The OSCAR PUBLISHING Services
American Journal of Applied Science and Technology

VOLUME Vol.06 Issue 05 2026
PAGE NO. 113-116
DOI 10.37547/ajast/Volume06Issue05-19

Models and Methods of Intelligent Processing of Web Documents Based on Artificial Intelligence Technologies

Husanov Sherzod Abdimannonovich

Senior Lecturer, Department of Multimedia Technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan

Received: 30 March 2026; **Accepted:** 26 April 2026; **Published:** 16 May 2026

Abstract: This article analyzes modern models and methods of intelligent processing of web documents based on artificial intelligence technologies. Algorithms for automatic processing, clustering, classification, and detection of unnamed events in textual, graphical, and structural data contained in web documents are considered. During the research, machine learning, deep learning, semantic analysis, and DOM tree-based models are studied, and their advantages and disadvantages are compared. In addition, the efficiency of K-means, DBSCAN, Hierarchical Clustering, and BiDist methods in intelligent processing of web documents is analyzed. As a result of the study, an effective model for automatic analysis of web data is proposed.

Keywords: Artificial intelligence, web document, DOM model, clustering, classification, machine learning, deep learning, BiDist, DBSCAN, K-means.

INTRODUCTION

Nowadays, the volume of web documents available on the Internet is growing at a very rapid rate. Effective analysis, processing, and extraction of useful knowledge from these data have become one of the most important challenges. Conventional software approaches are not sufficiently efficient for processing large-scale dynamic web data. Therefore, the use of artificial intelligence technologies is becoming increasingly important.

Intelligent processing of web documents refers to the process of automatically analyzing the information contained in web pages, grouping them, establishing semantic relationships, and generating useful knowledge for users. In these processes, machine learning, neural networks, natural language processing, and clustering algorithms are widely used.

Concept of Intelligent Processing of Web Documents

A web document is structured information located on the Internet in HTML, XML, or other formats. The main characteristics of web documents include the following:

- semi-structured representation;
- dynamic updating;
- consisting of text, image, video, and audio elements;
- representation based on the DOM (Document Object Model) tree structure.

Intelligent processing of web documents based on artificial intelligence is carried out through the following stages:

1. Scanning the web page;
2. Building the DOM tree;
3. Data cleaning;
4. Feature extraction;
5. Clustering or classification;
6. Evaluation of results.

Models for Analyzing Web Documents

DOM Model

The DOM (Document Object Model) represents the structure of a web document in the form of a tree. Each

HTML tag is considered a separate node.

For example:

```
<html>
  <body>
    <h1>Title</h1>
    <p>Text</p>
  </body>
</html>
```

Using the DOM model, it is possible to:

- identify important blocks on a web page;
- monitor user activities;
- detect unnamed events.

Vector Space Model

The Vector Space Model is used to represent web documents in mathematical form. In this model, each document is represented as a vector based on word frequencies.

The TF-IDF model is widely used:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

Where:

- $TF(t,d)$— the frequency of a term in a document;
- $IDF(t)$— the importance coefficient of a term across all documents.

Semantic Model

The semantic model is used to determine meaningful relationships between documents. In this model, deep learning methods such as the following are applied:

- Word2Vec;
- FastText;
- BERT;
- Transformer architecture.

Semantic models are important for:

- identifying similar documents;
- developing automatic recommendation systems;
- improving the efficiency of search engines.

Intelligent Processing Methods

Classification Methods

Classification is the process of assigning web documents to predefined categories or classes. In intelligent web document processing systems, classification algorithms are used to automatically analyze and categorize web pages according to their content, structure, and semantic meaning.

Several algorithms are widely applied in this process, including:

- Naive Bayes;
- Decision Tree;
- Random Forest;
- Support Vector Machine (SVM).

Among these methods, the Naive Bayes classifier is one of the most commonly used probabilistic classification algorithms. It is based on Bayes' theorem and calculates the probability that a document belongs to a particular class according to the occurrence of specific features or words.

