TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES

COLLEGE OF SCIENCE

DEVELOPMENT OF WEB-BASED INFORMATION RETRIEVAL SYSTEM OF ACADEMIC PAPERS WITH AUTOMATIC TAGGING USING PARTICLE SWARM OPTIMIZATION ALGORITHM IN THE TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES – MANILA

A Research Presented to the
Faculty of the College of Science
Technological University of the Philippines, Manila

In Partial Fulfillment
of the Requirements for the subject
CC303, Methods of Research in Computing

by:

Brillo, Alexandria Lee G.

Capistrano, Deazelle M.

Maximo, Raine Francesca

Velasquez, Erika F.

Villaruz, Maria Lourdes T.

Professor Dolores Montesines

S.Y 2024 – 2025

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

This chapter provides an overview of the context, the subject under consideration, and the significance of the study. The study covers the objectives, scope, and limitations of the research. This chapter will present a comprehensive summary of the research and a concise analysis of the concepts addressed in the study.

Introduction

This study presents the development of a web-based Information Retrieval System (IRS) for Academic Papers, also known as Research Papers. The system incorporates Automatic Tagging using the Particle Swarm Optimization (PSO) algorithm. Its goal is to improve research paper accessibility for students and faculty at the Technological University of the Philippines, Manila Campus. The system offers a convenient and efficient way for users to locate materials that are relevant to their respective courses. According to the study of Martin-Martin et al. (2020), academic research greatly depends on the ability to obtain and investigate scholarly articles, making Information Retrieval Systems (IRS) extremely important for researchers.

Background of the Study

In today's digital age, where information is readily available online, the ability to find and evaluate credible sources is essential (Long, C. et al., 2021). The internet has become the primary source of information for many individuals (Feng, S. et al., 2021). In academic research, information retrieval plays a crucial role allowing scholars to access and explore a wide range of scholarly publications (Zhang et al., 2020). Academic research heavily relies on accessing and exploring scholarly publications, making Information Retrieval Systems crucial for scholars

(Martín-Martín et al., 2021). These systems provide scholars with the ability to search for relevant papers in their specific area, helping them stay updated with the latest research and findings (Chen & Zhang, 2021). While search engines provide convenience and efficiency in information gathering (Giomelakis Dimitrios et al., 2021), students at the Technological University of the Philippines, Manila often face challenges in accessing high-quality academic research materials tailored to their specific courses.

The Technological University of the Philippines (TUP) is a state university in Manila, Philippines, established in 1901 as the Philippine School of Arts and Trades, which later became the Philippine College of Arts and Trades in 1937. TUP was officially established on July 11, 1978, through Presidential Decree No. 1518, which transformed the Philippine College of Arts and Trades into the Technological University of the Philippines. The university is mandated to provide higher and advanced vocational, technical, industrial, technological, and professional education and training in the industries, technology, and practical arts leading to certificates, diplomas, and degrees.

TUP is recognized for its excellence in engineering and technology education, offering a wide range of programs across various colleges, including the College of Engineering, College of Industrial Technology, College of Industrial Education, College of Architecture and Fine Arts, College of Science, and College of Liberal Arts. The College of Industrial Technology, in particular, has its roots in the Technical Department of the Philippine School of Arts and Trades, which was established in 1937. The university is committed to providing quality education, progressive leadership in applied research, developmental studies in technical, industrial, and technological fields, and production using indigenous materials. TUP also aims to effect technology transfer in the countryside and assist in the development of small and medium-scale industries in identified growth centers.

To facilitate research and training, TUP has established the Integrated Research and Training Center (IRTC), which aims to provide up-to-date training and research outcomes. The TUP Library System, which consists of libraries on four campuses, subscribes to the philosophy of cooperation and partnership, aiding students, and faculty in search of diverse academic library collections. Despite the vast amount of information available, students at TUP struggle to find current and relevant research papers.

One common challenge they face is the difficulty of locating current and pertinent research papers tailored to their specific academic courses. Postgraduate students and early career academics face challenges in choosing a practical approach, designing an efficient search strategy, locating relevant literature, determining the appropriate scope, and effectively synthesizing and critiquing the literature (Daniel, 2022). This struggle often arises due to the constant evolution of research in their field, making it a challenge to pinpoint the most up-to-date and relevant sources. In addition to this issue, students may also have limited access to past research conducted within the school or institution (Miller, 2023). To address these challenges, students often need to employ effective search strategies, leverage academic databases, and seek guidance from professors or librarians to ensure they access the most appropriate materials for their studies.

Recognizing the difficulty of accessing research at TUP Manila Campus, this study aims to develop an Information Retrieval System (IRS) for academic papers that can provide TUP students with a more efficient and effective way to search and access relevant information. By integrating advanced search algorithms, accessibility, and interoperability with other IRS and academic platforms, the proposed IRS aims to provide a more efficient and effective way for TUP students to search and access research papers within the school. This system will enhance students' research capabilities, enabling them to navigate the academic literature landscape with ease and precision, thereby fostering a more productive and engaging research experience.

Objective of the Study

The primary objective of this project is to develop a web-based Information Retrieval System (IRS) for the Technological University of the Philippines, Manila Campus, aimed at efficiently locating research papers conducted by students. This project specifically aims to:

- **Improve Accessibility**: Enhance the accessibility of research papers for both students and faculty members by implementing a user-friendly system that allows easy retrieval of relevant information.
- **Automatic Tagging with PSO Algorithm**: Implement the Particle Swarm Optimization (PSO) algorithm for Automatic Tagging of uploaded PDFs, facilitating effective categorization and organization of research papers based on content.
- **Search Functionality**: Develop a robust search functionality within the system to enable users to quickly and accurately locate specific research papers based on keywords, authors, topics, or other relevant criteria.
- **Systematic Organization**: Implement a systematic organization and categorization system for research papers, making it easier for users to navigate and find papers within specific subject areas or academic disciplines.
- **User Authentication**: Establish a reliable user authentication system with tiered permissions to control access to specific functionalities, ensuring that only authorized users can upload, edit, or delete research papers for enhanced security.

Significance of the Study

College students are required to complete a project known as a *thesis*, which often involves both written documentation and a practical component that varies depending on the specific course of study. Students build up and retain their skills and knowledge acquired during their four to five years stay in college in order to comply, demonstrate, and validate their abilities. This study will redound to benefit the following:

To Students, it provides easy access to past research conducted by TUP-Manila students, making it easier to find articles that are relevant to their courses. This resource reduces the process of finding appropriate materials, promoting a cooperative and enhanced academic atmosphere within the TUP-Manila community.

