academic publishers

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE (ISSN: 2692-5206)

Volume 04, Issue 10, 2024

Published Date: 27-12-2024

FUNCTIONAL STRUCTURE OF THE SYSTEM AIMED AT CREATING AN ELECTRONIC LIBRARY AND IDENTIFYING SIMILAR TEXTS

Jumamuratova Dilafruz

Library Information Activity Direction

Abstract: The creation of an electronic library coupled with the ability to identify similar texts plays a crucial role in the management and validation of content in various fields such as academia, publishing, and legal sectors. This paper discusses the functional structure of a system designed to store, organize, and retrieve documents while leveraging text comparison algorithms to detect similarities or potential plagiarism. The system integrates components such as a Library Management System (LMS), text processing and indexing modules, a similarity detection engine, and an intuitive user interface. By utilizing advanced text comparison algorithms like cosine similarity, Jaccard index, and machine learning models, the system enables efficient text similarity detection, thereby improving content validation and ensuring document integrity. The functional structure also emphasizes scalability, user experience, and the seamless integration of document storage, indexing, and comparison tasks.

Key Words: Electronic library, document management, text comparison, similarity detection, plagiarism detection, NLP, text indexing, cosine similarity, Jaccard index, machine learning, database management, scalability, user interface.

In the digital age, the management and organization of textual content have become essential for academic, legal, and publishing industries. The rapid increase in the volume of digital content necessitates the development of robust systems that not only store and organize documents but also facilitate content validation and integrity checks. One of the primary challenges in handling large collections of documents is ensuring the authenticity and originality of the content, particularly when it comes to identifying similarities or potential plagiarism between texts. An electronic library system that incorporates text similarity detection tools can significantly enhance content management by automating the process of detecting duplicated or similar text. Such systems leverage advanced algorithms and natural language processing (NLP) techniques to analyze and compare documents, facilitating the detection of even subtle instances of plagiarism or content overlap. This is particularly important in academic and legal contexts, where the integrity of content is crucial for maintaining credibility and avoiding legal ramifications

This paper explores the functional structure of a system designed to create an electronic library that not only stores and organizes documents but also identifies similar texts through various text comparison methods. The system is designed with the goal of improving the efficiency of document retrieval, enhancing content validation processes, and ensuring that document integrity is upheld. Key components of the system include a Library Management System (LMS), text processing and indexing modules, a similarity detection engine, and an interactive user interface for seamless interaction. The paper also discusses the technologies and algorithms employed in text comparison, such as cosine similarity, the Jaccard index, and deep learning-based models, to ensure high accuracy and scalability in identifying similarities across large datasets.

By combining document management with advanced text analysis techniques, this system offers a

comprehensive solution for managing electronic libraries while addressing challenges related to text similarity, plagiarism detection, and content verification. [1] The increasing demand for such systems across various sectors underscores the importance of integrating these functionalities into modern content management frameworks.

1. Electronic Library Management System (LMS)

An electronic library management system (LMS) forms the backbone of any content storage and retrieval system. The LMS manages the organization, storage, and retrieval of documents in the library. It ensures that all documents are easily accessible and well-categorized based on metadata such as titles, authors, keywords, publication dates, and document types. The main functionalities of the LMS include:

- **Document Upload and Import:** The system should support various formats (e.g., PDFs, DOCX, HTML) and have tools for extracting and standardizing the content for further processing. Batch uploads can also be enabled for efficient management of large datasets.
- **Metadata Management:** Each document should have associated metadata for quick categorization and searchability. Metadata includes attributes such as title, author, publication year, genre, and any other relevant tags for efficient document indexing.
- **Document Organization and Retrieval:** The system allows users to categorize documents into folders, apply tags, or create thematic collections. Full-text search functionality can be enabled by using indexing technologies like Elasticsearch or Apache Solr to retrieve documents quickly based on content, metadata, or keywords.

2. **Text Processing and Indexing Module**

Once documents are uploaded, the system processes the text to prepare it for comparison. This processing includes several steps such as text extraction, cleaning, and tokenization, which make the content ready for comparison and search. This module is vital for improving the efficiency of text retrieval and similarity detection. The key functions of this module include:

- **Text Extraction:** In the case of PDF, scanned documents, or images, Optical Character Recognition (OCR) technology (e.g., Tesseract) is used to extract readable text. This process ensures that even non-text formats are accessible for analysis.
- **Text Cleaning and Normalization:** The extracted text is cleaned to remove unnecessary characters, such as special symbols, extra spaces, or formatting inconsistencies. Additionally, stop words (commonly occurring words like "the", "and", etc.) are removed, and all text is converted to lowercase to ensure uniformity and accuracy in indexing.
- **Tokenization and Lemmatization:** The system divides the text into smaller units (tokens), such as words or phrases, and performs lemmatization, which reduces words to their base or root form. For example, "running" becomes "run". This helps normalize variations of the same word for better indexing and comparison.
- **Indexing for Retrieval:** The processed text is indexed using various algorithms like **TF-IDF** (Term Frequency-Inverse Document Frequency) or **word embeddings** for semantic indexing. These indices allow for fast retrieval based on content and support efficient similarity detection across the library.[4]

The functional structure of a system designed to create an electronic library and detect similar texts combines key components such as document management, text processing, and advanced similarity detection algorithms. By integrating these elements, the system not only offers efficient storage and retrieval of documents but also provides powerful tools for plagiarism detection and content verification. The use of text comparison techniques, from simple algorithms like cosine similarity to sophisticated machine learning models, ensures high accuracy in identifying similar or duplicate content. This system represents a comprehensive solution to the challenges faced by industries in managing large volumes of digital text while ensuring the authenticity and integrity of content.

References:

1. **Pervichko, I. A.** (2020). The Role of Psychological Support in Reducing Stress in Schoolchildren. Moscow: Russian Psychological Society.
2. **Vasilenko, T. P., & Pukinel, A. N.** (2019). Student Stress: Causes, Consequences, and Coping Strategies. St. Petersburg: Academic Publishing House.
3. **Kulikov, A. M.** (2018). Stress and Mental Health of Adolescents in Russian Schools: Current Challenges and Solutions. Journal of Russian Education and Psychology, 5(4), 102-110.
4. **Semyonova, T. S., & Galkin, V. V.** (2017). Academic Stress and its Influence on the Psychological State of Students in Higher Education. Russian Journal of Education Psychology, 29(3), 134-145.
5. **Shevchenko, N. M.** (2021). Mental Health and Stress Management Among Russian High School Students. Moscow: Institute of Educational Psychology.
6. **Mikhaylova, T. G., & Baranova, V. M.** (2022). Coping Strategies and Resilience in Russian High School Students. Psychological Research, 11(2), 74-83.