The Bayes classifier formula is expressed as:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(A)$$

$$P(B \mid A)$$

$$P(B \mid \neg A)$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \approx 0.68,\ P(B) \approx 0.25$$

P(B)=0.25P(B|A)P(A)=0.17P(A|B)~0.68Posterior = useful evidence / total evidence

Where:

- $P(A \mid B)$— posterior probability of class Agiven evidence B;
- $P(B \mid A)$— likelihood of observing evidence Bin class A;
- $P(A)$— prior probability of class A;
- $P(B)$— total probability of evidence B.

In practical web document classification tasks, the Naive Bayes algorithm estimates the probability of a document belonging to a certain category based on the frequency of words and features contained in the document. This approach is computationally efficient and highly effective for text-based classification problems.

Classification methods play an important role in:

- web page categorization;
- spam detection;
- user behavior analysis;

- recommendation systems;
- intelligent search engines.

These methods improve the automation and accuracy of intelligent web document processing systems.

Clustering methods are used to group documents according to their similarity level. In intelligent web document processing, clustering algorithms make it possible to automatically organize large amounts of web data into meaningful groups without predefined labels. Among the widely used clustering methods are DBSCAN, Hierarchical Clustering, and BiDist approaches.

K-means is one of the most popular unsupervised machine learning algorithms. This algorithm is used to divide data into specific groups, called clusters, based on their similarity level. Today, the K-means algorithm is widely applied in data analysis, artificial intelligence, data mining, web document processing, and recommendation systems.

The main objective of the K-means algorithm is to group data according to the similarity of elements. During the algorithm's execution, the distance between objects is calculated, and each object is assigned to the nearest cluster center. As a result, objects within the same cluster become highly similar to each other, while objects belonging to different clusters remain significantly different.

The distance between objects is commonly calculated using the Euclidean distance formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

BiDist Method. The BiDist method operates based on calculating bidirectional distances between web documents. This method determines similarity and distance by considering:

- DOM structure;
- semantic similarity;
- user behavior patterns.

The BiDist approach is particularly effective for:

- detecting unnamed events;
- analyzing user activities;
- optimizing web pages.

By combining structural and semantic analysis, the BiDist method improves the accuracy of intelligent web document processing systems and enables more effective identification of relationships between web objects and user interactions.

Detection of Unnamed Events

Unnamed events are events that are not explicitly defined by the programmer but occur as a result of user interactions. Examples of such events include page scrolling, closing modal windows, hiding elements, and opening dynamic content blocks.

The proposed model consists of the following stages:

- collecting web pages using a crawler;
- building the DOM tree;
- cleaning the content;
- vectorization based on TF-IDF;
- calculating BiDist distance;
- clustering using the DBSCAN algorithm.

Various web site datasets were analyzed during the research process. The obtained results are presented in the following table:

Method	Accuracy	Speed	Noise Resistance
K-means	High	Very fast	Low
DBSCAN	High	Medium	High
Hierarchical	Medium	Low	Medium
BiDist + DBSCAN	Very high	Medium	Very high

According to the results, the combination of BiDist and DBSCAN was identified as the most effective approach for intelligent processing and analysis of web documents.

CONCLUSION

In this study, models and methods for intelligent

processing of web documents based on artificial intelligence technologies were analyzed. The importance of the DOM model, semantic analysis, clustering, and classification algorithms in processing web data was highlighted. The research results showed that the approach based on BiDist and DBSCAN provides high accuracy in the efficient analysis of large-scale web data. In the future, applying this model in real-time systems, cybersecurity, and intelligent search systems is considered one of the promising research directions.

REFERENCES

1. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2012.
2. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. Pearson Education, 2021.
3. Manning C., Raghavan P. Introduction to Information Retrieval. Cambridge University Press, 2008.
4. Aggarwal C. Machine Learning for Text. Springer, 2018.
5. Witten I., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, 2016.
6. Ester M., Kriegel H.P., Sander J. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, 1996.
7. Blei D. “Probabilistic Topic Models”, Communications of the ACM, 2012.
8. Goodfellow I. Deep Learning. MIT Press, 2016.