IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

A Study On The Comparison Of Web Content Mining Models In Web Mining

[1]Kamaljeet Kaur, [2]Birinder Singh Sarao
[1,2] Assistant Professor, Mata Gujri College, Fatehgarh Sahib

Abstract
Web content mining is a core component of web mining that focuses on extracting meaningful information from the unstructured or semi-structured content of web pages. As the volume of web data continues to grow, effective classification and information retrieval techniques have become essential. This paper presents a comparative study of three widely used machine learning algorithms in web content mining: Naive Bayes (NB)**,** Support Vector Machine (SVM)**,** and Neural Networks (NN)**.** These algorithms are evaluated based on key factors such as classification accuracy, scalability, training time, and suitability for high-dimensional web data. The findings of this study aim to guide researchers and practitioners in selecting appropriate models for tasks such as web page classification, sentiment analysis, and content filtering.

Keywords: Web Content Mining, Naïve Bayes(NB), Support Vector machine(SVM), Neural Networks(NN)

1. Introduction

Web Mining is a specialized area of data mining that focuses on discovering patterns, trends, and useful knowledge from web-based data**.** With the rapid expansion of the World Wide Web, massive amounts of data are generated daily from websites, social media, e-commerce platforms, and other online sources. Web mining provides automated techniques to analyse this data and extract meaningful information, making it a crucial tool for improving search engines, personalizing user experiences, and supporting business intelligence

Web Content Mining is a fundamental branch of web mining that focuses on extracting useful information and knowledge from the content of web pages. This content includes a wide range of data types such as text, images, audio, video, metadata, and structured data like tables and lists. As the web continues to grow exponentially, the sheer volume of unstructured and semi-structured data presents both a challenge and an opportunity for intelligent data analysis.

The primary goal of web content mining is to automatically discover patterns, relationships, and insights within web documents to support tasks like information retrieval, content classification, summarization, sentiment analysis, and recommendation systems. It leverages techniques from fields such as natural language processing (NLP), machine learning, information extraction, and text mining.

2. Web Content Mining

Web content mining is a technique used to extract useful information from the content of web pages, which can include text, images, audio, video, and structured data like tables or lists. Unlike web structure mining, which analyzes the relationships between web pages, or web usage mining, which focuses on user behaviour, web content mining targets the actual data presented on websites. It plays a crucial role in making sense of the vast and ever-growing amount of online information by using tools and techniques from fields such as natural language processing, machine learning, and data mining. Applications of web content mining are widespread, ranging from product recommendation systems and sentiment analysis to news aggregation and academic research. However, it also comes with challenges, including dealing with unstructured content, handling dynamic web pages, and addressing ethical concerns related to data privacy and usage policies

3. Web Content Mining Algorithms

3.1 Naive Bayes

Naive Bayes is a family of simple yet powerful probabilistic classifiers based on Bayes' Theorem. It's called "naive" because it assumes feature independence — meaning each feature contributes independently to the probability of a certain outcome, which is rarely true in real-world data, but the method often performs surprisingly well.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. The "support vectors" are the critical data points closest to the hyperplane, and they directly influence its position and orientation. SVM aims to maximize the **margin** between classes — the distance between the hyperplane and the nearest support vectors — which helps improve generalization to unseen data. SVM can also handle non-linearly separable data using a technique called the **kernel trick**, which maps input data into a higher-dimensional space where separation is easier. Common kernel functions include linear, polynomial, and radial basis function (RBF). Due to its robustness and effectiveness, SVM is widely used in applications like image classification, text categorization, and bioinformatics.

3.3 Neural Networks(NN)

Neural Networks are a type of machine learning model inspired by the structure and functioning of the human brain. They consist of layers of interconnected units called neurons, which process and transmit information. A neural network typically includes an input layer that receives data, one or more hidden layers that perform computations, and an output layer that delivers the final result. Each neuron in the network applies a mathematical function to its inputs, combining them with adjustable weights and biases, and passes the result through an activation function to introduce non-linearity. Neural networks learn from data through a process called backpropagation, where errors are propagated backward from the output to adjust the weights and improve performance. This learning enables them to recognize patterns and make accurate predictions. Variants such as convolutional neural networks (CNNs) are used for image processing, while recurrent neural networks (RNNs) are designed for sequential data like text or time series. Neural networks are widely used in applications like image recognition, natural language processing, speech recognition, and autonomous systems due to their ability to model complex, non-linear relationships in data.