In the Field of Technology, is to provide an understanding of the creation of advanced technologies that are designed to make educational processes more efficient. The ultimate objective is to utilize innovation to greatly improve and elevate the overall standard of education by incorporating state-of-the-art technological solutions.

For Future Researchers, use this as a reference for further development and a guide for future studies, offering a blueprint to navigate the complexity of their research endeavors.

For Future Research Projects, it can function as a point of reference to improve and perfect the current systems. By utilizing the knowledge and techniques established in this research, researchers can expand upon a solid foundation, facilitating the development of innovative progress and the establishment of more precise systems. This reference point serves the purpose of not only duplicating successful models but also creating opportunities for customizing solutions to specific domains or enhancing discovered limitations during the research.

For the Higher Education Institutions, it serves as a method to centralize and exhibit all students' research papers. This study aims to showcase the combined scholarly work and assist students in finding relevant research for their courses. It promotes innovation and academic achievement within the university.

Scope and Delimitation

This project aims to develop a web-based IRS for research papers, specifically for TUP, Manila Campus. The primary focus is to enhance the accessibility of research papers for students and faculty by providing a convenient and efficient means of locating materials relevant to their respective courses.

An innovative aspect of this system is the implementation of the Particle Swarm Optimization (PSO) algorithm, which will be responsible for the Automatic Tagging of research papers based on their content. This will significantly improve the accuracy and speed of the retrieval process. However, this project is limited to the scope of the university's research papers and does not extend to external publications or resources. Additionally, Automatic Tagging will be constrained by the accuracy of keyword extraction and the predefined course categories established within the system.

CHAPTER II

CONCEPTUAL FRAMEWORK

This chapter presents the review of related literature, related studies, a conceptual model, and the definition of terms that are viewed by the researchers and have a significant bearing on the present study.

Review Related Literature

Information Retrieval System (IRS)

Information plays a crucial role in our everyday lives. Over time, the significance of storing and retrieving information has become widely recognized. With the swift advancement of technology, it has become simpler for individuals to store vast amounts of data and extract valuable information from it. Information retrieval involves enabling users to find relevant information within unstructured documents (Wable R. et al.,2021).

Information retrieval involves the science of locating information within documents, finding the documents themselves, and searching for metadata, as well as databases containing texts, images, or sounds (Wable R. et al.,2021). The purpose of an IRS is to provide accurate information to the user. To accomplish this, information is meticulously stored, collected, and organized across different subjects, making it easily accessible when required. In modern libraries and archives, information retrieval includes searching full-text databases, finding items from bibliographic databases, and document delivery through a network (Dr. Manjunatha S et al., 2022). In information retrieval, searches can utilize full-text or other forms of content-based indexing. As stated in the study of Agbele (2018), users frequently input ad-hoc keywords in their search queries, which are personalized terms not pre-defined within the system. It becomes the responsibility of information retrieval systems (IRS) to precisely grasp the user's information requirements and contextual nuances. Keywords play a pivotal role in information retrieval, as they dictate document relevance, assist in document classification, and streamline the indexing process (Pin Ni et al., 2020).

According to Sahal Manasia et al. (2023), the importance of archiving and retrieving information has been recognized for several years. With the widespread use of computers, the ability to extract valuable information from large collections has become essential. As a result, information retrieval has emerged as a significant research area within computer science and has gained prominence across various fields such as business, healthcare, agriculture, medicine, law, and others. Information retrieval involves locating material, often in unstructured document form, that contains the necessary information.

Moreover, the challenges identified in information retrieval among engineering students underscore the pressing need for advancements in search algorithms and user interfaces tailored to the specific requirements of diverse fields. This highlights the ongoing evolution and refinement of information retrieval techniques to address the intricacies of accessing relevant data in specialized domains.

In the study by Navitas et al. (2022), which analyzed the types of online database websites favored by students as their primary information sources, investigated the underlying reasons for their preference of these websites, and identified the challenges encountered during information seeking through online databases, it was found that engineering students primarily rely on search engines as their main information source due to their richness and flexibility. However, despite this reliance, these students often face challenges in effectively retrieving information using different keywords, primarily because the broad scope and complexity of engineering topics require precise and nuanced search queries, making it difficult to find relevant information amidst the vast amount of available data.

Moreover, in the study of Wu et al. (2022), it is emphasized that a computer-based legal information retrieval system serves as a pivotal tool in navigating the complexities of legal data. Through its comprehensive search functionality, users can efficiently access a plethora of legal resources, ranging from statutes to case law. Moreover, the system offers curated legal information content, ensuring users have access to accurate and up-to-date data essential for decision-making processes. Additionally, by providing robust management services, the system facilitates seamless organization and retrieval of information, contributing to enhanced productivity and efficiency. As highlighted in their study, the adaptability of such systems to changing demands underscores their significance in driving technological advancements and fostering growth in our economic society.

Furthermore, according to Lima et al. (2022), in their paper titled "Information Storage and Retrieval System: An Analysis of the Impact of Variables and Measures Aimed at the Organization and Retrieval of Information Centered on the User," the authors assert that the efficacy of an information retrieval system relies on several key components. Chief among these is the quality of organization within the system, which directly influences its ability to manage and categorize data effectively. Additionally, they emphasize the importance of timely and precise retrieval of the most relevant information, ensuring that users can access the data they need efficiently. Furthermore, the authors highlight the significance of adopting a systemic perspective, with the user positioned as the central focus of the information retrieval process. This user-centric approach acknowledges the diverse needs and preferences of users, thereby enhancing the usability and overall effectiveness of the system. In summary, the authors suggest that meticulous organization, targeted retrieval strategies, and a user-centric approach are critical for optimizing the performance of information retrieval systems.

Moreover, according to Yu, B. (2019), the integration of domain ontology improves information retrieval systems. The proposed model includes document processing, ontology document retrieval, an information database, and query transition and retrieval agent modules. A key innovation is the use of a genetic algorithm to optimize weighted factors of word frequency, which improves precision and recall rates compared to simulated annealing. This leads to a better understanding of users' requirements. For future research, Yu suggests using data mining to establish an ontology database to address the challenges of ontology creation. Additionally, developing personalized query preferences aims to tailor retrieval results to individual user demands, enhancing system effectiveness and user satisfaction.

