digital universe and the application of web mining techniques [13][7] to extract knowledge from it. The authors categorize web mining into three distinct categories: Web content mining, Web Structure Mining, and Web Usage Mining. The primary objective of a Web Mining process is to discover interesting patterns and rules from data collected within the Web space. Web usage mining operates on the data from server access logs, information from users' registration application forms, users' profiles, and transactions. The authors also discuss the application areas of Web Usage Mining, which include personalization, system improvements, modification of web site based on discovered web user navigation patterns, business intelligence, and characterization of use. The paper also discusses the use of pattern discovery techniques in the mining process and the application of web usage mining in an E-Learning scenario. The authors conclude that as the web and its usage continue to grow, the opportunity to analyze web data and extract all manner of useful knowledge from it also grows.

Xiaoling Shu in [18] discusses the application of data mining and machine learning [1]in the field of social science research. It cites several studies that have used these techniques to analyze social science data, such as the use of a K-means clustering algorithm to identify distinct Mexican-to-United States migrant clusters and latent class analysis to examine gender attitudes. The paper also discusses the role of big data in knowledge discovery, with examples of studies that used text mining and deep learning to extract knowledge from large datasets. Various machine learning [5] techniques used in social science research are overviewed, including supervised and unsupervised learning, decision trees, random forests, artificial neural networks, and deep learning. The paper also discusses the challenges and limitations of using data mining and machine learning [1][5]in social science research, such as the complexity of models, the difficulty of interpretation, the risk of overfitting, and the use of convenience samples.

3. Methodology:

Efficient classification and clustering techniques are fundamental for knowledge discovery [4] in web content mining. Here's a methodology that outlines the steps to achieve this:

1. Problem Definition and Data Collection:

- Clearly identify the project objectives for online content mining.
- Identify data collection sources (e.g., websites, social media, forums).
- Collect a representative dataset for training and testing.

2. Data preprocessing:

- Remove noise and extraneous information from the dataset, such as HTML elements and special characters.
- Normalize text data by converting to lowercase, deleting stop words, and stemming/lemmatizing words.
- Extract textual features for machine learning[1] methods like TF-IDF and word embeddings.

3.Classification Techniques:

- Choose suitable classification techniques for the task (e.g., Naive Bayes, Support Vector Machines, Decision Trees, Neural Networks).

- Divide dataset into training and testing sets for model assessment.
- Use cross-validation techniques to train and optimize hyperparameters in classification models.
- Evaluate model performance using measures including accuracy, precision, recall, and F1-score.
- Select the best performing model for deployment.

4. Clustering Techniques:

- Choose suitable clustering algorithms depending on the characteristics of the data (e.g., K-Means, Hierarchical Clustering, DBSCAN).
- Apply dimensionality reduction techniques like PCA or t-SNE if dealing with high-dimensional data.
- Cluster the data points into groups based on similarity.
- Evaluate the quality of clusters using metrics like silhouette score or Davies–Bouldin index.
- Analyze the clusters to extract meaningful insights and patterns.

5.Integration and Interpretation:

- Integrate the classification and clustering results to gain a comprehensive understanding of the web content.
- Interpret the findings to extract actionable knowledge and insights.
- Visualize the results using plots, charts, or dashboards to communicate findings effectively.

6.Iterative Refinement:

- Fine-tune the models and algorithms based on feedback and insights gained from the initial results.
- Continuously monitor the performance of the models and update them as new data becomes available.

7.Documentation and Reporting:

- Document the entire methodology, including data preprocessing steps, model selection criteria, and evaluation metrics.
- Prepare a comprehensive report summarizing the findings, insights, and recommendations derived from the web content mining process.

Novel contribution:

To create a novel contribution methodology for efficient classification and clustering techniques in web content mining need to integrate cutting-edge approaches with traditional methods. Here's a proposed methodology:

1. **Hybrid Feature Representation**:

- Develop a hybrid feature representation technique that combines traditional bag-of-words models with advanced deep learning-based embeddings.
- Utilize pre-trained language models like BERT or GPT to extract rich contextual representations of web content.
- Incorporate domain-specific knowledge into the feature representation process to enhance classification and clustering accuracy.

