I.J. Intelligent Systems and Applications, **2024, 2, 13-23**
Published Online on April 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2024.02.02

A New Approach to Improving Search Efficiency in Digital Libraries

Irada Alakbarova
Ministry of Science and Education of Azerbaijan, Institute of Information Technologies, Baku, Azerbaijan
E-mail: airada.09@gmail.com
ORCID iD: https://orcid.org/0000-0002-9876-3035
*Corresponding Author

Dilbar Alizada
Ministry of Science and Education of Azerbaijan, Institute of Information Technologies, Baku, Azerbaijan
E-mail: dilberilhamgizi@gmail.com
ORCID iD: https://orcid.org/0000-0001-9331-8066

Received: 24 November 2023; Revised: 29 December 2023; Accepted: 17 January 2024; Published: 08 April 2024

Abstract: The development of Internet technologies influences the activities of libraries and changes their nature. The volume of content collected in digital libraries is growing rapidly. This requires the use of new technologies to search and obtain electronic materials (text, video, images, sound files) stored in the e-library. Today, using the new capabilities of network technologies and intelligent systems, the proper organization of the digital library, and increasing the efficiency of library services are the main factors leading to an increase in the number of readers and their satisfaction. The main objectives of digital libraries are to ensure efficient retrieval of electronic resources and collaboration between users. While researching various scientific articles on library and information sciences (LIS), we did not encounter approaches using cluster analysis in combination with wiki technologies. To collaborate users in digital libraries and their involvement in organizing electronic resources, we propose using an open database managed by wiki technologies. To effectively search for electronic resources in these open databases, it is proposed to use the clear clustering method. The clear clustering method also allows you to control the quality of clustering. The proposed method is important when creating intelligent (smart) libraries that are easy to manage and automate certain tasks. The research aims to create not just a smart library, but a smart library based on wiki technologies.

Index Terms: Digital Library, Wiki Technologies, Intelligent Systems, Clustering, Smart Library.

1. Introduction

Digital libraries are various types of electronic publications, electronic archives, electronic databases, electronic textbooks teaching aids, etc. Is an organization with specialized personnel that selects, structures, offers intellectual access, interprets, disseminates, preserves integrity, and ensures continuity over time of collections of electronic works. The term digital library or digital library is used in different subject areas [1].

User interfaces for digital libraries include interface technologies used in computer systems. A key challenge for digital libraries is the availability of software systems that manage information storage and retrieval. Digital library systems are built by freely integrating applicable and existing tools or by extending existing systems that support library catalogs and library automation [2].

Digital libraries have many advantages in modern information infrastructure. Standards, advanced technologies, and reliable systems can serve users with different areas of interest, providing a wide range of specialized services. The digital library focuses on infrastructure, personnel, facilities, etc., rather than a concept based on service and delivery to the consumer. This concept tends to structure the library by function type, separating functions that must be partially or fully supported by library staff and that can be implemented using existing hardware and software systems.

Digital libraries have allowed readers to share resources and obtain the necessary information from different sources of knowledge, regardless of location and time. Advances in information technology have made the collection, storage, management, and security of knowledge much more efficient [3].

Digital libraries, like regular libraries, have many bibliographic categories and a large collection of electronic books, magazines, articles, and newspapers. Despite the development of information technology in our time, most digital library work with these collections and categories using traditional methods. Although most libraries have an automated information system, information security, identification, and authentication systems are organized effectively, in many cases, network speed and search efficiency are not enough, and the optimal arrangement of books and other materials (magazines, newspapers, articles, etc.) requires a large amount of labor of library workers [4].

The problems faced by digital libraries are also related to the rapid growth of Internet technologies. Today, Internet users want an open, smart, and flexible library experience. For this reason, digital libraries must create new, modern types of services. In other words, digital libraries must be smart, accessible, and dynamic.

Digital libraries must be sustainable, use modern technology, and be updated regularly to serve users better. Digital library users need accurate and timely information, which can only be provided by trained specialists in this field. It is necessary to use technologies that, on the one hand, will facilitate the work of specialists, and on the other hand, would ensure user satisfaction. Also, modern digital libraries should be accessible and open.

Today industry 4.0 affects also including digital libraries [5]. The main requirement is that the digital library should be free, with a searchable collection of hundreds of millions of books available to everyone via the Internet. The issue of security of user's data plays an important role here. In most developed countries, innovations in digital libraries are based on big data, cloud computing, intelligent systems, etc. These possibilities led to the creation of a new digital library model. This artificial intelligence-based model is called a "smart library."

Today, artificial intelligence is used in many areas: medicine, business, education, games, libraries, etc. The idea of creating artificial intelligence systems in libraries appeared back in 1990 [6]. Libraries driven by artificial intelligence are called smart libraries or intelligent libraries. The smart library has overcome the limitations of traditional libraries thanks to advanced technologies such as artificial intelligence, cloud technology, Internet of Things [7]. Artificial intelligence is a key factor in creating a new generation of digital libraries [8].

A smart library mainly means efficiency in finding information, organizing, and making work easier through the use of modern information technology and process automation. Smart libraries are created on the basis of intelligent programs that replace library staff and serve readers. Applications of artificial intelligence in library systems include descriptive cataloging, subject indexing, reference services, technical services, collection development, information security, etc.

To enable users of digital libraries to participate in the organization of electronic resources, to create collective knowledge, and to collaborate, we propose to use an open database managed using wiki technologies.