Automatic Tagging

According to Krishnapriya et al., (2017), when there are multiple research papers in a tagging mechanism, manual tagging becomes time-consuming and prone to error. This paper introduces a document-centered approach that automates the tagging of research papers, returning precise tags associated with the papers. Prior to tagging, we categorize each research document into its respective domain. Because the majority of research papers consist of multiple domains, it is necessary to specify both the primary domain and the sub-domains to which each uploaded paper pertains. As classification improves the tagging process's efficiency and refinement, it has no effect on the tagging procedure. Research papers within the same domain generally share a greater number of common tags than papers spanning different domains. The majority of currently implemented tag recommendation systems utilize term frequency-inverse document frequency

(TF IDF) as the information retrieval weight. We dynamically update the training dataset with the highest-scoring keywords deemed pertinent to the research paper to minimize the need for manual keyword annotation. Given the laborious and mistake-prone nature of adding new phrases, we regard this supplementary functionality as one of the system's primary benefits.

According to Hult A. and Megyesi B. (2019), the objective of automatic text categorization is to classify a document into one of a predetermined set of categories. The predominant methodology is supervised machine learning, which trains an algorithm using documents from predetermined categories. Before any learning occurs, the documents must be converted into a format that the learning algorithm can understand. We thus apply a trained prediction model to assign the categories to previously unseen documents. To accomplish a text categorization task, one must make two significant choices: how to represent the text and which learning algorithm to use to construct the prediction model. In most research studies, the most effective representation is the complete text, with the tokens in the document maintained in a distinct manner, specifically as unigrams. However, recent years have seen several experiments evaluating richer representations. This is shown by the work of Caropreso et al. (2001), Moschitti et al. (2004), Kotcz et al. (2001), and Mihalcea and Hassan (2005), which use complex nominals in their bag-of-words representation and run experiments where the representation is fed automatically extracted sentences. Of these three instances, sentence extraction appears to be the only one where the automatic text categorization performance improved. Furthermore, we specified that the term "extracted keywords" signifies the inclusion of the chosen terms in their exact form in the document. A keyword may comprise a single token or multiple tokens. A keyword may also consist of an entire expression or phrase, as in the case of "snakes and ladders."

In the study of Choe et al., (2019), the implementation of tagging or linking across HTML texts has been vital to the success of the Internet ever since the founding of the World Wide Web. During the latter part of the 1990s, Google recognized the significance of in-text content hyperlinks (HTML tags) and developed a successful algorithmic enhancement known as PageRank, which revolutionized the web page search industry. Web2.0 has additionally elevated the prominence of tags, or tagging, as one of the most significant concerns in Internet society. Tagging, as a phenomenon, is consistent with the Web 2.0 philosophy that users can not only generate content but also use a more robust, responsive, and adaptable method for navigating and searching new and existing media. However, due to the nascent methodologies in the field of natural language processing, the majority of tagging tasks remain manual. For instance, despite its recognition as a highly successful Web 2.0 service, the manual process of tagging or cross-linking related items within Wiki articles persists. A bibliography database containing articles from scientific journals is one example. The authors typically designate a KEYWORDS section in bibliographical data. This section allows for the incorporation of flourished hypertext links for keywords in the abstract or even the full text into the bibliographic information. The inclusion of a scientific terminology dictionary could potentially improve scholarly articles' legibility and intuitive navigation.

The implementation of automated keyword extraction is critical when organizing vast collections of text documents. Keywords or tags can significantly enhance information retrieval (IR) tasks by providing more efficient and high-quality search results. Additionally, content-based recommendation systems can implement them to offer personalized content suggestions to users. Despite the large number of tags in text mining applications, automatic tag extraction remains an extremely difficult problem to solve. A prospective resolution to this issue could involve the construction of machine learning models employing supervised methodologies. Nevertheless, supervised methods frequently necessitate humans pre-tagging an entire corpus of documents. Factors such as variations in the grammatical syntax and vocabulary employed by human taggers hinder this endeavor, despite the availability of sufficient resources to manually tag every document. Turney et al. (2000) identify the repetition of candidate phrases in their work. Repetition of the same keywords often leads to over-tagging, which also limits the identification of new tags within the document corpus (Pandya et al., 2020).

Furthermore, they asserted that automatic keyword extraction has been a captivating subject of investigation in the domains of text mining, information retrieval, and natural language processing (NLP) due to its substantial applications. Researchers evaluated a variety of methodologies to extract keywords from a vast corpus of documents. Utilizing document corpus-specific intrinsic statistics, such as word frequency and tf*idf, is the most straightforward method. Others have extracted keywords using graph-based methods. Many researchers have also evaluated the implementation of supervised and unsupervised methods for keyword or topic detection. Previous attempts have attempted to tackle the issue of automatic keyword extraction by employing straightforward statistical techniques like word frequency and tf*idf. By analyzing word-document statistics, these methods generate document-specific keywords with relative ease. The tf*idf method computes two crucial statistics: term frequency and inverse document frequency. Inverse document frequency pertains to the number of documents that contain a particular term, whereas count denotes the number of occurrences of a term in a document.

It is common practice for a document to contain a collection of words or phrases that provide a concise summary of its primary subject matter. A wide variety of documents, including news articles, academic papers, digital media, and other types of documents, make extensive use of keywords in order to facilitate people's ability to efficiently manage and retrieve the information they seek. The use of keywords has become an essential tool for users to search for content of interest within a vast amount of information as a result of the exponential growth of information in the age of the Internet. As a result, numerous companies, including Google and Baidu, have developed search engines that are based on keywords (Shen, 2021).

In-text Categories

The field of automatic text categorization has seen significant advancements over recent years, driven by the increasing volume of online information and the need for efficient information retrieval. According to Hult and Megyesi (2019), the primary objective of automatic text categorization is to classify a document into one of a predetermined set of categories. This process involves analyzing the content of the text and assigning it to the most appropriate category based on predefined criteria.

Numerous companies, including Google and Baidu, have developed search engines that rely heavily on keyword-based categorization systems. Shen (2021) notes that these search engines use sophisticated algorithms to index and retrieve information based on keyword searches, enabling users to find relevant documents quickly and efficiently. This keyword-based approach forms the backbone of many modern search engines, highlighting the importance of accurate text categorization in improving search results.

Tags can serve as informal metadata for objects such as web pages and multimedia data, as described by Choe et al. (2019). Tags provide an additional layer of categorization, allowing users to organize and retrieve information based on user-generated labels. This method of tagging enhances the searchability and organization of content, making it easier for users to find relevant information.

Silla and Freitas (2019) provide a precise definition of hierarchical classification and propose a unifying framework for classifying tasks across different domains. Hierarchical classification is particularly useful for complex categorization tasks where categories are organized in a tree-like structure, allowing for multi-level categorization. This framework helps in managing and categorizing large datasets more effectively by recognizing the inherent relationships between different categories.