3.4 Decision Tree

Decision Trees are a popular and intuitive machine learning algorithm used for both classification and regression tasks. They work by recursively splitting the dataset into subsets based on the value of input features, creating a tree-like structure of decisions. Each internal node in the tree represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a final decision or prediction. The goal is to create the most efficient splits to minimize impurity, often measured by metrics such as Gini index or information gain. Decision trees are easy to interpret and visualize, making them useful for understanding the decision-making process behind predictions. However, they can be prone to overfitting, especially with complex or noisy data, which is why techniques like pruning or ensemble methods such as Random Forests are often used to improve their performance and generalization.

4. Comparison of Web Content Mining Algorithms

Various parameters are considered for concluding the results in a table which represents the comparison of different algorithms used in web content mining. Table 4.1 represents the different measures for the comparison.

Table 4.1 Comparison on the basis of different aspects

Feature	Naive Bayes	SVM	Neural Network	Decision Tree
Accuracy	Moderate	High	Very High	Moderate
Training Speed	Very Fast	Moderate	Slow	Fast
Interpretability	High	Medium	Low	High
Handles Non-linearity	Poor	Good	Excellent	Moderate
Resource Requirement	Low	Medium	High	Low
Risk of Overfitting	Low	Moderate	High	High
Training Time	Very Fast	Moderate to slow (depends on size of dataset)	Slow(Vary according to dataset size)	Fast
Resource Usage	Low	Moderate to high	High	Low to moderate

5. Conclusion and Future Aspects

In most scenarios, the best algorithm depends on the specific problem, dataset size, complexity, interpretability needs, and resource availability. For quick and simple tasks, Naive Bayes or Decision Trees work well. For high accuracy with moderate data, SVM is a strong choice. For large-scale, complex tasks, Neural Networks are often the most effective.

The future aspects of machine learning algorithms like Naive Bayes, Support Vector Machines (SVM), Neural Networks, and Decision Trees point toward greater integration, optimization, and application across diverse domains. Naive Bayes, valued for its simplicity and speed, is expected to remain relevant in real-time and interpretable systems, with future enhancements focusing on handling feature dependencies and hybridization with more complex models.

6. References

1. Julie Abadji, Pedro Suàrez, Javier Ortiz, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event). Leibniz-Institut für Deutsche Sprache, 1--9.
2. Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. (Jan. 2022). arxiv: 2201.06642
3. Julian Alarte, David Insa, Josep Silva, and Salvador Tamarit. 2018. Main Content Extraction from Heterogeneous Webpages. In Web Information Systems Engineering -- WISE 2018. Springer International Publishing, 393--407.
4. Julián Alarte, Josep Silva, and Salvador Tamarit. 2019. What Web Template Extractor Should I Use? A Benchmarking and Comparison for Five Template Extractors. ACM Trans. Web, Vol. 13, 2 (March 2019), 1--19.
5. Martin Armstrong. 2021. Infographic: How many websites are there? https://web.archive.org/web/20230131222529/https://www.statista.com/chart/19058/number-of-websites-online/Captured: 31 Jan, 2023.

6. Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021. Association for Computational Linguistics, Online, 122--131.
7. Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. CleanEval: a competition for cleaning webpages. In Lrec. European Language Resources Association (ELRA), Marrakech, Morocco.
8. Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003. Extracting Content Structure for Web Pages Based on Visual Representation. In Web Technologies and Applications, 5th Asian-Pacific Web Conference, APWeb 2003, Xian, China, April 23-25, 2002, Proceedings. Springer Berlin Heidelberg, 406--417.