The emergence of wiki technologies in the virtual space can be considered an information revolution in the field of information technology. The word "Wiki" means "quick" in Hawaiian [9]. With the help of wiki technologies, websites are created whose content can be changed by any Internet user. An example of these projects is Wikipedia, Wiki Knowledge, Wikibooks and other popular sites that today have millions of users [10,11].

Today, the concept of wiki has given impetus to the development of new views covering various spheres of society, such as wiki-economy, wiki-democracy, wiki-parliament, wiki-medicine, etc. [12,13]

The main goal of wiki technology is to help every Internet user create a web page on any topic, change pages and upload files of different formats to pages at any time, without leaving the browser. Wiki technologies allow users to create content on the Internet as full-fledged web specialists, regardless of their specialization and interests [11,14]. At the same time, wiki technologies are a convenient tool for building social relationships. They allow people to participate in online discussions and express their opinions on content (web pages, articles, comments, etc.) created or modified by other users, resulting in new social relationships in society.

Wiki technologies allow not only the collaborative creation and editing of information, but also the rapid search and use of resources, the use of convenient navigation, and the creation of appropriate metadata for navigation and management.

The use of wiki technology in digital libraries promotes the active participation of book lovers in the formation of a new information space. During this process, a complete reorganization of the relationship between the library and users occurs, and fundamentally new forms of information services appear [11]. The concept of an electronic wiki library is based on the following ideas:

- Creating an open environment – certain services and technological developments reach their highest level and allow the development of an intelligent environment.
- Mobile access – availability of all types of electronic services anywhere in the world. These services must be individually tailored to each user.
- Creation of new knowledge: based on collective creativity; with the involvement of expert groups; using social networks.
- Active content – it is not enough to simply put content into the repository for activity. All objects must be connected to each other. In turn, the quality of the repository can be controlled using ready-made applications. For example, API (Application Programming Interfaces). Such interfaces follow standards (typically HTTP and REST) that are easily accessible and understandable. APIs play a key role in data collection and analysis [15]. Like any other software product, a modern API has its own software development lifecycle, including design, testing, creation, and versioning.
- Adaptability – creating an individual set of services to suit user requests. A large number of sources and a maximum variety of media (audio, video, graphics) allow you to quickly and easily adapt to the level and

needs of the user [16].

The article explores the possibilities and problems of digital libraries. The basic principles of development and capabilities of wiki technology are determined. The prospects for introducing wiki technologies and artificial intelligence into digital libraries have been identified. A general scheme of a smart library has been developed. To increase the efficiency of searching in digital libraries, we propose to use the cluster analysis method.

2. Related Works

The multimedia nature of digital libraries has interested librarians and information researchers since the end of the last century. Improving digital libraries requires a transition from simple search for information using keywords to more advanced technologies that allow the use of intelligent algorithms for text analysis, identification and processing of images, and speech recognition. To solve the problem of big data, information organization and access in digital libraries, an approach to developing information agents was proposed. Agents providing various services such as information retrieval, data extraction and data filtering ensure the efficient operation of digital libraries.

A study of library strategies in the United States revealed that library management focuses on collections, collaboration, and learning. Researchers have sought to understand how intelligent algorithms and related technologies are used in academic libraries. The research was conducted in two very different contexts: the UK and China [17].

The article relates to the detection of outliers in text documents. Due to unstructured data and big data, solving this problem is very difficult. and the authors propose a hybrid density-based approach using a local outlier algorithm. By using normalized mutual information in combination with a density-based algorithm, good results were obtained, especially in high-dimensional data sets. Because of big data, libraries are forced to use intelligent algorithms in information retrieval and other services [18]. Because artificial intelligence and related technologies can enable the transition to smart libraries.

In [19], the authors attempted to identify the factors influencing the overall performance of digital libraries, user satisfaction, etc. The authors developed a new approach using Delone and McLean's information system success model. Data were collected through a survey of Indonesian Open University (IOU) students. The results show that service quality plays the most important role in user loyalty to electronic services. The purpose of the study is to improve digital library services.

In [20], a synthesis of empirical studies on the application of artificial intelligence and machine learning in digital libraries was carried out. The data for the experiment was collected from the LISA, LISTA, Web of Science, and Scopus databases. The authors argue that the activities of digital libraries (content analysis, indexing, cataloging, classification, text recognition, information security, user identification, etc.) are supported by artificial intelligence and machine learning technologies. Machine learning methods such as deep learning and neural network algorithms have also proven to be powerful tools for searching and analyzing information. Clustering and classification allow automatic extraction of titles from documents, dividing documents by topic and by directory, which is useful when searching for documents [20, 21].

In [22] examined global trends in digital library research. The chronological growth of literature and the format of documents are analyzed. For the analysis, data extracted from the Scopus database was used. These data were taken as performance indicators. These include the number of publications, citations, H-index, and overall link strength. The study used a clustering method to analyze the co-authorship of organizations and countries. To assess the productivity of journals, authors, organizations, and countries, Lotkas's law was applied.

Research has shown that cluster analysis methods are widely used in document processing, mainly in the management and "reconnaissance" of digital libraries [23]. The number of studies using the cluster analytical approach in library and information science (LIS) was about 5 studies per year in the early 2000s, and increased to an average of 35 studies per year in the mid-2010s. The journal Scientometrics has the largest number of articles published on LIS research. The cluster analytical approach was shown to be used in 102 studies. Researchers argue that cluster analysis can make LIS research more accessible by providing an original, interesting, and educational knowledge discovery process [23].