With the rapid growth of online information, text classification (also known as text categorization or text labeling) has become a central task in Natural Language Processing (NLP). Attieh and Tekli (2023) emphasize the importance of text classification in NLP applications, ranging from spam detection and sentiment analysis to topic labeling and recommendation systems. The advancements in machine learning algorithms have significantly improved the accuracy and efficiency of text classification systems, enabling more precise categorization of large volumes of text data.

Moreover, the development of deep learning techniques has further enhanced the capabilities of text classification systems. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have been widely adopted for text categorization tasks due to their ability to capture contextual information and complex patterns within text (Kim, 2020). These models have demonstrated remarkable performance in various text classification challenges, setting new benchmarks in the field.

Text Classification

One common approach for organizing documents is to assign a specific label to each individual document. However, this method is only effective when users already know the exact labels they are searching for. This tactic does not adequately address the more general issue of locating documents related to a particular topic or subject area. In situations where users are seeking information on broader themes, a more efficient solution is to group documents based on shared, generic topics and then assign a descriptive name to each group. These groups, each identified with a meaningful label, are referred to as categories or classes (Hong, Y., & Ratner, K., 2020). This categorization not only simplifies the search process for users but also enhances the overall organization and accessibility of the information within the system. By focusing on common topics and labeling groups accordingly, the system can provide a more intuitive and user-friendly experience, ensuring that users can efficiently find the documents they need even if they are unfamiliar with the specific labels of individual documents.

Document classification is the systematic process of categorizing documents into predefined groups through the application of supervised learning techniques (Aggarwal, D., & Sharma, A., 2022). This categorization can be achieved either manually by subject matter experts or automatically using various classification algorithms. Alternative approaches such as Particle Swarm Optimization (PSO) have gained attention in recent years. Inspired by natural social behaviors observed in organisms like birds and fish (Qawqzeh Y. et al., 2021), PSO adjusts classification models to minimize a specified cost function, aiming to improve the accuracy and efficiency of document categorization (Huda, R., & Banka, H., 2019). This dynamic optimization process allows for the fine-tuning of classification parameters, ensuring precise assignment of documents to their respective categories based on content and context.

Classification of documents involves sorting them based on different attributes like authorship, document type, publication date, or subject matter (Surovtseva, N., 2022). Among the approaches used for document classification, two main methods stand out: the content-based approach and the request-based approach. In the content-based approach, the classification of a document is determined by the emphasis placed on its subjects within the document itself (Simplilearn, 2024) while in automatic classification, the number of times given words appears in a document determine the class (Xiaoxiao Li et al., 2020). In Request-oriented classification, the expected user requests influence the classification of documents (Fortra, 2023). Automatic document classification tasks can be categorized into three types (Zaizoune M. et al., 2023):

1. Unsupervised document classification, also known as document clustering, involves autonomously categorizing documents based solely on their inherent characteristics, without relying on external data or predefined categories. Algorithms analyze the content and structure of documents to group them into clusters, aiding in the discovery of natural patterns and relationships within unlabeled text data.
2. In semi-supervised document classification, portions of documents are annotated using external methods, while the remaining sections are left unlabeled for autonomous classification. This hybrid approach combines the benefits of external guidance with autonomous classification, improving accuracy, especially when labeled data is limited.
3. Supervised document classification utilizes external methods, such as human feedback or labeled training data, to guide the classification process and determine document categorization. Algorithms learn from labeled examples, making predictions on unseen documents based on patterns gleaned from the provided labels, commonly used when labeled data is abundant or domain expertise is crucial.

Particle Swarm Optimization (PSO) Algorithm

Stated by Liu et al., (2017), Kennedy and Eberhart (1995) introduced the Particle Swarm Optimization (PSO) algorithm, a stochastic optimization method that utilizes swarm dynamics. The PSO algorithm simulates the social behavior of insects, fish, herds, and birds. These swarms adhere to a cooperative strategy for locating food, with each member continuously modifying the search pattern in response to his own and other members' learning experiences. Millonas, through the lens of artificial life theory, proposed five fundamental principles (Van Den Bergh 2001) for the development of cooperative swarm artificial life systems, which computers can implement.

1. Proximity: The swarm should be able to execute basic space and time computations.

2. Quality: The swarm ought to possess the capability to perceive and react to variations in environmental quality.

3. Diverse response: The swarm should not limit its approach to resource acquisition to a specific scope.

4. Stability: the mode of behavior of the swarm should not fluctuate in response to environmental changes.

5. Adaptability: The swarm should adjust its behavior mode when necessary.

It is worth noting that the fourth and fifth principles represent contrasting aspects of the same coin. The fourth and fifth principles encompass the fundamental attributes of artificial life systems and have evolved into guiding principles for the development of swarm artificial life systems. Particles in PSO are capable of adjusting their velocities and positions in response to changes in their surroundings; thus, it satisfies the criteria for both quality and proximity. Furthermore, the PSO swarm operates without any movement restrictions and perpetually explores the possible solution space in search of the optimal solution. Particles in PSO are capable of maintaining a constant motion within the search space while adjusting their mode of motion to accommodate environmental changes. Thus, particle-swarm systems satisfy the five principles mentioned.

Considerable alterations to the initial particle swarm optimization (PSO) algorithm have been suggested ever since it was developed. The modifications frequently manifest as algorithmic components that contribute to enhanced performance. Added constants in the particles' velocity update rule and stand-alone algorithms that are utilized as components of hybrid PSO algorithms are examples of these algorithmic components (Birattari, 2018).

The use of evolutionary techniques in scientific fields has suddenly gained popularity. Their widespread usage is a result of their adaptability as design instruments and their exceptional capability to locate the optimal solution in complicated multimodal search spaces. Genetic algorithms and particle swarm optimization have surfaced as algorithms that are both effective and efficient in addressing intricate optimization challenges. Charles Darwin established the principles that form the foundation of the genetic algorithm, an approach to learning. Working with a PSO is comparable to how birds or animals behave in a flock or herd. A socio-cognitive study investigating the concept of collective intelligence in biological populations during the mid-1990s involved an attempt to replicate the graceful, well-orchestrated motion of a flock of birds. This endeavor led to the creation of the PSO technique. It was classified as an evolutionary technique shortly after its inception (Juneja & Nagar, 2016).