2. **Self-Supervised Learning for Pretraining**:

- Apply self-supervised learning techniques to pretrain models on unlabelled web data.
- Leverage methods such as contrastive learning or masked language modelling to learn meaningful representations from raw web content.
- Fine-tune the pretrained models on labelled data for specific classification and clustering tasks, leading to improved generalization and efficiency.

3. **Attention Mechanisms for Contextual Understanding**:

- Integrate attention mechanisms into classification and clustering models to capture the contextual dependencies within web content.
- Design attention mechanisms that adaptively focus on relevant parts of the text, considering the hierarchical structure and semantic relationships present in web documents.

4. **Ensemble Learning for Robustness**:

- Employ ensemble learning techniques to combine multiple classification and clustering models.
- Aggregate predictions from diverse models trained with different algorithms or feature representations to improve robustness and generalization performance.
- Implement techniques like stacking or boosting to further enhance the predictive power of the ensemble.

5. **Active Learning for Data Efficiency**:

- Implement active learning strategies to intelligently select informative instances for annotation.
- Develop algorithms that actively query the most uncertain or informative samples from the

unlabelled data pool, reducing the annotation effort while maintaining classification and clustering performance.

6. **Interpretability and Explainability**:

- Emphasize interpretability and explainability in classification and clustering models to facilitate understanding and trustworthiness.
- Integrate methods for generating human-interpretable explanations of model predictions, enabling stakeholders to comprehend and validate the discovered knowledge.

7. **Continuous Model Adaptation**:

- Establish a framework for continuous model adaptation that dynamically adjusts classification and clustering models to evolving web content.
- Implement mechanisms for incremental learning and online updating of models to accommodate changes in the characteristics and distribution of web data over time.

8. **Evaluation Metrics Beyond Accuracy**:

- Extend the evaluation criteria beyond traditional accuracy metrics to capture the quality and relevance of discovered knowledge.
- Introduce novel evaluation metrics that consider factors such as novelty, diversity, and actionable insights extracted from the classified and clustered web content.

Result and Discussion:

Table 1. Comparison of existing methods for data extraction from heterogeneous networks

Method	Description	Pros	Cons
Node Embeddings	Representation learning techniques aiming to encode nodes as dense, low-dimensional vectors while preserving network structure and properties.	Captures both structural and semantic information Scalable to large networks- Embeddings can be used as features	Computationally intensive May does not capture long-range dependencies well
Graph Neural Networks	Deep learning models are designed specifically to operate on graph-structured data. They can perform both node-level and graph-level predictions.	Can learn hierarchical representations Handle different data types Powerful for various tasks	- Limited interpretability Sensitive to graph structure and parameters
Meta-path-based Methods	Utilizes sequences of node and edge types, called meta-paths, to define semantic relationships between nodes. Features are extracted based on these paths.	Explicitly captures semantic relationships Interpretable Can handle heterogeneous data types	Requires predefined meta- path May not capture complex relationships
Heterogeneous Information Networks (HIN) Mining	Focuses on extracting patterns and knowledge from heterogeneous networks. Techniques include network motif mining, heterogeneous graph mining, and subgraph mining.	Can discover interesting patterns and relationships Useful for knowledge discovery Exploratory analysis	Computationally expensive May suffer from scalability issues
Feature Engineering	Traditional approach involving extracting handcrafted features from nodes, edges, or subgraphs based on domain knowledge or heuristics.	Simple and interpretable Domain-specific features capture important characteristics Computationally efficient	Limited by domain knowledge May overlook complex patterns

Most methods in Table 1 for their computation and testing purposes use data from Kaggle at https://www.kaggle.com/code/antoniohidalgo/keras-neuralnet-in-heterogeneous-field-data-set for their computation. Upon closer inspection it can be observed that clustering, even though is best suited for unsupervised pattern extraction system, may not be the best fit for the given problem. Hence it is better to look for other methods such as ANN, CNN, or GAN for the same.