In digital libraries, to achieve high quality services, especially in the areas of collaboration, content management and efficient information retrieval, it is advisable to use the capabilities of wiki technologies. Research has shown that wikis are a good tool for content classification and collaboration. For example, at [24] discusses the prospects for using Wiki technologies in digital libraries. The role of wikis in digital libraries is explored, especially in collaboration and information retrieval. Researchers propose using Wiki as an interactive tool for searching, creating an open database, and managing content. Wiki is considered the most successful tool for collaboration in digital libraries.

In [25], the author analyzed 30 digital libraries that used wikis. Libraries were classified using a classification scheme that has four categories: collaboration between intermediaries; interaction between library staff and visitors; collaboration between libraries; and collaboration among library staff. The author argues that Wiki can be used for various purposes: collecting digital resources, sharing information, creating digital repositories, supporting related work, web content management, the creation of intranets (documents on the same topic but in different languages), helping support, the creation of a knowledge base, reference books, and reader data. It was found that digital libraries mainly use MediaWiki software.

In [26] researchers introduce wikis as a new technology related to Web 2.0 that allows a group of people to create a collaborative digital library. Wikis facilitate communication between groups of people and provide annotated guides to collections and resources. In the case of a digital library, a wiki can be used for collaboration between library staff and users.

Research shows that the combined use of wiki technologies and machine learning methods to improve the quality of digital libraries is not proposed in any work. We propose that digital libraries use both wikis and machine learning techniques. In this case, electronic databases managed using wiki technologies will be open, and each user will be able to freely place electronic documents (books, magazines, newspapers, etc. materials) in them directly from the browser. Thanks to clustering, these documents are automatically divided into catalogs. This approach, that is, the combined use of a wiki and the clustering method, will facilitate searches, as well as ensure user satisfaction and simplify the work of library staff.

3. General Scheme of a Smart Library Using Wiki Technologies

The intelligence of information systems or digital libraries reflects their ability to perform a specific task in the face of variability and adjust their actions as the situation changes. Examples of artificial intelligence in a system are speech recognition, natural language processing, machine learning, deep learning, and robotics. The advantage of artificial intelligence is that the system can effectively recognize patterns at a scale and speed that are inaccessible to humans. The main function of smart libraries is the systematic formation of collections, storage, and systematization of information and knowledge in digital form, and providing easy access to it via the Internet [27]. In smart libraries, all library functions are automated, and the central library will digitize all information.

A smart library is an information center connected to other libraries and city services in a larger information ecosystem. The main goal of smart libraries is to create a more intelligent, convenient library environment that satisfies the readership. A smart library is an innovative system that combines modern information technologies to meet functional requirements and provide users with intelligent knowledge services through various smart devices.

The smart library contains various aspects such as smart services, smart economy, smart mobility, smart management, smart information security, etc. [28].

The capabilities of smart libraries improve the quality of traditional library functions and satisfy more and more information needs of users [27,28].

The user can access any information through e-books, newspapers, and search tools. Not all libraries need to become smart. A library can be modern by improving the quality of its services, developing new services, and introducing new information technologies and intelligent systems. In smart libraries, intelligent systems solve complex problems, such as selecting books according to the user’s age and interests, and returning, storing, and ordering goods without the participation of employee libraries. Technologies associated with smart libraries include machine learning, tablet-based kiosks, mobile applications, and the Internet of Things [27].

However, these innovative tools and services are only smart if they are user-friendly and user focused. Effective classification of electronic resources that today create big data in digital libraries requires the introduction of wiki technology and intelligent algorithms into the system. As can be seen from the scheme (fig. 1), a smart wiki library must have an open database. This wiki-based database creates new opportunities for users to incorporate electronic resources stored in user personal archives into the digital libraries database.

The proposed smart library using wiki technologies consists of three sections:

- Smart Library organization;
- Aspects of a smart library;
- Digital collection.

Digital libraries already have extensive experience in collecting information to create centralized repositories of both metadata and content of various types. Such systems are well suited to take advantage of the capabilities of search engine technology. The data warehouse cleans and aggregates unstructured data entered into the system (into the digital library database), which facilitates the work of analysts and increases the efficiency of the search engine. It also contains metadata, which stores information about the data. When working with traditional data warehouses, aggregated data is used to solve certain problems, such as setting up inconsistencies between data, etc. [29].

The digital library software is based on two technologies: a barcode system and a search engine. The barcode system is based on a unique code assigned to each resource in the library, which is then scanned using a barcode scanner.

The presence of online search in the wiki library allows users to find the information they need quickly and reliably, helping them in their research activities, and enhancing their outlook and professionalism. It should be noted that the pace of implementation of wiki technologies and intelligent algorithms in libraries may vary. This is due to a number of reasons: inconsistency with the needs of the library, lack of the necessary program and technical support, lack of qualified personnel, etc.

A smart search engine is essential to manage the data workflow. The main challenge in managing an digital library is searching through a wide range of information types, such as images, videos, or audio files. With such diversity of information, there is a clear need to more dynamically incorporate new search methods across different types of information. Dividing a digital collection into categories as shown in Figure 1, allows for very fast and relevant searches of large volumes of information found in disparate databases. The search service provides access to the entire range of digital library information, including books, periodicals, still images, audio, and video. It serves as the basis for statistical purposes as well as management information.