Study by Lin J. (2018), PSO is predicated on the notion that individuals should collaborate and share information in order to achieve the optimal solution. The fundamental algorithm of PSO is characterized by its ease of comprehension and implementation, as well as its rapidity and robustness. PSO has achieved excellent results in numerous fields such as power system optimization, neural network training, system identification, and PID parameter optimization. Similar to the genetic algorithm and other evolutionary algorithms for global optimization, premature convergence is a characteristic of the PSO algorithm, particularly for complex multimodal search problems. Researchers proposed numerous improved algorithms to prevent premature convergence. A novel hybrid PSO algorithm utilizes gradient information to achieve faster convergence without encountering local optima.

The PSO algorithm implements function stretching, a novel technique, to transform the objective function and mitigate local optima. One important thing that causes the PSO algorithm to converge too quickly is that the optimal particle found by the swarm often ends up in a local minimum. This paper presents a proposed enhanced PSO (IPSO) algorithm. During the iteration of the IPSO algorithm, we use the roulette wheel method to select a particle from among several particles with the highest fitness rather than simply selecting the one with the highest fitness as the optimal one. By preventing this, we reduce the likelihood of premature convergence to local optima and prevent all particles from approaching a potential local optimum very quickly.

Conceptual Model of the Study

Input	Process	Output
1. Academic Papers and Metadata **a. Collection of academic papers** **b. Metadata (author, title, publication date, etc.)** **2. Keywords for Tagging** **3. Particle Swarm Optimization (PSO) Algorithm Parameters** **a. Swarm size** **b. Number of iterations** **4. User Queries** **a. Search terms**	1. Data Collection - Gather all academic papers and their metadata 2. Document Preprocessing - Clean and normalize academic papers 3. Feature Extraction - Identify and extract key terms and phrases from academic papers 4. Algorithm Application - Apply PSO algorithm for automatic tagging: i. Initialization: Configure PSO parameters and initial tags ii. Optimization: Use PSO to refine and optimize tags based on document features 5. Indexing - Organize academic papers for efficient retrieval 6. Web Interface Development - Create a user-friendly web interface for search and retrieval 7. Search Functionality Implementation - Develop backend to process user queries and rank relevance of results	1. Repository of Academic Papers a. Centralized database of collected academic papers 2. Cleaned and Standardized Academic Papers 3. Extracted Features - Key terms and phrases from academic papers 4. Optimized Tags - Tags generated by the PSO algorithm for each academic paper 5. Indexed Academic Papers - Organized for efficient retrieval 6. User-Friendly Web Interface - Platform for users to search and retrieve academic papers 6. Search Results - List of academic papers relevant to user queries

Figure 1. Conceptual Model

Input

The development of a web-based information retrieval system for academic papers with automatic tagging using the Particle Swarm Optimization (PSO) algorithm at the Technological University of the Philippines – Manila involves several key inputs, processes, and outputs. The inputs include the academic papers themselves, along with their metadata such as author, title, and publication date. Keywords for tagging, PSO algorithm parameters, user queries, and user credentials for access control are also essential inputs.

Process

The process begins with the collection and preprocessing of academic papers to ensure they are clean and standardized. Feature extraction follows, where key terms and phrases are identified. The PSO algorithm is then applied to automatically generate optimized tags for each academic paper. Next, the academic papers are indexed to facilitate efficient retrieval based on user queries. The development of a web-based interface is crucial for providing a user-friendly platform where users can search for and retrieve academic papers. This involves implementing robust search functionality to process user queries and rank the relevance of search results. Additionally, user authentication and access control measures are implemented to ensure secure access to the system.

Output

The outputs of this process include a centralized repository of academic papers, cleaned and standardized documents, extracted features, and optimized tags. The indexed academic papers enable efficient search and retrieval, while the user-friendly web interface provides a seamless experience for users. Finally, secure access ensures that only authorized users can interact with the system, maintaining the integrity and confidentiality of the academic papers.

Architecture of the Study

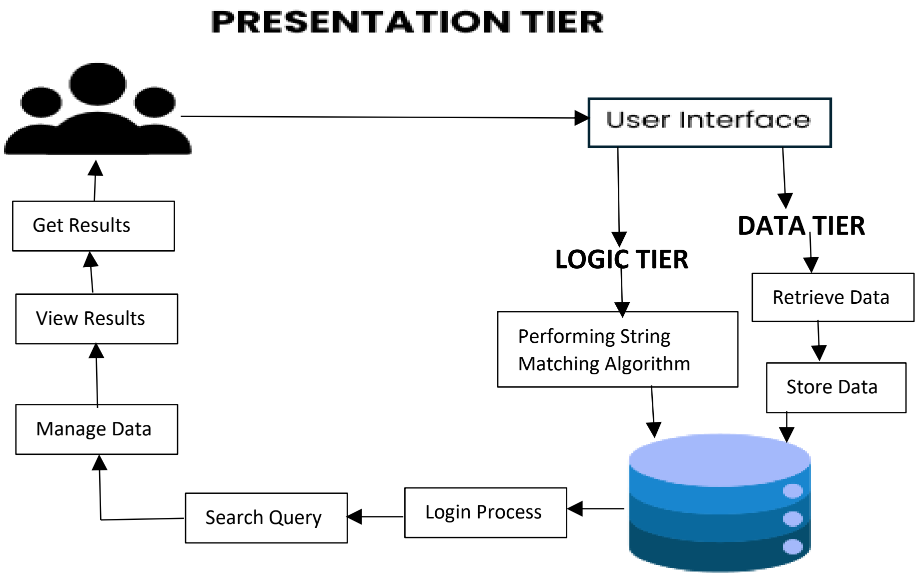

Fig. 2: Three-tier Architecture of the Information Retrieval System

Presentation Tier

The Presentation Tier constitutes the user interface layer, where students/user engage with the system through web pages or forms. It encompasses interfaces for the login process, searching academic papers, displaying search results, managing data (including adding, deleting, and approving/denying access requests), and accessing detailed paper information.

Logic Tier

The Logic Tier, positioned beneath the presentation layer, is dedicated to performing a string-matching algorithm as part of its responsibilities. It manages the system's logic processes, processes user inputs from the presentation tier, handles authentication and authorization of users, and coordinates communication between the presentation and data tiers.

Data Tier

The Data Tier serves as the foundational layer, responsible for retrieving and storing critical system data. This includes the Academic Papers Database, which stores information such as titles, authors, abstracts, and full documents of academic papers. Additionally, the Data Tier manages the Student Database, storing essential student information such as emails, TUP IDs, and passwords. This foundational layer plays a pivotal role in supporting the overall functionality of the system by efficiently handling the storage and retrieval of academic and student-related data.

