УДК 004.9

**ВИБІР ЗАСОБІВ МОВИ R ДЛЯ РОЗВ’ЯЗАННЯ ЗАДАНОЇ   
ПРИКЛАДНОЇ ЗАДАЧІ СИСТЕМНОГО АНАЛІЗУ**

**Дратований Михайло**,аспірант кафедри системного аналізу, комп’ютерного моніторингу та інженерної графіки,

**Довгополюк Сергій**, аспірант кафедри системного аналізу, комп’ютерного моніторингу та інженерної графіки,

Вінницький національний технічний університет, Україна

Є багато важливих прикладних задач системного аналізу, наприклад, задача аналізу часових рядів даних про стан системи [1, 2]. Часові ряди, як відомо, є найбільш поширеним класом даних моніторингу стану будь-яких систем, процесів та явищ. Велика кількість задач в техніці, економіці, екології та ін. галузях потребує їх аналізу, спрямованого, передусім на забезпечення можливості їх прогнозування за тих чи інших умов. Накопичено вже значний досвід у розв’язанні цих задач. В наш час найбільшу прикладну цінність викликає не стільки розвиток теорії аналізу часових рядів, скільки розвиток готових до застосування комп’ютерних засобів для автоматизації такого аналізу.

Останні роки у сфері комп’ютеризованої обробки даних домінуюче місце у світі займає мова програмування R. Мова R — [мова програмування](https://uk.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B2%D0%B0_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F) і програмне середовище для [статистичних](https://uk.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) обчислень, аналізу та представлення даних у графічному вигляді.

R розповсюджується безкоштовно за ліцензією [GNU General Public License](https://uk.wikipedia.org/wiki/GNU_General_Public_License) у вигляді вільнодоступого вихідного коду або відкомпільованих бінарних версій більшості операційних систем: [Linux](https://uk.wikipedia.org/wiki/Linux), [FreeBSD](https://uk.wikipedia.org/wiki/FreeBSD), Microsoft Windows, [Mac OS X](https://uk.wikipedia.org/wiki/Mac_OS_X), [Solaris](https://uk.wikipedia.org/wiki/Solaris). R використовує текстовий інтерфейс.

Для аналізу часових рядів станом на початок 2017 року в мові R є 228 пакетів засобів для аналізу часових рядів, в кожному з яких є від 3 до 80 функцій. Для розв’язання заданої прикладної задачі з аналізом певних видів часових рядів є проблематичним швидко визначити які саме функції яких саме пакетів мови R є оптимальними для розв’язання задачі із заданими умовами. (<http://web.mit.edu/~r/current/arch/amd64_linux26/lib/R/doc/manual/R-lang.pdf>).

Щоб вручну вибрати бібліотеки із заданими функціями, потрібно вивчити опис біля 230 R-пакетів, тому раціональніше застосувати сервіс пошуку в текстових масивах за ключовими словами. Основною проблемою аналогів таких сервісів є їхня лише умовна безкоштовність або взагалі використання на платних умовах. Інша проблема — це частота оновлення бази даних таких сервісів. Найбільш популярною системою-аналогом є «Site Content Analyzer 3» **(**http://www.cleverstat.com/ru/sca-website-analysis-software-index.htm**)**. Цей сервіс є платним. Іншим сервісом такого типу є Fastkeywords.biz, він – умовно безкоштовний, але інформація в базі даних рідко оновлюється.Такі сервіси працюють з текстовими документами та веб-сайтами, але не з pdf [3].

# Інший підхід полягає у вивченні спеціальної веб-сторінки “Cran Task Views” (https://cran.r-project.org/web/views/), яка містить інформацію про основні можливості основних R-пакетів, причому ця сторінка регулярно оновлюється.

Альтернативним способом вибору R-пакетів є аналіз тексту pdf-файлів з описом можливостей кожного такого пакету. Часто для аналізу текстових масивів використовують латентно-семантичний аналіз (ЛСА) – метод обробки інформації природною мовою, що дозволяє проаналізувати взаємозв'язок між колекцією документів і термінами, які в них зустрічаються. Зіставляє деякі фактори (теми) всім документам і термам [4]. А для представлення критеріїв пошуку та сортування часто використовують cемантичну мережу — інформаційну модель предметної області, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної області, а ребра задають відносини між ними. Об'єктами можуть бути поняття, події, властивості, процеси [5].

Ми пропонуємо ще один підхід. Користувач формує масив pdf-документів та масив ключових слів, за якими проводиться пошук у документах. Програма формує звіт, в якому вказується файл, сторінка та рядок із ключовими словами. Також формується звітний файл ранжування pdf–документів за частотою згадування ключових слів у них. Таку програму можна написати також на R з використанням засобів R-пакету «pdfsearch»**.**

Отже, проведено аналіз сучасних підходів до вибору засобів мови R для розв’язання заданої прикладної задачі системного аналізу та запропоновано авторський підхід.

**Список використаної літератури**

1. Мокін Б. І. Математичні методи ідентифікації динамічних систем : навчальний посібник / Б. І. Мокін, В. Б. Мокін, О. Б. Мокін. – Вінниця : ВНТУ, 2010. – 260 с.
2. Мокін В. Б. Ідентифікація математичної моделі гідрологічних процесів на гідропості "Селище" річки Південний Буг / В. Б. Мокін, А. Р. Ящолт // Вісник Вінницького політехнічного інституту. — 2005. — № 6. — С. 85–88.
3. Shianghau Wu (2016): The application of 3D fruit fly optimization algorithm to the keywords analysis of Macau’s international relations, Intelligent Automation & Soft Computing.
4. Формальні методи образного аналізу та синтезу природно–мовних конструкцій : монографія / О. В. Бісікало. – Вінниця : ВНТУ, 2013. – 316 с.
5. Розробка комплексної моделі інформаційно-пошукової веб-системи відкритих числових даних / В. Б. Мокін, С. О. Довгополюк, М. П. Боцула, М. В. Коханський // Вісник Вінницького політехнічного інституту. — 2017. — № 1 — С. 62-69.