Вочевидь, що для кожного растрового зображення треба створити індивідуальну текстову мапу. Ручна побудова їх – трудомісткий процес. А використання редактора E-Colorer може бути ефективним лише при наявності простого способу створення текстових мап. Тому нами у межах проекту E-Colorer розроблено редактор текстових мап E-Mapedit, який автоматизує процес їх створення. Редактор E-Mapedit розпізнає текст вихідного зображення і в автоматичному режимі будує цифрову мапу. Результат роботи завантажується у вікно редагування, в якому користувач має можливість у ручному режимі відкоригувати текстову мапу та виправити помилки розпізнавання, які можуть мати місце при поганій читаності вихідного документа. Демонстраційна версія редактора E-Mapedit доступна для завантаження з електронного ресурса за адресою: ftp://micrography.gov.ua/pub/demo/ecolorer. Для встановлення демонстраційної версії редактора на персональний комп'ютер необхідно скористатися інструкцією howto-install.pdf, яка знаходитися за вказаною вище адресою.

Підхід до формалізації атрибутів розпізнаного тексту растрових зображень, запропонований нами, відкрив можливість реалізувати механізм повнотекстового пошуку і наочного відображення його результатів на растрових зображеннях текстів, що містяться в електронних документах. Основними перевагами реалізованого механізму є:

– простота у використанні;

– невимогливість до обчислювальних ресурсів операційної системи;

– використання відкритого формату XML.

Крім використання реалізованого механізму для поліпшення якості забезпечення документами користувачів електронних архівів, перспективними також є можливості його використання для створення електронних бібліотек і в більшості напрямків формування та використання страхового фонду документації України.

Борис Березін, Дмитро Ланде

ДОСЛІДЖЕННЯ СТАНУ ОПТИЧНИХ НОСІЇВ ПРИ ДОВГОТЕРМІНОВОМУ ЗБЕРІГАННІ ЦИФРОВОЇ ІНФОРМАЦІЇ

Серед основних напрямків дослідження стану оптичних носіїв можна виділити дві групи моделей: прискореного та природного старіння. При прискореному старінні для тестування оптичних носіїв

з метою оцінки строку служби при довготерміновому зберіганні як правило використовуються моделі прискореного старіння Ейрінга або Арреніуса[1].

Проведений аналіз досліджень з прискореного та природного старіння оптичних носіїв показав обмеженість даних щодо характеристик природного старіння CD та особливо DVD, BD, UDO дисків. З метою дослідження характеристик природного старіння DVD дисків, порівняння їх із даними попередніх досліджень та визначення можливостей подальшого використання для управління інфраструктурою довготермінового зберігання інформації нами створено базу даних, що відповідає колекції DVD дисків. По кожному диску база містить його номер, час запису інформації, об'єм інформації, тип носія, ідентифікатор виробника диску, час тестування, значення показника помилок, оцінку зовнішнього стану диску тощо. В 2012 р. проведено тестування вибіркового масиву приблизно з 150 носіїв із колекції DVD дисків, записаних у 2006–2012 рр. Для вимірювання щільності помилок використовується значення PIE (Parity Inner Error) – кількість рядків парності блоку ECC із помилками (Error Correction Code – код корегування помилок), а точніше PI Sum 8 – значення для 8 послідовних ECC з блоку. Максимальне допустиме значення PI Sum 8 складає 280 помилок.

Для виявлення особливостей розподілу характеристик DVD-дисків при природному старінні дані про вибірку з 150 носіїв було проранжировано за кількістю помилок. Отриманий розподіл подано на рис. 1. Він може бути апроксимований за допомогою степеневої функції (Power Law) із степеневим показником – 0,827 та достовірністю апроксимації 0,87.

Для порівняння характеристик CD та DVD дисків при природному старінні аналогічне ранжирування виконано для частини даних, розрахованих на основі результатів дослідження колекції CD-дисків Бібліотеки Конгресу США в 1999–2003 роках[2]. Отриманий розподіл теж може бути апроксимований за допомогою степеневої функції із степеневим коефіцієнтом – 0,724 та достовірністю апроксимації 0,91.

