Система штучного інтелекту трактуватиметься як технічна система, аналогічна загальновідомим системам автоматичного регулювання. Така система використовує моделі, апарат і прийоми, запозичені з різних дисциплін: психології, лінгвістики, інформатики, дискретної математики, системного програмування, науки обчислень і ін.

Висновки

Таким чином, в основу інтелектуальної автоматизованої системи оцінювання знань покладено алгоритм перевірки граматики, семантики і прагматики. Для нечіткого порівняння окремих слів у відповіді в алгоритмі використовується метрика Левенштейна, однак для ефективнішої перевірки додатково здійснюється аналіз структури речення. Експериментальна перевірка показала достатню ефективність запропонованого алгоритму при перевірці питань тесту, в яких відповідь потрібно навести у вільному викладенні. Однак для подальшого розвитку пропонується використання семантичного аналізу з подальшим порівнянням семантичних мереж зразка і текстової відповіді.

У роботі алгоритмів задіяні моделі штучного інтелекту. Такі алгоритми, в яких використані елементи штучного інтелекту, у частині побудови семантичних мереж сприятимуть суттєвому підвищенню ефективності роботи системи оцінювання знань студентів.

ЛІТЕРАТУРА

1. *Вакарчук О.С.* Екзаменаційна сесія як вона є. Режим доступу: http://forum.osvita.org.ua

2. *Шидло Г.М.* Использование аппарата теории нечетких множеств для реализации алгоритма оценки обучаемого: материалы Междунар. науч.-техн. конф. [«Информационные системы и технологии»] (Новосибирск, Россия 22-25.04. 2003 г.) / Изд. НГТУ, 2003. – С. 79–84.

3. *Рудинский И.Д.* Интеллектуальная система контроля знаний – новый подход к автоматизации педагогического тестирования. – Режим доступу: http://lib.convdocs.org/docs

4. *Рудинский, И.Д.* Создание интегрированной автоматизированной системы контроля знаний / *И.Д. Рудинский* // Информатика и образование. – 2005. – №2. – С. 117–122.

5. *Аскеров Э.М., Емелин М.А* и др. Принципы и технологии создания интегрированной автоматизировнной системы контроля знаний. – КГТУ, 2008. – 54 с.

6. *Ермаков А.Е.* Извлечение знаний из текста и их обработка: состояние и перспективы / *А.Е. Ермаков* // Информационные технологи. – М: Новые технологии, 2009. – С. 50–55.

УДК 004.63:004.75

МОДЕЛЮВАННЯ ЖИВУЧОСТІ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ ПІД ЧАС ДОВГОТЕРМІНОВОГО ЗБЕРІГАННЯ ВЕЛИКИХ ОБСЯГІВ ДАНИХ


Д.В. Ланде, *докт. техн. наук,*
Б.О. Березін

Постановка проблеми і її актуальність. За оцінками IDC, обсяг інформації, що створювалася у світі в 2006 р., був порівнянним з обсягом ресурсів, доступним для її зберігання. Але вже у 2007 р. інформації було створено більше, ніж засобів для її зберігання. У 2008 р. з'являється термін Big Data, пов'язаний спочатку з проблемою зростання і різноманіття наукових

даних. До Big Data відносять нові технології, що об'єднують, аналізують дані й забезпечують добування цінних знань з великих обсягів даних різного типу шляхом швидкого доступу. Зростання загального обсягу інформації, що зберігається, спричинює зростання обсягів даних, які мають зберігатися довготерміново.

Довготермінове зберігання інформації передбачає послідовність дій, стратегій, необхідних для гарантування доступу до цифрового контенту протягом потрібного терміну, незважаючи на відмови носіїв і обладнання, зміни технологій тощо. Серед основних загроз довготермінового зберігання даних слід відзначити такі: пошкодження носіїв інформації; старіння носіїв/обладнання, програмного забезпечення/форматів; помилки операторів; мережні атаки; природні катастрофи тощо. Для подолання цих загроз застосовуються стратегії збільшення середнього часу помилок, зменшення середнього часу відновлення, аудиту, реплікації, міграції, різноманітності, прозорості й ін.