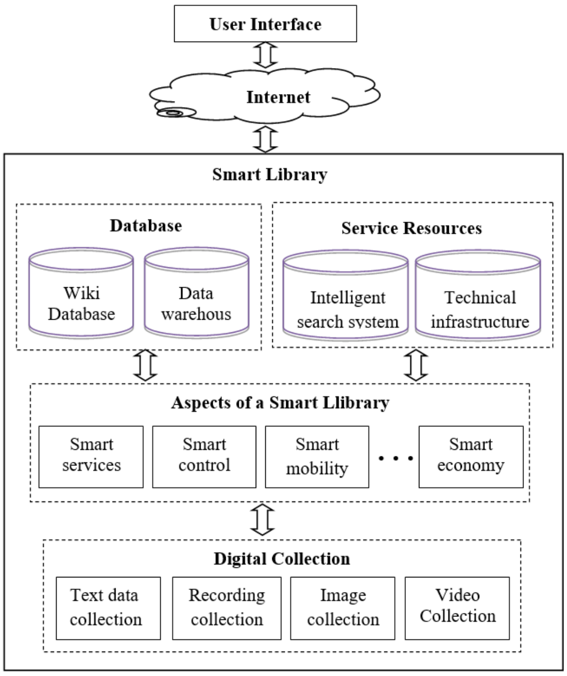


Fig.1. General scheme of a smart library using wiki technologies

To effectively search electronic documents, you can use the clustering method. By combining similar objects, we can quickly search document collections, easily understand different topics, and run queries across multiple applications. There are five main stages of work after selecting a resource: data loading, preprocessing, indexing, access, and navigation [30]. Technical infrastructure can speed up the process of creating a high-quality virtual resource.

4. Methodology

The proposed smart library using wiki technologies consists of three sections (Fig. 1):

- Smart Library organization;
- Aspects of a smart library;
- Digital collection.

The proposed approach for solving problems with big data in digital libraries and automatic distribution of documents without operator participation into catalogs consists of the following stages:

- Uplade data into the system (electronic library) using various technologies, mainly using wiki technology.
- The process of loading data into the data warehouse;
- Clustering of text documents;
- Data presentation.

Using clustering, we can process and analyze large amounts of data and thus find valuable information [31, 32]. To cluster text documents, the following steps are required:

- Pre-treatment. At this stage, the words included in the set of documents are subjected to morphological analysis, and commonly used working words are excluded from the text. 2. Description of documents. At this stage, each document (in this case, each document of the set DWL) is represented as a vector. For this, a common vector model (Vector Space Model) is used. Using this model, each document is represented as a vector in *m*-dimensional Euclidean distance.

- Selecting a proximity metric. There are different metrics for calculating the proximity between two vectors. For example: cosine metric, Chebyshev metric, Levenshtein metric, etc. Here we will use Euclidean distance.
- Determination of the clustering method. This stage is considered the most difficult stage of clustering, and the choice depends on the required criteria.

5. Clustering Electronic Documents

The clustering of electronic documents and web pages is attracting more and more attention as a key basic method for the uncontrolled organization of document flow, automatic pattern recognition, and operational search [32, 33]. Clustering can be defined as dividing a given set of objects into clusters in such a way that objects belonging to the same cluster are as similar as possible to each other, and objects in different clusters are different from each other [34]. Document clustering has been researched as a fundamental work for many fields such as data mining, information retrieval, text abstracting, etc. Clustering algorithms are used to solve the following problems: efficient division of data into clusters, determining the optimal number of clusters and determining the necessary clusters [35].

Clustering is performed in two different modes: clear and fuzzy [36]. With clear clustering, an object can belong to strictly one class. In the case of fuzzy clustering, an object can belong to all classes with different degrees of membership.

This paper examines clear clustering and presents a control-constrained clustering model formulated as a linear optimization problem with Boolean variables. The proposed model allows you to control the quality of clustering by adjusting the appropriate parameters. The quality of clustering refers to “compactness,” “similarity,” and “distance” in clustering [31,37].

The ability to download various types of web files, e-books, magazines, and various documents directly from the browser offered by wiki technologies makes the work of operators easier but requires precise control. The main task is to automatically place text documents in the appropriate directories. In other words, we need to group text documents by topic. Suppose a wiki library (DWL) contains *n* documents:

$$DWL = \{d_1, d_2, ..., d_n\}$$

It is known that when clustering, electronic documents must be in a suitable form: the proximity between documents must be determined. To do this, the text data is cleared and initial processing occurs. After this, for clustering, you need to represent documents using different mathematical models: vector model, cosine metric, etc. The most widely used model for representing texts is the vector model [35].

In this model, each e-document in the m-dimensional Euclidean space is displayed at a certain point, where m is the number of words.

Let $T = \{t_1, t_2, ..., t_m\}$ be the set of words used in a variety of DWL. Then, each document d_i on the vector model can be written in the form $d_i = [w_{i1}, w_{i2}, w_{im}](i = 1,...,n)$, where w_{ij} – weight of the word l_j in d_i. Weight w_{ij} is calculated according to the scheme *tf-idf* (term frequency – inverse document frequency) [30]:

$$w_{ij} = tf_{ij} log \frac{n}{n_j}; \, i = 1,...,n \tag{1}$$

where tf_{ij} – frequency of using the word t_j in the document d_i;

n_j – number of documents, containing the word t_j;

n – the total number of words.

