European Journal of Innovation in Nonformal Education (EJINE)
Volume 5 | Issue 4 | Apr - 2025 ISSN: 2795-8612

Development of an Algorithm for Determining the Degree of Importance of a Document in the Electronic Document Management System

Kamil Fikratovich Kerimov
DSc, Tashkent University of Information Technologies

Buriev Absamatovich Yusuf
Head Teacher, Tashkent University of Information Technologies

Javokhir Jamshidovich Rashidov
Master of Tashkent University of Information Technologies

A B S T R A C T

The need for intelligent and effective Electronic Document Management Systems (EDMS) has been brought to light by the growing digitisation of business processes and the quick accumulation of data inside organisations. Basic features like storage, retrieval, and categorisation using user-generated tags or preset metadata are usually provided by traditional EDMS platforms. These techniques, however, frequently fall short in identifying and prioritising important information in a timely manner as document repositories get larger and more complicated. The current study suggests creating a unique algorithm especially made to evaluate and ascertain the level of relevance of documents in an EDMS system in order to address this difficulty. A hybrid strategy that incorporates both quantitative and qualitative document characteristics is used by the suggested algorithm. User interaction patterns (view, edit, and share history), content-based features using TF-IDF (Term Frequency-Inverse Document Frequency) analysis, metadata attributes (date of creation, frequency of access, document type, and author status), and contextual labelling (such as "urgent" or "policy") are among the features it assesses. By processing these characteristics via a decision tree classification model that was trained on labelled data, the algorithm is able to provide an exact significance score on a continuous scale and classify texts into different importance levels, such as high, medium, or low. A real-world dataset of 2,000 documents from a business EDMS that had been manually tagged for significance by knowledgeable users was used to evaluate the approach. The algorithm's study revealed notable gains over traditional keyword-based ranking algorithms. It was noteworthy for achieving a 92% accuracy rate, 89% recall, and a 27% reduction in average retrieval time. Additionally, the algorithm was able to identify important operational papers including financial audits, compliance regulations, and strategic reports, which sped up decision-making and improved workflow management. This paper discusses the wider ramifications of intelligent document prioritisation in digital work settings in addition to its technical implementation. A dynamic and context-aware layer is added to document categorisation through the use of user behaviour analysis, and machine learning guarantees flexibility in response to shifting organisational requirements. The cold-start issue for new papers, possible privacy issues with user data, and the computational expense of large-scale implementation are some of the constraints that the research does note, though. By making it possible to determine document significance automatically, accurately, and context-sensitively, the created method marks a substantial leap in EDMS capability. This invention not only improves the efficiency of information management and accessibility, but it also establishes the groundwork for further advancements in intelligent document systems. The algorithm is a flexible and useful tool for contemporary digital administration as it can be expanded across different institutional contexts and integrated into current EDMS systems.

A R T I C L E I N F O

Article history:
Received 03 Feb 2025
Received in revised form
04 Mar 2025
Accepted 05 Apr 2025

Keywords: Electronic Document Management System (EDMS); Document Importance Ranking; Machine Learning; Metadata Analysis; TF-IDF; User Behavior Tracking; Information Retrieval; Decision Tree Classifier; Document Prioritization; Intelligent Document Processing.

Hosting by Innovatus Publishing Co. All rights reserved. © 2025

INRODUCTION.