Таким чином, унаслідок значного зростання обсягів інформації, яка має довготерміново зберігатися в електронному вигляді, підвищуються вимоги до живучості інформаційних об'єктів, файлів, тобто спроможності об'єкта виконувати свої основні функції в умовах негативних впливів, тимчасово або постійно відмовляючись від виконання менш важливих функцій [1].

Аналіз досліджень і публікацій. Світовий досвід свідчить, що для вирішення цієї проблеми і зменшення витрат на зберігання недостатньо розвитку традиційного апаратного і програмного забезпечення. Необхідно створювати нові засоби – математичні моделі зберігання і побудовані на їхній основі інструментальні засоби, програмні пакети для вибору стратегій, правил, для планування й оптимізації процесу зберігання. Більшість відомих досліджень, спрямованих на вирішення проблем довготермінового зберігання великих обсягів даних, присвячені моделям надійності систем зберігання і різних видів носіїв. Зважаючи на те, що довготермінове зберігання відбувається в умовах негативних впливів, пов'язаних із відмовами і старінням систем зберігання, носіїв, програмного забезпечення (ПЗ), форматів даних тощо, це дослідження присвячене вивченню проблеми живучості інформаційних об'єктів (ІО).

Серед моделей довготермінового зберігання можна відзначити ефективну стратегію управління надійністю збереження даних у хмарних дата-центрах [2]. Для налагодження зв'язку між потоками відмов і тривалістю зберігання в ній було вибрано експоненціальний розподіл. Для довготермінового зберігання в разі зміни форматів і міграції розроблено інструментальний засіб ReproSim [3], призначений для імітаційного моделювання й оцінки цифрових репозиторієв протягом певного часу (використовує нормальний розподіл або розподіл Вейбулла). Послідовна міграційна стратегія є одним із можливих шляхів підтримки інформаційних об'єктів. Об'єкти конвертуються в нові формати у визначені інтервали часу, і частота таких міграцій залежить від дійсності і доступності відповідних форматів або їхніх сімейств. Для моделі [4] характерною є особливість довготермінового зберігання великих обсягів даних, пов'язана з необхідністю міграції даних, обміну даними внаслідок старіння форматів тощо. Під час планування способів зберігання великих обсягів даних необхідно враховувати час, що витрачається на міграцію даних, оскільки він може обчислюватися місяцями або навіть роками. У роботі [4] запропоновано відповідну математичну модель для оцінки часу, яка використовується для тестування режимів міграції даних у Національній бібліотеці Норвегії.

Проведений аналіз досліджень довготермінового зберігання [5; 6] показав обмеженість даних щодо характеристик природного старіння оптичних носіїв. Більшість досліджень, присвячених відмовам під час зберігання даних, використовують надійнісні, експоненціальні моделі, моделі Вейбулла, Пуасона

[2; 3; 8]. Моделі Парето, які застосовуються в економіці, для аналізу зберігання використовуються недостатньо. З точки зору живучості інформаційних об'єктів у цій роботі розглядаються моделі природного старіння DVD- і CD-дисків, а також зберігання в розподілених системах з використанням степеневих законів розподілу.

Живучість інформаційних об'єктів під час довготермінового зберігання

З метою дослідження характеристик природного старіння DVD-дисків, порівняння їх з даними попередніх досліджень і визначення можливостей подальшого використання для управління інфраструктурою довготермінового зберігання інформації було створено базу даних колекції DVD-дисків. База містить номер кожного диска, час запису інформації на диск, обсяг інформації, тип носія, ідентифікатор виробника диска, час тестування, значення показника помилок, оцінку зовнішнього стану диска тощо [5; 6]. У 2012 р. було проведено тестування вибіркового масиву приблизно зі 150 носіїв із колекції DVD-дисків, записаних у 2006–2012 рр. Для вимірювання щільності помилок використовується значення PIE (Parity Inner Error) – кількість рядків парності блока ECC із помилками (Error Correction Code – код корегування помилок), а точніше PI Sum8 – значення для восьми послідовних ECC з блока. Максимально допустиме значення PI Sum8 становить 280 помилок.