After the presentation of documents to determine proximity between two vectors is defined to be the metric. To determine the semantic similarity of documents, we use the metric of the cosine. According to this metric, the proximity of the vectors $d_i = [w_{i1}, w_{i2}, w_{im}]$ and $d_l = [w_{l1}, w_{l2}, w_{lm}]$ calculated as follows:

$$sim(\mathrm{d}_i, \mathrm{d}_l) = \frac{\sum_{j=1}^{m} w_{ij} w_{lj}}{\sqrt{\sum_{j=1}^{m} w_{ij}^2 \cdot \sum_{j=1}^{m} w_{lj}^2}}, \; i, l = 1,2...,n \tag{2}$$

The proposed clustering model between documents is based on the principle of clear clustering.

Documents *WL* must be grouped between clusters $C = \{C_1, C_p, ..., C_k\}$. According to the definition of clear clustering, the following conditions must be met:

- Each document (e-document) must belong to a cluster, in other words, there should not be any documents outside the clusters:

$$\bigcup_{p=1}^{k} C_p = DWL = \{d_1, d_2, ..., d_n\}$$

- Each cluster must have at least one document:

$$C_p \neq \phi, \; p = 1,...,k;$$

- Each document must belong to only one cluster:

$$C_p \bigcap C_q \neq \phi, \; \forall p \neq q;$$

To solve the clustering problem, it is necessary to clarify some notations:

$$x_{ip} = \begin{cases} 1, \text{ if } \mathrm{d}_i \in C_p \\ 0, \text{ other wise} \end{cases}$$

O – center of the of the documents $DWL = \{d_1, d_2, ..., d_n\}$ and $O = \{O^1, O^1, ..., O^m\}$
l -th coordinate O^l is calculated as:

$$O^l = \frac{1}{n} \sum_{i=1}^{n} w_{il}, \; l = 1, 2, ..., m; \tag{3}$$

$O_p = [O_p^1, O_p^2, ..., O_p^m]$ – center of the documents C_p,
l -th coordinate O_p^l is defined as follows:

$$O_p^l = \frac{1}{|C_p|} \sum_{i=1}^{n} w_{il} x_{ip} \tag{4}$$

$|C_p| = \sum_{i=1}^{n} x_{ip}$ – number of documents in the cluster $C_p (l = 1,...,m; p = 1,...,k)$.

When clustering, you need to minimize the degree of proximity between the cluster centers and the center of all sets of documents. Then the clustering problem is formulated as follows:

$$f(x) = \sum_{p=1}^{k} \frac{1}{|C_p|} \sum_{i=1}^{n} sim(\mathrm{d}_i, R_p) \, x_{ip} \to \max \tag{5}$$

$$sim(R_p, R_q) < \varepsilon \, \forall p \neq q = 1, 2, ..., k$$

$$sim(R_p, R) < \delta \, \forall p = 1, 2, ..., k$$

$$\sum_{p=1}^{k} x_{ip} = 1 \; \forall i = 1,2,...,n$$

$$x_{ip} = \{0,1\} \; \forall i = 1,2,...,n \text{ и } \forall p = 1,2,...,k$$

Using the ε and δ parameters, we can control the clustering properties. $R = [R^1, R^2, ..., R^g]$ – document collection center DWL. Here the coordinate l is defined as follows:

$$R^l = \frac{1}{n}\sum_{i=1}^{n} w_{il};\ l = 1,2,...g \tag{6}$$

$R_p = [R_p^1, R_p^2, ...R_p^g]$ – center cluster C_p .

R_p^g is calculated as:

$$R_p^g = \frac{1}{\left|C_p\right|}\sum_{i=1}^{n} w_{il} x_{ip}, (g = 1,..., m) \tag{7}$$

$\left|C_p\right| = \sum_{i=1}^{n} x_{ip}$ – number of documents in the cluster C_p .

After clustering, the process of loading documents into the corresponding electronic catalogs follows. To automatically establish a connection between electronic documents (in our case, a wiki page) and catalogs, machine learning methods can be used.

The choice of the type of clustering algorithm is the main task affecting the accuracy of clustering. There are different types of clustering algorithms: hierarchical, partitioned, density-based, grid-based, etc. Each clustering algorithm has its disadvantages and advantages. One of the main factors affecting the accuracy of clustering algorithms is the setting of parameters such as the number of clusters, the level of detail, density threshold, distance measure, speed and scalability, and connection method.

The proposed approach is planned to be implemented in the "Electronic Library" system on the AzScienceNet platform. For the experiment, 250 documents (books, articles, magazines) for 2020-2021, stored in the Electronic Library system and not cataloged, were used. This dataset corresponds to five relevant areas: politics, technology, biology, sports, and education. To evaluate the clustering results, the "purity" criterion was used.

Each cluster may contain electronic resources belonging to different groups. The purity coefficient of cluster C_p is determined as follows [38]:

$$purity(C_p) = \frac{1}{\left|C_p\right|} max_{p^+=1,...,k^+} \left|C_p \cap C_{p^+}^+\right|, p = 1,2,...,k, p = 1,2,...k^+ \tag{8}$$

The purity coefficient always takes values from the interval $\left[\frac{1}{k^+}, 1\right]$.

k^+ – the number of groups. C^+ – is a set of groups. k – number of clusters.