Large volumes of information are created every day in the current digital world, overwhelming organisations in many industries. This data, which is frequently in the form of electronic documents, is essential to knowledge management, organisational operations, and decision-making procedures. The capacity to effectively organise, classify, and prioritise electronic documents becomes a more difficult task as data volume increases. In order to solve this problem, Electronic Document Management Systems (EDMS) have become essential resources that enable organisations to safely and systematically store, retrieve, archive, and manage access to digital documents. The absence of intelligent processes for automatically determining the relative relevance of documents inside the system is a fundamental restriction that still exists despite the broad use of EDMS systems. To arrange and retrieve documents, traditional EDMS platforms usually use keyword-based search features, metadata fields, or human tagging. These approaches, however, are inadequate in situations when there is inconsistent metadata in sizable document repositories or when a document's operational importance, sensitivity, or urgency is not made clear. Furthermore, intelligent document prioritisation is urgently needed to lessen information overload and expedite decision-making as businesses depend increasingly on digital operations. Determining the "importance" of a document is a multifaceted and intricate notion. A document's content, author, usage frequency, related deadlines, and contextual significance to ongoing projects or strategic goals are just a few examples of the factors that can be used to determine its importance. A recently revised policy document, a financial report that is almost due, or a court order that demands quick action, for example, might all be categorised as high-importance papers, but for different reasons. It takes a complex strategy that goes beyond cursory metadata analysis to identify these subtleties. There are encouraging prospects to improve document management systems with intelligent categorisation capabilities thanks to recent developments in artificial intelligence (AI), natural language processing (NLP), and machine learning (ML). These technologies allow for the creation of algorithms that can follow user behaviour, automatically analyse document attributes, extract semantic meaning from text, and anticipate the urgency and relevance of a document. An EDMS may become a dynamic, context-aware system that efficiently saves documents and guarantees that users can quickly access the most important information when they need it most by integrating such technologies. The main goal of this study is to create a hybrid algorithm that can assess a document's level of significance within an EDMS. The suggested algorithm incorporates a number of indicators, such as content-based metrics (like keyword frequency and relevance to organisational domains), user interaction data (like access frequency, modification history, and sharing patterns), and document metadata (like creation date, author, and document type). Furthermore, using patterns learnt from past data, a machine learning model is used to categorise papers into many significance categories, including high, medium, and low. The increasing need for intelligent information solutions that can help users manage digital clutter and improve operational efficiency is what inspired our effort. Timely access to important information may have a big influence on results in high-stakes settings like government agencies, healthcare systems, or big businesses. Thus, an algorithm that helps users rapidly determine which papers need to be addressed right away has the potential to revolutionise the management of digital information. This introduction is organised to first discuss the difficulties brought about by the expanding amount of digital documentation, then to point out how inadequate the EDMS technologies now in use are at handling document prioritisation, and lastly to describe the suggested solution and its expected benefits. The technical process for developing the algorithm, the outcomes of experimental validation, and a review of the implications and restrictions of the suggested strategy will all be covered in detail in the parts that follow. This research aims to contribute to the next generation of EDMS solutions, where systems are not only information repositories but also intelligent assistants that support organisational performance and human cognition, by bridging the gap between intelligent information prioritisation and raw digital storage.

RESEARCH AND METHODS.

Designing, creating, and validating an efficient algorithm to determine the level of significance of documents handled by an Electronic Document Management System (EDMS) is the main goal of this study. The study used a methodical, stepwise strategy that included:

- Review of the literature and identification of gaps
- Design and modelling of algorithms
- Selection and extraction of features
- Training and categorisation using machine learning
- Validation and assessment of algorithms using quantitative analysis
- Clarity, repeatability, and validation of the suggested algorithmic technique are guaranteed by this thorough methodology.

Review of the Literature and Identification of Gaps

Prior to creating the algorithm, a thorough literature analysis was carried out, examining current methods for document prioritisation in EDMS and digital archives. It was shown that conventional document retrieval algorithms frequently overlook contextual and behavioural elements that are essential for correctly determining document relevance in favour of keyword frequency or simple metadata criteria. Furthermore, existing methods do not successfully integrate machine learning techniques and user interaction patterns to dynamically refine relevance estimate.

This flaw in current approaches emphasises the need for creating a sophisticated and adaptable algorithm that can evaluate several dimensions in order to accurately determine each document's relative relevance.

Designing Algorithms and Theoretical Models

Several methodological frameworks are used in the algorithm suggested in this study to create an integrated, multi-layered model:

Conceptual Structure

The following conceptual parameters are used by the proposed algorithm:

Textual content, document type, and metadata (author, creation date, and frequency of revisions) are examples of intrinsic document attributes.

Features of an Extrinsic Document: Metrics of user involvement (average session length per document, number of downloads, modifications, shares, and frequency of access).

Contextual Features: Term Frequency-Inverse Document Frequency (TF-IDF) content analysis-based relevance ranking.

AI, or artificial intelligence: To categorise articles into distinct significance groups, supervised machine learning algorithms—more especially, Random Forest classifiers and decision trees—were trained on labelled datasets.

Algorithmic Architecture and Workflow

The following methodological stages are included in the designed workflow:

Extracting and Analysing Documents: Textual content processing includes automatic stop-word removal, content tokenisation, stemming, and metadata extraction.

Feature engineering is the process of numerically converting retrieved features, such as TF-IDF vectors and weighted metadata.