Definition of Terms

Information Retrieval System - a system that extracts relevant documents from digital document repositories by considering the user's query and its relationships with other words and syntactic roles in the sentence (Mahapatra et al., 2020).

Database – A database is an electronic framework that stores, manages, and retrieves information, often governed by a database management system (Sridevi, R., & Srimathi, S., 2020).

PHP - a programming language that allows users to create dynamic web pages and applications (Power, D., 2021).

SQL - a structured language used to communicate with data stored in a database management system, using dynamic query commands for processing, and controlling data (Nethravathi, et al., 2020)

CHAPTER III

Project Design

The primary objective of this project is to develop a web-based Information Retrieval System (IRS) for the Technological University of the Philippines, Manila Campus. The system aims to facilitate the sorting and viewing of research papers, enabling students and professors to easily access relevant materials for academic purposes. Additionally, the system aims to automatically categorize scanned research papers according to their respective courses, streamlining the retrieval process for users.

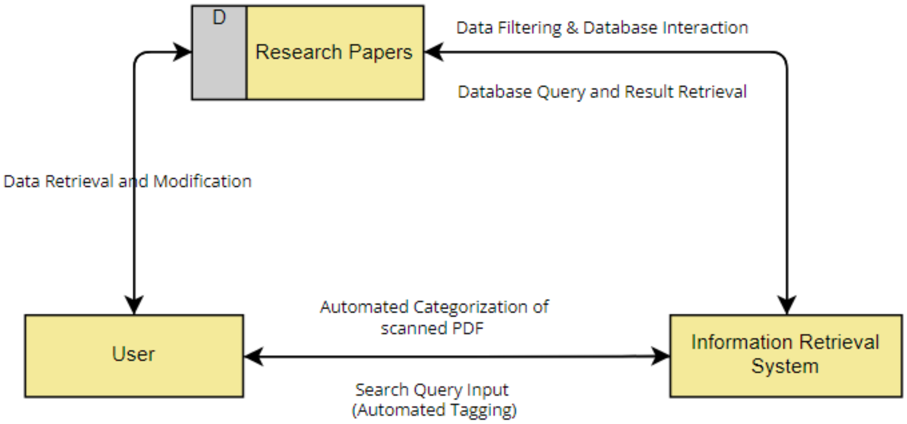
D
Research Papers
Data Filtering & Database Interaction
Database Query and Result Retrieval
Data Retrieval and Modification
Automated Categorization of scanned PDF
User
Information Retrieval System
Search Query Input (Automated Tagging)

Fig 3. Context Diagram

The context diagram illustrates the high-level interaction between users and the IRS (Information Retrieval System). Users input search queries into the IRS, which then filters data from the research paper database and returns sorted results to the users. Users can also upload research papers directly into the system. Once uploaded, the system automatically scans and categorizes the papers based on the relevant course or topic. This automation ensures the database is always up-to-date and accurately organized.

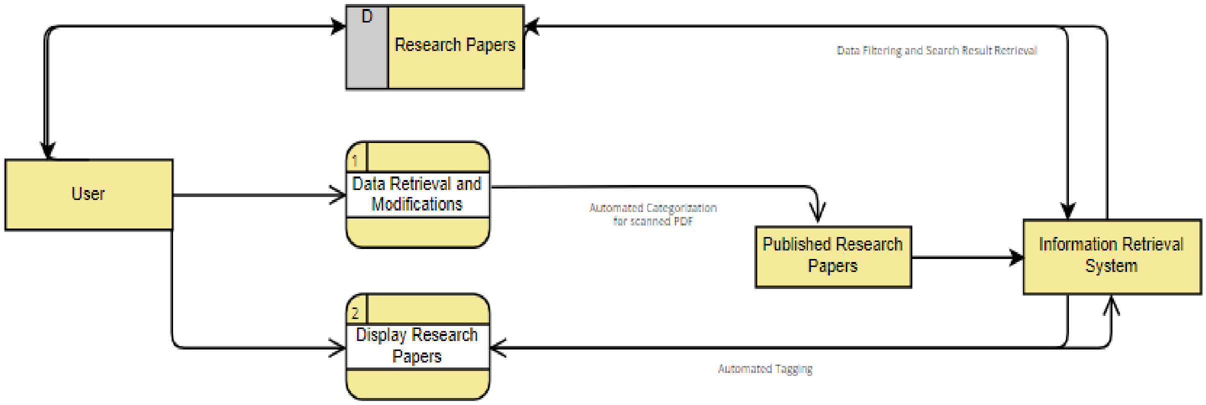
D
Research Papers
Data Filtering and Search Result Retrieval
User
1
Data Retrieval and Modifications
Automated Categorization for scanned PDF
Published Research Papers
Information Retrieval System
2
Display Research Papers
Automated Tagging

Fig 4. Data Flow Diagram (DFD) Level 0

At the top level, users initiate the process by entering search queries into the IRS. The system filters data based on these queries, retrieves relevant papers from the research paper database, and returns sorted results to the users. Additionally, users can upload research papers directly into the system. Once uploaded, the system automatically scans and categorizes the papers based on the relevant course or topic. Authors of the research papers have the authority to retrieve and modify their own papers.

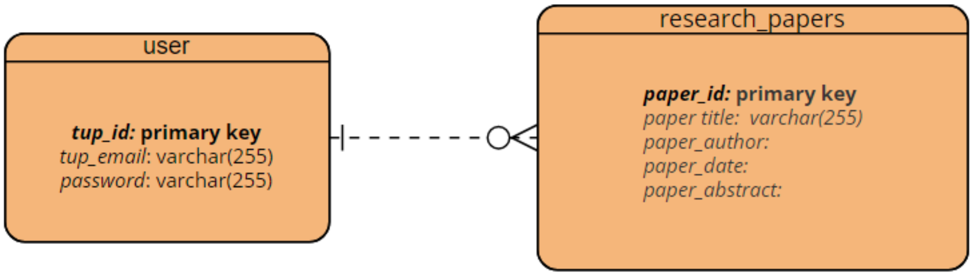

Fig 5. Entity Relationship Diagram (ERD)

The relationships between essential entities in the system are represented by the ERD. Each user has the ability to add, delete, amend, and view multiple research papers, as evidenced by the one-to-many relationship between users and those documents.

The objective of this design is to develop a user-friendly and robust Information Retrieval System that is specifically designed to meet the requirements of the Technological University of the Philippines – Manila Campus. This system will ensure security and accountability while facilitating the expedient access to research papers.