To speed up the calculation, you first need to find and remove semantically similar words. To do this, the synonyms of each word and their number are determined from the WordNet lexical database. WordNet is a virtual database that identifies semantic relationships between words and is accessible by the Internet user. Using this network, we can easily identify synonyms, hypertexts, hyponyms, etc. WordNet is also important for word sentiment analysis [39].

If a set of documents contains semantically similar words, documents with similar content may fall into different clusters when clustering. This will lead to a decrease in the quality of clustering. To eliminate such cases, it is proposed to identify semantically similar words in a set of documents, leave only one of them, and delete the rest. We calculate the proximity between words using formula (2) using the cosine metric.

Semantic close words from the set of terms the removal did not hurt the quality of clustering, on the contrary, the purity coefficient received a fairly high value. Clustering accuracy is described in Table 1.

Table 1. Purity coefficient of clustering at different values of α

Number of resources	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**
	Purity				
50	0,94	0,87	0,8	0,81	0,87
100	0,98	0,96	0,97	0,97	0,97
150	0,99	0,92	0,94	0,99	0,98
200	0,97	0,87	0,91	0,98	0,98
250	0,98	0,97	0,98	0,95	0,96

This shows that the quality of clustering is high. The clustering purity coefficient was calculated for different values of (0.1, 0.2, 0.3, 0.4, 0.5). These values correspond to the selected five themes: politics, technology, biology, sports and education. Calculations show that as the number of resources increases, the number of semantically similar words increases accordingly. As can be seen from the table, as the number of library resources increased, the cleanliness factor also got a high value.

The purity coefficient indicates the ratio of the size of the dominant class in the cluster to the size of the cluster. High quality of clustering can be determined by a high purity coefficient.

6. Conclusions

There are quite a lot of electronic libraries, and these systems do not lose their relevance and become popular thanks to the latest technologies used in them. Some of these popular technologies can be called wiki technologies and intelligent algorithms.

Smart wiki libraries can play a big role in protecting the integrity and availability of data. They have greater capabilities than conventional electronic libraries. These features are as follows:

- exchange of information, joint discussion of problems and issues related to the work of the library;
- joint writing and storage of professional documentation in a publicly accessible place;
- speed and ease of information search;
- involving users in the creation of web pages on various topics;
- involving users in the compilation of catalogs;
- automatic checking of directories;
- automatic systematization and categorization of information;
- automatic organization of public archives;
- ease of creating photo galleries and adding video materials to pages;
- sending information about updating resources, invitations to vote or creating a new category using ready-made templates.

The collection and creation of documents (e-books, magazines, articles, etc.) in open electronic libraries managed by wiki technologies is mainly carried out by users. Automatic distribution of documents without operator participation in catalogs can be achieved through clear clustering. This method can make it easier and faster to find information. Solving the issue using the clustering method also provides control over documents in digital libraries.

The proposed approach can help information technology specialists create smart libraries that are easy to manage and automate certain tasks. The ease of finding information, the ability to exchange information, and the organization of a public archive are proof that smart wiki libraries are projects that meet the requirements of modern Internet users.

One of the services of the AzScienceNet network (http://www.azsciencenet.az), located at the Institute of Information Technologies of the Republic of Azerbaijan, is the “Electronic Library” service. The service was created using the Alefino software from the Exlibris company. One of the Alefino EBS services, located on the AzScienceNet network platform, is the “Intelligent Information Service” system.

The functions of the “Electronic Library” system are implemented in several modules and the menu includes “Information”, “Information”, and “Electronic catalog”. “Electronic resources”, and “Administrator menu” are grouped into sections. Applications are written in C# language based on the ASP.net platform. The “Electronic Library” uses the MARC21 database system. "

Cataloging is carried out in accordance with the rules of RAK, RSWK, and AACR. The Thesaurus module is based on the DIN 1463 standard. Alephino also has an interface for receiving and transmitting standard data from PND, SWD, and GKD. Thanks to the use of cloud technologies, it is possible to easily analyze the profiles of electronic library users and electronic resources in the library.

The document clustering methods described in our work are still relevant. They are suitable for any digital library because they do not require large supercomputers and expensive software. The use of machine learning methods, mainly clustering and classification methods, to achieve intelligent and convenient maintenance of electronic libraries allows achieving good results in searching and organizing electronic resources.

References

[1] Ignatius M. Ezeani, “Digital library deployment in a university: Challenges and prospects”, Library Hi Tech., Vol.29, No.2, pp.373-386, 2011. DOI:10.1108/07378831111138233

[2] Gary Cleveland, “Digital Libraries: definitions, issues and challenges”, International Federation of Library Associations and Institutions: UDT Core Programme, pp.1-10, 1998.