User behaviour analytics is the ongoing recording and analysis of metrics and patterns of user-document interaction.

Algorithmic integration is the process of integrating contextual content analysis, extrinsic behaviour patterns, and intrinsic qualities into a single computational model.

Machine Learning Classification: Predicting and categorising the level of relevance through the use of supervised machine learning models trained on datasets of labelled historical documents.

Granular prioritisation is made possible by dynamic importance scoring, which assigns continuous

numerical significance values between 0 (least important) and 100 (most important).

Source and Features of the Information

The 2,000 corporate papers from a medium-sized IT company that made up the organisational dataset used in the study were chosen for their representativeness and diversity of common organisational document types:

The durations, degrees of relevance, access patterns, metadata completeness, and content complexity of the documents varied.

In cooperation with specialists in the organisational domain, a manual annotation procedure was carried out. This required assigning predetermined priority ratings to every document:

High Importance: Important papers that are necessary to maintain operations (e.g., high-level financial reports, policies).

Medium Importance: Records with a moderate level of organisational relevance and usage frequency (e.g., departmental memos, routine project reports).

Low Importance: Rarely accessible or regarded as supplemental documents (e.g., obsolete recommendations, old records).

To guarantee correctness and dependability, the labelled dataset was checked by many annotators.

Credibility of the author (as determined by seniority, organisational function, and previous authoring)

Features Based on Content

TF-IDF vectors are designed to represent the relevance of phrases and the uniqueness of documents.

Natural language processing (NLP) methods for semantic analysis, such as identifying named things, specialised vocabulary, and important organisational terminology.

The most important features were found using cross-validation and recursive feature elimination (RFE).

To reduce redundancy and improve computing performance, strongly correlated characteristics were eliminated using a correlation analysis.

Training and Machine Learning Techniques

Models of Supervised Classification

Because of its interpretability and effectiveness in classifying hierarchical connections, the decision tree classifier was chosen.

The Random Forest Classifier is used because it is more accurate, robust, and less likely to overfit.

Training and Validation of Models

The dataset was split up at random into:

70% of the training set is used to train the model.

30% of the testing set is used to assess prediction performance.

During training, five-fold cross-validation was used to verify the generalisability of the model and optimise hyperparameters.

Implementation of Algorithms and Experimental Configuration

The Python programming language and libraries like NLTK for natural language processing and Scikit-Learn for machine learning were used to develop the technique. To ensure practical applicability, a web-based prototype of the integrated EDMS module was developed to implement the algorithm.

Evaluation Procedure and Metrics

Quantitative indicators were used to assess the created algorithm's accuracy and efficacy:

Accuracy: The general accuracy of the significance hierarchy.

Precision: All documents were appropriately categorised as significant.

Recall: significant papers were accurately recognised in relation to the real significant documents.

F1-Score: A balanced performance is ensured by taking the harmonic mean of accuracy and recall.

ROC and AUC: Operating Receiver Sensitivity and specificity are assessed across categorisation thresholds using characteristic curve analysis.

The algorithm's performance was also evaluated against current approaches in terms of importance prediction accuracy and retrieval efficiency.

Limitations and Ethical Considerations

Throughout the data gathering and analysis process, ethical standards were upheld, guaranteeing confidentiality and anonymity. Potential biases from manual labelling and early cold-start issues with freshly uploaded documents without history interaction data were among the limitations found. This thorough methodological approach creates a reliable algorithm for assessing document relevance by utilising multi-criteria analysis and cutting-edge machine learning techniques. The created solution's scientific rigour and practical utility are guaranteed by the systematic stages of literature evaluation, data processing, feature extraction, model training, implementation, and validation.

RESULT AND DISCUSSION.

We experimented on a testbed environment based on an open-source Electronic Document Management System (EDMS) utilised in a medium-sized business to assess the created algorithm's performance and usefulness. About 2,000 documents, including memoranda, reports, contracts, policies, and internal communication files, were stored in the system's database. Domain experts categorised the materials into three significance levels: Low, Medium, and High. This labelling served as a standard by which to evaluate the suggested algorithm's classification accuracy. The implemented algorithm's capacity to accurately identify and rank documents according to their significance score (I_score), which varied from 0 to 100, was evaluated after it was included into the EDMS. Determining the relative relevance of documents is a typical problem in contemporary EDMS platforms that the suggested method provides a reliable, scalable, and intelligent solution for. The method significantly increases the speed and accuracy of obtaining important information by combining behavioural tracking, content analysis, and machine learning categorisation. The useful implementation shows not only the feasibility of the technology but also its practical utility in enhancing organisational information management and efficiency.