Project Development

Phase 1: Project Initiation

The objectives and scope of the Information Retrieval System (IRS) at the Technological University of the Philippines – Manila Campus will be established during this phase of the project. In order to identify extant gaps in the field, researchers will conduct a literature review and conduct a stakeholder analysis. Additionally, explicit success criteria will be implemented.

Phase 2: System Design and Architecture

A high-level system architecture will be established, and the technology stack will be selected by researchers, who will also develop detailed requirements. Through stakeholder feedback, a fundamental prototype will be refined and developed.

Phase 3: System Development

The primary objectives of this phase are to establish databases, develop fundamental functionality, implement security measures, and implement the user interface. Powerful systems are guaranteed through comprehensive testing, which encompasses both unit and integration testing.

Phase 4: Testing and Deployment

User acceptance testing, deployment of the IRS to production, and testing of the integrated system comprise the final phase. A seamless transition to operational use will be facilitated by comprehensive documentation and training sessions.

Testing and Operating Procedure

Table 1. User Interface

The user interface is designed to facilitate efficient information retrieval for research papers at the Technological University of the Philippines – Manila Campus. The following procedures outline the testing and operation for the user interface:

System Function	Procedure	Expected Output
1. Log in as a student	1. Input username/tup-id 2. Input password 3. Click log in button	1. Incorrect credentials should generate an error message "Invalid User or Password" and prompt the user to input the correct details. 2. Upon successful login the user must be redirected to the home screen.
	1. Input Data/words 2. Click Search button	1. Upon successful search results, the list of Research Papers must be display, according to selected filters.
2. Search Research Papers	3. Click Filter button (Title, Author, Abstract) 4. Select any research paper	2. Upon successful view results, the Research Paper Title must be clearly seen with the abstract below, other information like authors, course, year and department should be visible.

3. Add Research Paper	1. Upload and scan research paper.	1. Automatically categorize according to course.
4. Delete Research Paper	2. Input Research Paper Title 3. Select Research Paper 4. Click Delete button	2. Before proceeding to delete action, generate confirmation message "Are you sure you want to delete this Research Paper? Title: &Title, this action cannot be undone. [Cancel] [Confirm]" 1. Upon successful delete, generate display message "Research Paper Title: &title deleted successfully", reload the page.

4. Update Research Paper	1. Select row (Update Title, Year, Author, Abstract) 2. Click Update button	1. Upon successful search results, the list of Research Papers must be display, with default sorting of (Title - Asc.) 2. Upon successful update, generate display message “Updated successfully “and reload the page.

End - Users:

Student Community Engagement: In order to improve the end-user experience, researchers suggest a multifaceted approach. Initially, students will participate in appointed usability testing sessions and provide detailed feedback on the intuitiveness of the interface, the relevance of search results, and their overall satisfaction. A user-friendly feedback mechanism will be implemented within the system interface to encourage continuous improvement. This mechanism will enable users to report issues or propose improvements with ease. Furthermore, in order to facilitate the exchange of distinct use cases and challenges that students may encounter during the research paper discovery process, researchers will initiate focus groups or online forums.

Experts:

Algorithmic Optimization: In the context of auto-categorization of research papers, collaborate with information retrieval experts to conduct an exhaustive evaluation of the system's search algorithms. The objective is to enhance the speed of query processing and the relevance of search results.

Industry Best Practices Integration: Maintain a current understanding of industry best practices by collaborating with experts in educational technology and information retrieval systems. Incorporate feedback to ensure that the IRS is in alignment with current and emergent trends in academic information retrieval, and implement identified best practices to improve the system's overall performance and user satisfaction.

Cutting-Edge Feature Development: Establish a collaborative development team that includes system developers and experts. Generate and prioritize features that capitalize on emerging technologies. Before the full implementation, conduct beta testing of cutting-edge features with a restricted group of users to collect feedback. Ensure that the existing system is seamlessly integrated by iterating on features based on expert insights and user responses. Scalability, Security, and Technological Integration: Collaborate closely with security and scalability experts to guarantee that the IRS can manage escalating data volumes while simultaneously preserving system performance. Conduct security audits on a regular basis and implement recommendations to protect user data. Investigate the potential for the integration of emerging technologies, such as blockchain for secure document validation or cloud-based solutions for scalability. Establish a roadmap for future technological advancements in collaboration with experts to guarantee that the IRS remains at the vanguard of academic information retrieval systems.

REFERENCES

Agbele, K., Ayetiran, E. and Babalola, O. (2018). A Context-Adaptive Ranking Model for Effective Information Retrieval System. International Journal of Information Science. 8(1), 1-12

Aggarwal, D., & Sharma, A. (2022). Deep Learning Approaches for Document Classification: An Insight. 2022 2nd International Conference on Intelligent Technologies (CONIT), 1-5. https://doi.org/10.1109/CONIT55038.2022.9848163.

Attieh, J., & Tekli, J. (2023). Supervised term-category feature weighting for improved text classification. Knowledge-based Systems, 261, 110215. https://doi.org/10.1016/j.knosys.2022.110215

Choe, H.-S., Kim, J., Jin, D.-S., & Kim, K. (2019, September 22). Automatic In-Text Keyword Tagging based on Information Retrieval. Research Gate. https://www.researchgate.net/profile/Ho-Seop-Choe/publication/220635657_Automatic_In-Text_Keyword_Tagging_based_on_Information_Retrieval/links/0f317538c213d8bc54000000/Automatic-In-Text-Keyword-Tagging-based-on-Information-Retrieval.pdf

Farek L and Benaidja A. (2023). Feature redundancy removal for text classification using correlated feature subsets. Computational Intelligence. 10.1111/coin.12621. 40:1 https://onlinelibrary.wiley.com/doi/10.1111/coin.12621

Hong, Y., & Ratner, K. (2020). Minimal but not meaningless: Seemingly arbitrary category labels can imply more than group membership.. Journal of personality and social psychology. https://doi.org/10.1037/pspa0000255.

Huda, R., & Banka, H. (2019). New efficient initialization and updating mechanisms in PSO for feature selection and classification. Neural Computing and Applications, 32, 3283-3294. https://doi.org/10.1007/s00521-019-04395-3.

Hulth, A., & Megyesi, B. (2019). A Study on Automatically Extracted Keywords in Text Categorization (pp. 537–544). Association for Computational Linguistics.