[3] Fakhar M. Manesh, Massimiliano M. Pellegrini, Giacomo Marzi, Marina Dabic, “Knowledge Management in the Fourth Industrial Revolution: Mapping the Literature and Scoping Future Avenues”, IEEE Transactions on Engineering Management, No.68, pp.1-22, 2021. DOI:10.1109/TEM.2019.2963489

[4] Li Jiahui, Wang NingXing, Duan Chao, “The Design of Smart Library Based on 5G”, Journal of Physics Conference Series

(JPCS). Vol.1606, pp.1–7, 2020. DOI:10.1088/1742-6596/1606/1/012011

[5] Yingshen Huang, Andrew M. Cox, John Cox, “Artificial Intelligence in academic library strategy in the United Kingdom and the Mainland of China”, The Journal of Academic Librarianship, Vol.49, No.6, pp.1-10, 2023. DOI:10.1016/j.acalib.2023.102772

[6] Asefeh Asemi, Adeleh Asem, “Artificial Intelligence (AI) application in Library Systems in Iran: A taxonomy study”, Library Philosophy and Practice. Vol.2, No.1, pp.1–11, 2018. https://core.ac.uk/reader/189479400

[7] Li Xuemin, Ji Mingxin, Li Hongwei, Li Junyang, “Research on the development path of university smart library based on the Internet of things”, Applied Mathematics and Nonlinear Sciences. Vol.7, No.2, pp.1075-1084, 2022. DOI:10.2478/amns.2021.2.00310

[8] Daphne P. Koller, Y. Shoham, Michael P. Wellman, Edmund H. Durfee, W.P. Birmingham, J. Carbonell, “The role of AI in digital libraries”, IEEE Expert. Vol.11, No.3, pp.8-13, 1996. DOI:10.1109/64.506746.

[9] Mohammad F. Manesh, Massimiliano M. Pellegrini, Giacomo Marzi, Marina Dabic, “Knowledge Management in the Fourth Industrial Revolution: Mapping the Literature and Scoping Future Avenues”, IEEE Transactions on Engineering Management, Vol.68, No.1, pp.289-300, 2021. DOI:10.1109/TEM.2019.2963489

[10] Andrea J. Hester, Judy E. Scott, “A conceptual model of wiki technology diffusion” Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS-41 2008), Waikoloa, Big Island, HI, USA, pp.32-32, 7-10 January 2008, DOI:10.1109/HICSS.2008.10

[11] Đorđe Stakić, “Wiki technology – origin, development and importance”, Journal of Informatics and Librarianship, Vol.10, No.1-2, pp.61-69, 2009.

[12] Christian Wagner, Karen S.K. Cheung, Rachael Ip, Stefan Böttcher, “Building Semantic Webs for e-government with Wiki technology”, Journal of Electronic Government, Vol.3, No.1, pp.36-55, 2006. DOI:10.1504/EG.2006.008491

[13] Ulrike Cress, Joachim Kimmerle, “A systemic and cognitive view on collaborative knowledge building with wikis”, International Journal of Computer-Supported Collaborative Learning (IJCSCL), Vol.3, No.2, pp.105-122, 2008. DOI:10.1007/s11412-007-9035-z

[14] Margaret M. Luo, Sophea Chea, “Wiki use for knowledge integration and learning: A three tier conceptualization”, Computers & Education, Vol.154, pp.1-16, 2020, DOI:10.1016/j.compedu.2020.103920

[15] Fernando N.V. Vlist, Anne Helmond, Marcus Burkhardt, Tatjana Seitz, “API Governance: The Case of Facebook’s Evolution”, Social Media + Societ, Vol.8, No.2, pp.1-24, 2022. DOI:10.1177/20563051221086

[16] Heather Ford, Shilad Sen, David R. Musicant, Nathaniel Miller, “Getting to the source: where does Wikipedia get its information from?”, Proceedings of the 9th International Symposium on Open Collaboration, No.9, pp.1-10, 2013. DOI:10.1145/2491055.2491064

[17] Laura Saunders, “Academic libraries’ strategic plans: Top trends and under-recognized areas”, Journal of Academic Librarianship, Vol.41, No.3, pp.285-291, 2015. DOI:10.1016/j.acalib.2015.03.011

[18] Sandy Hervieux, Amanda Wheatley, “Perceptions of artificial intelligence: A survey of academic librarians in Canada and the United States”, Journal of Academic Librarianship, Vol.47, No.1, pp. 1-11, 2021. DOI:10.1016/j.acalib.2020.102270

[19] Jeremy Frumkin, “The Wiki and the digital library”, OCLC Systems & Services, Vol.21, No.1, pp.18-22, 2005. DOI:10.1108/10650750510578109

[20] Rajesh K. Das, Mohammad Sh.U. Islam, “Application of Artificial Intelligence and Machine Learning in Libraries: A Systematic Review”, Library Philosophy and Practice, 6762, pp.1-18, 2021. DOI:10.48550/arXiv.2112.04573

[21] Beena AL., Humayoon S. Kabir, “Machine Learning and Security Applications in Digital Library”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol.9, No.1, pp.2165-2168, 2019. DOI:10.35940/ijitee.AL4718.119119

[22] Dhruba J. Borgohain, Sohaimi Zakaria, Manoj K. Verma, “Cluster Analysis and Network Visualization of Global Research on Digital Libraries during 2016–2020: A Bibliometric Mapping”, Science & Technology Libraries, Vol.41, No.3, pp.266-287, 2021. DOI:10.1080/0194262X.2021.1993422

[23] Brady D. Lund, Jinxuan. Ma, “A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering”, Performance Measurement and Metrics, Vol.22, No.3, pp.161-173, 2021. DOI:10.1108/PMM-05-2021-0026

[24] Yasir Riady, Muhammad Sofwan, Mailizar Mailizar, Turki M. Alqahtani, Lalu N. Yaqin, Akhmad. Habibi, “How can we assess the success of information technologies in digital libraries? Empirical evidence from Indonesia”, International Journal of Information Management Data Insights, Vol.3, No.2, pp.1-10, 2023. DOI: 10.1016/j.jjimei.2023.100192