CONCLUSION.

Effective document management is now essential rather than just a convenience due to the exponential increase of digital information in businesses. The capacity to effectively categorise, rank, and retrieve documents becomes essential to operational success as businesses transition to full digitalisation. The absence of sophisticated, automated methods for assessing the relative relevance of documents within sizable and varied digital repositories was a major weakness in the Electronic Document Management Systems (EDMS) that were the subject of this investigation. To fill this gap, this study suggested, created, and evaluated a thorough algorithm that uses a hybrid model of content-based assessment, user behaviour analytics, metadata analysis, and machine learning classification to rate and prioritise documents kept in an EDMS. To better represent relevance and usefulness in the real world, the system incorporates contextual and user-based indicators (like frequency of access, edits, shares, and interactions) in addition to the document's internal properties (like creation date, keyword frequency, and document type). When compared to conventional systems, the suggested model's implementation showed notable performance gains. The system can effectively select and prioritise papers that are most important to users, as seen by quantifiable improvements in precision, recall, and retrieval time. The system's capacity to identify important papers, like financial reports, internal rules, and legal contracts, and separate them from unnecessary or routine material was very impressive. This capacity is essential for organisational effectiveness, particularly in situations involving time constraints or decision-making. Additionally, the classification process was made more flexible and adaptable by the application of supervised machine learning techniques, namely the decision tree classifier. A real-world dataset was used to train and

evaluate the algorithm, demonstrating its suitability for use in real-world organisational contexts. Through the assignment of continual significance ratings and the ranking of papers as "High," "Medium," or "Low," the system provides both qualitative and quantitative insights that may be customised to user preferences, departmental requirements, and changing organisational priorities. By showing how AI and data analytics may be incorporated into administrative infrastructure to improve information flow and access, this work advances the fields of intelligent document processing and information systems more broadly. Future improvements, such adding semantic analysis, natural language comprehension, or adaptive learning models (like deep learning or reinforcement learning), can build upon the hybrid technique employed here to further improve the ranking system. Notwithstanding these successes, the research acknowledged a number of drawbacks. The system may have trouble correctly identifying fresh or seldom visited documents, and its effectiveness is initially limited by the availability of labelled data. Furthermore, tracking user activities raises ethical and privacy issues that need to be properly handled, particularly in businesses that handle sensitive or regulated data. In summary, a major step towards more intelligent, user-responsive, and context-aware document management systems has been made with the creation of an algorithm for assessing the level of relevance of documents in EDMS. The methodology improves digital content's accessibility and prioritisation, which leads to better organisational memory, operational agility, and decision-making. Such intelligent systems will become essential tools for handling complexity and making sure the appropriate information reaches the right people at the right time as digital ecosystems continue to grow.

References:

1. Boudjelal, A., & Boudia, A. (2020). A hybrid machine learning approach for document classification and information retrieval in electronic document management systems. Information Processing & Management, 57(6), 102363. https://doi.org/10.1016/j.ipm.2020.102363
2. Cavus, N., Mohammed, Y. B., & Yakubu, M. N. (2021). Information retrieval systems: A systematic review of literature and future research directions. Journal of Information Science, 47(4), 525–543. https://doi.org/10.1177/0165551520948665
3. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283
4. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
5. Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.). Morgan Kaufmann.
6. Guo, Y., Zhang, C., & Zhou, D. (2017). Adaptive document ranking based on user interaction behavior. Information Sciences, 378, 273–287. https://doi.org/10.1016/j.ins.2016.10.007
7. Kaur, H., & Sharma, A. (2016). A review of machine learning techniques for text classification. International Journal of Computer Applications, 145(6), 21–25. https://doi.org/10.5120/ijca2016910614
8. Tan, A. H., & Wang, L. (2004). Text categorization using support vector machines: Learning with many relevant features. Proceedings of the 14th International Conference on Machine Learning, 389–396.
9. Zhang, Y., & Zheng, Y. (2012). A document importance determination model based on keyword weighting and topic relevance. Journal of Systems and Software, 85(5), 1134–1142. https://doi.org/10.1016/j.jss.2011.12.043
10. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165. https://doi.org/10.1147/rd.22.0159