Отримані значення показників стану носіїв для виборок колекції цифрової інформації, записаної в 2006–2012 рр., дають змогу порівнювати стан DVD – дисків колекції та вибирати необхідні моменти міграції файлів колекції на нові носії.

Висока ступінь достовірності при апроксимації отриманих розподілів степеневою функцією підтверджує відповідність процесу ста-

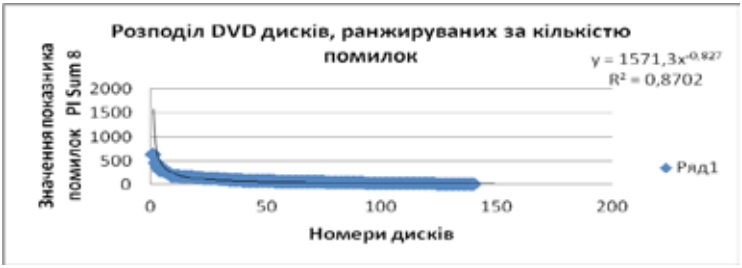

Рис. 1. Дані про вибірку з 150 DVD дисків, ранжирувані за кількістю помилок з апроксимацією степеневою функцією

ріння CD та DVD дисків та розподілу помилок закономірностям, що визначаються універсальною закономірністю Парето, завдяки якій можна зробити оцінку живучості інформаційних об'єктів, поданих на цих носіях. Відомо, що живучість інформаційного об'єкта оцінюється як ймовірність того, що об'єкт буде непошкодженим протягом визначеного періоду часу t при визначених умовах[3].

Якщо інформаційний об'єкт зберігається частинами на n носіях інформації, то ймовірність руйнування цього об'єкта оцінюється як:

У цьому добутку $F_i(t)$ – ймовірності руйнування i –го носія за час t.

Відповідно живучість оцінюється як:

Маючи на увазі те, що ймовірність виникнення помилок на носіях пропорційна часу існування цих носіїв, що доведено даними вимірів, і те, що розподіл помилок має степеневий розподіл, можна вважати доцільним і обґрунтованим дослідження моделі із степеневим розподілом помилок, що принципово відрізняється від підходів, в яких використовується пуасонівський потік помилок (теорія систем масового обслуговування) та розподіл помилок за Вейбулом[4]. У цьому випадку, живучість можна оцінювати як:

$$S_n(t) = 1 - \prod_{i=1}^{n} F_i(t) = 1 - \prod_{i=1}^{n} Ct^{-\beta} = 1 - C^n t^{-n\beta},$$

де C, β – деякі константи.

Виявлені закономірності статистичного розподілу помилок дають підстави робити висновки, пов'язані з живучістю інформаційних об'єктів, що розміщуються на оптичних носіях даних, а саме врахову-

вати явища самоподібності, нерегулярності виникнення помилок, наявність «товстого хвосту» в розподілі, що характеризує надзвичайно велику кількість носіїв із незначною кількістю помилок тощо.

Подані вище залежності для оптичних дисків разом з відповідними характеристиками інших видів носіїв можна використати при побудові інструментальних засобів управління інфраструктурою довготермінового зберігання інформації для підвищення її ефективності та живучості.

---

[1] Standard ECMA-396. Test Method for the Estimation of Lifetime of Optical Media for Long-term Data Storage, 2010. – 44 p.

[2] *Shahani C. J., Manns B., Youket M.* Longevity of CD Media Research at the Library of Congress / Preservation Research and Testing Division Library of Congress. – Washington DC, USA, 2005. – 14 p.

[3] *Li Y., Miller E. L., Long D. D. E.* Understanding Data Survivability in Archival Storage Systems // Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR 2012), June 4–6, 2012, Haifa, Israel.

[4] Ibid.