[25] Mattew Bejune, “Wikis in Libraries”, Information Technology and Libraries (ITL), Vol.26, No.3, pp.26-38, 2007, DOI:10.6017/ital.v26i3.3273

[26] Ruslan A. Baryshev, Olga I. Babina, Pavel A.. Zakharov, Vera P. Kazantseva, Nikita O. Pikov, “Electronic Library: Genesis, Trends: From electronic library to samrt library”, Journal of Siberian Federal University (Humanities & Social Sciences), Vol.6, No.8, pp.1043-1051, 2015. DOI:10.17516/1997-1370-2015-8-6-1043-1051

[27] Nahak Brundaban, Padhi Satyajit, “The Role of Smart Library and Smart Librarian for E- Library Services”, Proceedings of the 12th International CALIBER-2019, Bhubaneswar, Odisha, pp.89-97. 28-30, November, 2019, https://ir.inflibnet.ac.in/handle/1944/2338

[28] Joachim Schöpfel, “Smart Libraries”, Infrastructures, Vol.3, No.4:43, pp.1–11, 2018. DOI:10.3390/infrastructures3040043

[29] Linda Cloete, "Metadata and its Applications in the Digital Library: Approaches and Practices", Library Hi Tech, Vol.27, No.2, pp.313-314. DOI:10.1108/07378830910968281

[30] Friedrich Summann, Norbert Lossau, “Search Engine Technology and Digital Libraries”, D-Lib Magazine, Vol.10, No.9, 2004. DOI:10.1045/september2004-lossau

[31] Michael Z. Zgurovsky, Yuriy P. Zaychenko, “The Cluster Analysis in Big Data Mining”, Big Data: Conceptual Analysis and Applications, Vol.58, pp.1-42, 2020, DOI:10.1007/978-3-030-14298-8_1

[32] Irada Alakbarova, "Determining the interests of Social Network Users", International Journal of Education and Management Engineering, Vol.13, No.4, pp. 1-8, 2023.

[33] Swagatam Das, Amit Konar, “Automatic image pixel clustering with an improved differential evolution”, Applied Soft

Computing, Vol.9, No.1, pp.226-236, 2009, DOI:10.1016/j.asoc.2007.12.008

[34] Jiarui Li, Yukio Horiguchi, Tetsuo Sawaragi, "Cluster Size-Constrained Fuzzy C-Means with Density Center Searching", International Journal of Fuzzy Logic and Intelligent Systems (IJFLIS), Vol.20, No.4, pp.346-357, 2020. DOI: 10.5391/IJFIS.2020.20.4.346

[35] Svetlana Simić, Zorana Banković, Jose R. Villar, Dragan Simić, Svetislav D. Simić, "A hybrid fuzzy clustering approach for diagnosing primary headache disorder", Logic Journal of the IGPL, Vol.29, No.2, pp.220-235, 2021. DOI:10.1093/jigpal/jzaa048

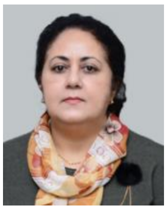
[36] Sara I.R. Rodríguez, Francisco A.T. Carvalho, "Fuzzy clustering algorithms with distance metric learning and entropy regularization", Applied Soft Computing, Vol.113, Part A, pp.1-22, 2021. DOI:10.1016/j.asoc.2021.107922

[37] Adil M. Bagirov, Ramiz M. Aliguliyev, Nargiz Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients", Pattern Recognition, Vol.135, pp.1-15, 2023, DOI:10.1016/j.patcog.2022.109144

[38] Ramiz M. Aliguliyev, "Performance evaluation of density-based clustering methods", Information Sciences, Vol.179, No.20, pp.3583-3602, 2009. https://doi.org/10.1016/j.ins.2009.06.012

[39] Hanane Ezzikouri, Youness Madani, Mohammed Erritali, Mohamed Oukessou, "A New Approach for Calculating Semantic Similarity between Words Using WordNet and Set Theory", Procedia Computer Science, Vol.151, pp.1261-1265, 2019. DOI:10.1016/j.procs.2019.04.182

Authors' Profiles



Irada Yavar Alakbarova in 1984 she graduated from the Faculty of Automation of production processes, Azerbaijan Institute of Oil and Chemistry named after M. Azizbayov. In the same year, she was accepted for employment at the Institute of Information Technology. In currently holds the post of Sector Chief of the Institute of Information Technology. In 2018, the defense of the dissertation on the "Development of methods and algorithms for analysis of information war technologies in a wiki environment" and she received his Ph.D. (2018). In currently conducts research in the field of Social Network Analysis, Text Analysis, Clustering, Social Credit Analysis, and Big Data Analytics. She is the author of 48 articles and three books.

Dilbar Ilkham Alizada works as a junior researcher at the Institute of Information Technologies, Baku. Scientific interests: e-government, e-governance, intelligent systems. Author of 5 articles.

How to cite this paper: Irada Alakbarova, Dilbar Alizada, "A New Approach to Improving Search Efficiency in Digital Libraries", International Journal of Intelligent Systems and Applications(IJISA), Vol.16, No.2, pp.13-23, 2024. DOI:10.5815/ijisa.2024.02.02