4. Conclusion:

To conclude, this paper has presented a comparison of the most followed methods for homogeneous data extraction in heterogeneous networks including IoT, CCN, GSM and others. The methods are iterated and compared. After thorough comparison it can be seen that there is still a large possibility for research in this domain due to the increasing penetration of networks in our lives. This paper proposes Generative adversarial networks (GAN) for the purpose of experimentation.

References

[1] Shu, Xiaoling & Ye, Yiwan. (2022). Knowledge Discovery: Methods from data mining and machine learning. Social Science Research. 110. 102817. 10.1016/j.ssresearch.2022.102817

[2] Allahyari, Mehdi & Pouriyeh, Seyedamin & Assefi, Mehdi & Safaei, Saied & Trippe, Elizabeth & Gutiérrez, Juan & Kochut, Krys. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.

[3] Dash, Yajnaseni. (2013). A Review of Clustering and Classification Techniques in Data Mining.

[4] P. Madhura, M. Padmavathamma, 2015, A Web Mining Process for Knowledge Discovery of Web usage Patterns, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCACI – 2015

[5] Xiaoling Shu, Yiwan Ye,Knowledge Discovery: Methods from data mining and machine learning, Social Science Research, Volume 110, 2023,102817,ISSN 0049-089X,

[6] Antonia Kyriakopoulou,"Text Classification Aided by Clustering: a Literature Review" in "Tools in Artificial Intelligence" doi: 10.5772/6083

[7] Ngai, Eric & Xiu, Li & Chau, Dorothy. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Appl. 36. 2592-2602. 10.1016/j.eswa.2008.02.021.

[8] Lorena Siguenza-Guzman, Victor Saquicela, Elina Avila-Ordóñez, Joos Vandewalle, Dirk Cattrysse, Literature Review of Data Mining Applications in Academic Libraries, The Journal of Academic Librarianship, Volume 41, Issue 4, 2015, Pages 499-510, ISSN 0099-1333,

[9] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, Saeed Ur Rehman, Research on particle swarm optimization based clustering: A systematic review of literature and techniques, Swarm and Evolutionary Computation, Volume 17, 2014, Pages 1-13, ISSN 2210-6502,https://doi.org/10.1016/j.swevo.2014.02.001.

[10] International Journal of Scientific Research in Computer Science, Engineering and Information Technology A Survey on Text Mining - Techniques, Application 2023

[11] K. Mohan, "A survey on web structure mining," *International Journal of Advanced Computer Research*, vol. 1, no. 1, pp. 715–720, 2017.

[12] S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie revenue prediction based on purchase intention mining using YouTube trailer reviews," *Information Processing & Management*, vol. 57, no. 5, Article ID 102278, 2020.

[13] Prem Sagar Sharma, Divakar Yadav, R. N. Thakur, "Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey", Mobile Information Systems, vol. 2022, Article ID 7519573, 19 pages, 2022. https://doi.org/10.1155/2022/7519573

[14] http://paginas.fe.up.pt/~ec/files_0506/slides/06_WebMining.pdf [Accessed on Feb. 6,2013]

[15] Chintandeep Kaur, Rinkle Rani Aggarwal," Web Mining Tasks & Types: A Survey", International Journal of Research in IT & Management, Volume 2, Issue 2 (ISSN 2231-4334), February 2012, Pp-547-559.

[16] K. Wang and H. Liu. Discovering T ypical Structures of Documents: A Road Map Approach. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 146–154, 1998.

[17] Bassiou, N. and Kotropoulos, C. 2006. Color Histogram Equalization using Probability Smoothening. Proceedings of XIV European Signal Processing Conference

[18] Shu, Xiaoling & Ye, Yiwan. (2022). Knowledge Discovery: Methods from data mining and machine learning. Social Science Research. 110. 102817. 10.1016/j.ssresearch.2022.102817.