Juneja, M., & Nagar, S. K. (2016). Particle swarm optimization algorithm and its parameters: A review. 2016 International Conference on Control, Computing, Communication and Materials (ICCCCM). doi:10.1109/icccm.2016.7918233

Keswani A, Jain T and Sharma B. (2023). Multi-Class Text Classification using Machine Learning & Deep Learning 2023 2nd International Conference on Futuristic Technologies (INCOFT). 10.1109/INCOFT60753.2023.10425423. 979-8-3503-0884-6. (1-6). https://ieeexplore.ieee.org/document/10425423/

Kim, Y. (2020, August 25). *Convolutional neural networks for sentence classification*. [1408.5882] Convolutional Neural Networks for Sentence Classification (arxiv.org)

Khalilian, M., & Hassanzadeh, S. (2019). Document classification methods. ArXiv, abs/1909.07368. How to Correctly classify your data in 2022. (n.d.). Tripwire. https://www.tripwire.com/state-of-security/how-to-correctly-classify-your-data

Krishnapriya, M. S., Thushara, M. G., & Nair, S. S. (2017). A model for auto-tagging of research papers based on keyphrase extraction methods. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). doi:10.1109/icacci.2017.8126087 https://aclanthology.org/P06-1068.pdf

Kumar, S., Garg, A., & Haider, M. (2023). PSO based Web Documents Prioritization for Adaptive Websites using multi-Criteria. 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 1-6. https://doi.org/10.1109/ISCON57294.2023.10112161.

Li, X., Al-Zaidy, R., Zhang, A., Baral, S., Bao, L., Giles, C., & , L. (2020). Automating Document Classification with Distant Supervision to Increase the Efficiency of Systematic Reviews. ArXiv, abs/2012.07565.

Li, Y., Gui, W., Yang, C., & Li, J. (2018). Improved PSO algorithm and its application. Journal of Central South University of Technology, 12(1), 222–226. doi:10.1007/s11771-005-0403-4

Lima, G., & Campos, M. (2022). Information Storage and Retrieval System: an analysis of the impact of variables and measures aimed at the organization and retrieval of information centered on the user. RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação. https://doi.org/10.20396/rdbci.v20i00.8667925/28663.

Manasia, S., Rai, S., Kalwar, M., Yadav, V., & Shinde, V. (2023). Information Retrieval System for College Search. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/cseit2390224.

Mohanrasu S, Janani K and Rakkiyappan R. (2024). A COPRAS-based Approach to Multi-Label Feature Selection for Text Classification. Mathematics and Computers in Simulation. 10.1016/j.matcom.2023.07.022. 222. (3-23). https://linkinghub.elsevier.com/retrieve/pii/S0378475423003129

Montes de Oca, M. A., Stutzle, T., Birattari, M., & Dorigo, M. (2018). Frankenstein's PSO: A Composite Particle Swarm Optimization Algorithm. IEEE Transactions on Evolutionary Computation, 13(5), 1120–1132. doi:10.1109/tevc.2018.2021465

Navitas, P., & Febrianti, Y. (2022). Analysing engineering students' information retrieval behaviour using online databases as information sources. Jurnal Kajian Informasi & Perpustakaan. https://doi.org/10.24198/jkip.v10i2.35264.

Ni, P., Li, Y., & Chang, V. (2020). Research on Text Classification Based on Automatically Extracted Keywords. Int. J. Enterp. Inf. Syst., 16, 1-16. https://doi.org/10.4018/ijeis.2020100101.

Qawqzeh, Y., Alharbi, M., Jaradat, A., & Sattar, K. (2021). A review of swarm intelligence algorithms deployment for scheduling and optimization in cloud computing environments. PeerJ Computer Science, 7. https://doi.org/10.7717/peerj-cs.696.

Pandya, M. R., Reyes, J., & Vanderheyden, B. (2020). Method for Customizable Automated Tagging: Addressing the problem of over-tagging and under-tagging text documents. 2020 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata50022.2020.9378048

Pittaras N, Giannakopoulos G, Stamatopoulos P and Karkaletsis V. (2023). Content-based and Knowledge-enriched Representations for Classification Across Modalities: A Survey. ACM Computing Surveys. 55:14s. (1-40). https://doi.org/10.1145/3583682

Polatgil M and Kekül H. (2023). The Effect of Document Length on Machine Learning Success in Text-Based Data 2023 Innovations in Intelligent Systems and Applications Conference (ASYU). 10.1109/ASYU58738.2023.10296594. 979-8-3503-0659-0. (1-6). https://ieeexplore.ieee.org/document/10296594/

S, D. (2022). DIGITALISATION OF LIBRARIES (DIGITAL LIBRARIES) AND ITS SCOPE. EPRA International Journal of Research & Development (IJRD). https://doi.org/10.36713/epra10724.

Silla, C.N.; Freitas, A.A. A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. 2019, 22, 31–72

Shen, X. (2021). Automatic Keyword Tagging With Machine Learning Approach. https://spectrum.library.concordia.ca/id/eprint/990187/1/Shen_MASc_S2022.pdf

Surovtseva, O. (2018, October). Text classification based on machine learning methods. In 2018 International Conference on Advanced Computer Information Technologies (ACIT) (pp. 240-244). IEEE. Text classification based on machine learning | IEEE Conference Publication | IEEE Xplore

Surovtseva, N. (2022). Classification of documents as a theoretical problem in office work and in the archive. Herald of an archivist. https://doi.org/10.28995/2073-0101-2022-3-756-771.

Simplilearn. (2024, May 16). Data Classification: Overview, types, and examples. Simplilearn.com. https://www.simplilearn.com/data-classification-overview-types-examples-article

Wable, R. (2021). Information Retrieval in Business. https://doi.org/10.36227/TECHRXIV.14709456.V1.

Wang, D., Tan, D., & Liu, L. (2017). Particle swarm optimization algorithm: an overview. Soft Computing, 22(2), 387–408. doi:10.1007/s00500-016-2474-6

Wu, F., Ji, Y., & Shi, W. (2022). Design of a Computer-Based Legal Information Retrieval System. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/6942773.

Yu, B. (2019). Research on information retrieval model based on ontology. EURASIP Journal on Wireless Communications and Networking, 2019, 1-8. https://doi.org/10.1186/s13638-019-1354-z.

Zaizoune, M., Fakhri, Y., & Boulaknadel, S. (2023). Automatic emails classification. 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), 1-4. https://doi.org/10.1109/WINCOM59760.2023.10322962.

Zangari A, Marcuzzo M, Rizzo M, Giudice L, Albarelli A and Gasparetto A. (2024). Hierarchical Text Classification and Its Foundations: A Review of Current Research. Electronics. 10.3390/electronics13071199. 13:7. (1199). https://www.mdpi.com/2079-9292/13/7/1199