УДК 026:004-0.25.27

**Олександр Кузнєцов**,

ORCID https://orcid.org/0000-0002-9902-1295,

Завідувач відділу наукового комплектування та опрацювання бібліотечних фондів,

Державна науково-технічна бібліотека України (Київ, Україна)

вул. Антоновича 180, Київ, 03150

e-mail: nkof@dntb.gov.ua

**Віктор Заїка**,

ORCID https://orcid.org/0009-0003-6582-6524,

кандидат фізико-математичних наук,

провідний інженер відділу бібліометрії і наукометрії,

Національна бібліотека України імені В. І. Вернадського (Київ, Україна)

просп. Голосіївський, 3, Київ, 03039

e-mail: victor.zayika@gmail.com

**КЛАСТЕРИЗАЦІЯ НАДХОДЖЕНЬ ДО ЕЛЕКТРОННОЇ БІБЛІОТЕКИ**

Розглянуто та практично застосовано метод кластеризації для виділення подібних за тематикою груп вхідних документів, що може бути корисним при створенні та наповненні електронної бібліотеки (ЕБ). В якості характеристики, за якою проведено порівняння та групування електронних документів (ЕД) використано міру косинусної подібності, для знаходження якої кожен електронний документ було векторизовано, тобто, знайдено відповідний частотний масив лексем. Для кожного набору двох векторів: частотного масиву файлу та частотного масиву корпусу, розраховано коефіцієнт тематичного напрямку (КТН). Побудовано розподіл КТН на площині та графіко-аналітично виявлено кластери ЕД.

*Ключові слова*: електронний документ, електронна бібліотека, систематизація, векторизація, кластер, кластеризація.

Опрацювання великих масивів електронних документів неможливо без використання сучасних підходів та програмних інструментів.

В традиційній бібліотечній сфері термін опрацювання документу обов’язково включає систематизацію, тобто, присвоєння індексу УДК (універсальної десяткової класифікації), предметної рубрики, тощо. Іншими словами, результатом опрацювання документа є віднесення його до певної групи подібних документів, об’єднаних за тематичним напрямком.

Наявність документів в цифровому вигляді відкриває ряд можливостей, зокрема, дозволяє класифікувати тексти за певними категоріями програмно, в автоматичному режимі. З цією метою використовується кластерний аналіз або кластерізація – статистична обробка множини вхідних даних (електронних документів), результатом застосування якої є виділення відносно однорідних груп, або кластерів, за подібністю певних характеристичних ознак. У випадку електронної бібліотеки критерієм за яким визначається кластер або групуються ЕД, доцільно взяти зміст або тематичний напрям електронного документа.

У традиційних бібліотеках такий процес називається систематизацією і здійснюється систематизатором виключно в ручному режимі. Систематизатор інтелектуально (аналітико-синтетично) опрацьовує документ та виокремлює його змістовні та формальні у вигляді класифікаційних індексів. Як правило, систематизатор не має змоги ознайомитися зі змістом документу повністю, а формулює свої узагальнення лише на основі аналізу назви та анотації.

Беззаперечними перевагами застосування кластерного аналізу до масиву електронних документів є швидкість самого процесу та глибина опрацювання ЕД у порівнянні з традиційною систематизацією, що є особливо актуальним в сучасних умовах лавиноподібного зростання кількості ЕД в електронній бібліотеці.

Для демонстрації можливостей кластерного аналізу було використано перші тридцять п’ять повнотекстових файлів, що надійшли до ДНТБ України в 2023 році. Тематичний напрямок аналізованих файлів заздалегідь не був відомий. КТН кожного текстового файлу визначався відносно свого корпусу, текстовий вектор якого формується з усіх файлів крім аналізованого.

Побудова вектору КТН всіх досліджуваних текстів – їхнього розподілу за зростанням КТН – дозволив виявити кластери – групи текстів, подібних за змістом та тематикою. Результати представлено на Рис. 1.

Рис. 1. Розподіл КТН розглянутих файлів ЕБ ДНТБ

Достовірним критерієм віднесення документів до одного й того ж кластеру є належність КТН файлів достатньо вузькому діапазону значень.

Таким чином,

* використовуючи метод кластерного аналізу проаналізовано повні тексти ЕД, а не лише анотації та назви.
* знайдено КТН кожного файлу та виявлено 5 кластерів.
* число кластерів, не відоме заздалегідь, визначено об’єктивно, в якості критерію використано тематичний напрям документа.
* тематичну рубрику кожного кластеру можна визначити зі спільної частини індексу УДК файлів, що входить до одного й того ж кластеру.
* кількість тематичних напрямів визначається кількістю кластерів.

**UDC** 026:004-0.25.27

**Oleksandr Kuznetsov**,

https://orcid.org/0000-0002-9902-1295,

Head of department of Scientific Document Processing and Cataloging,

The State Scientific and Technical Library of Ukraine (Kyiv, Ukraine)

180, Antonovycha str., Kyiv, 03150

e-mail: nkof@dntb.gov.ua

**Victor Zayika**,

https://orcid.org/0009-0003-6582-6524,

Candidate of physical and mathematical sciences,

Engineer of the department of bibliometrics and scientometrics,

V. I. Vernadskyi National Library of Ukraine (Kyiv, Ukraine)

3, Holosiivskyi Avenue, Kyiv, 03039, Ukraine

e-mail: victor.zayika@gmail.com

**CLUSTERIZATION OF RECEIPTS TO THE ELECTRONIC LIBRARY**

The clusterization for selecting groups of input documents similar in subject matter is considered and practically applied, which can be useful when creating and filling an electronic library. The cosine similarity measure was used as a characteristic by which the comparison and grouping of electronic documents was carried out. To find cosine similarity measure each electronic document was vectorized, i.e., a corresponding frequency array of lexems was found. For each set of two vectors (the first one is the frequency array of the file and the second one is the frequency array of the corpus) the coefficient of thematic direction is calculated. The distribution of coefficient of thematic direction on the plane was plotted and electronic documents clusters were identified graphically and analytically.

*Keywords*: electronic document, electronic library, systematization, vectorization, cluster, clusterization.