Для виявлення особливостей розподілу характеристик DVD-дисків під час природного старіння дані про вибірку зі 150 носіїв було проранжировано за кількістю помилок (рис.1). Отриманий розподіл може бути апроксимований за допомогою степеневої функції (Power Law) із степеневим показником 0,827 і достовірністю апроксимації 0,87. Для порівняння характеристик CD- і DVD-дисків під час природного старіння аналогічне ранжирування було виконано для частини даних, розрахованих на основі результатів дослідження колекції CD-дисків Бібліотеки Конгресу США в 1999 р. [7]. Отриманий розподіл наведено на рис. 2. Він теж може бути апроксимований за допомогою степеневої функції зі степеневим коефіцієнтом 0,724 і достовірністю апроксимації 0,91.

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

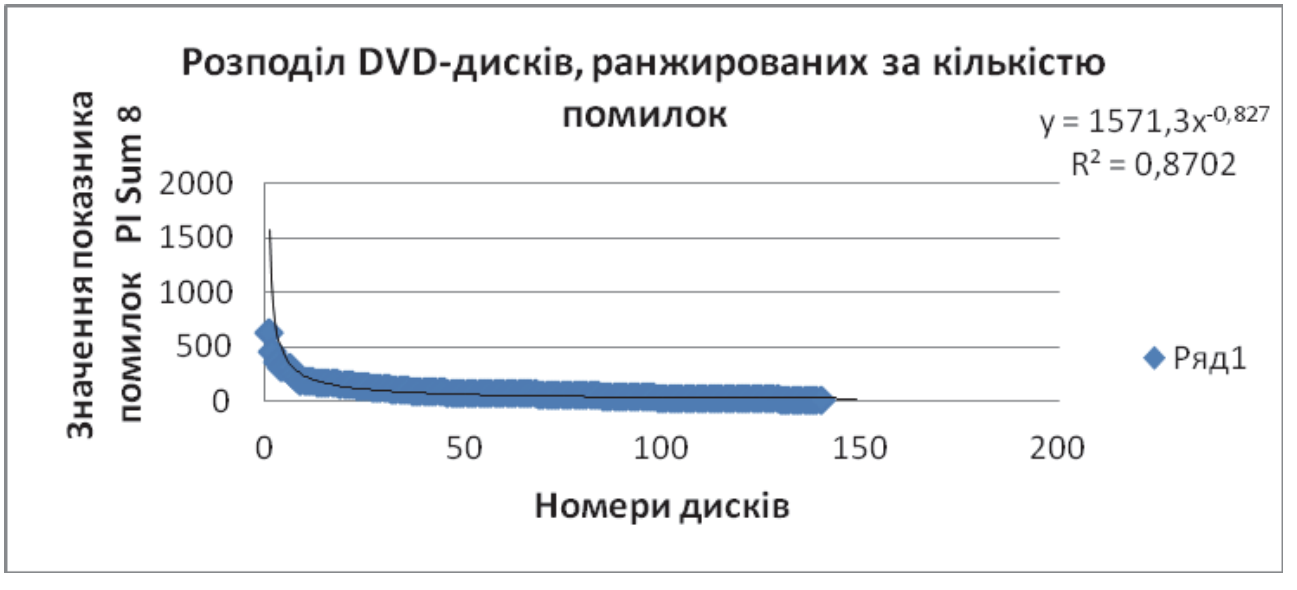

Рис. 1. Дані про вибірку зі 150 DVD-дисків, ранжировані за кількістю помилок з апроксимацією степеневою функцією

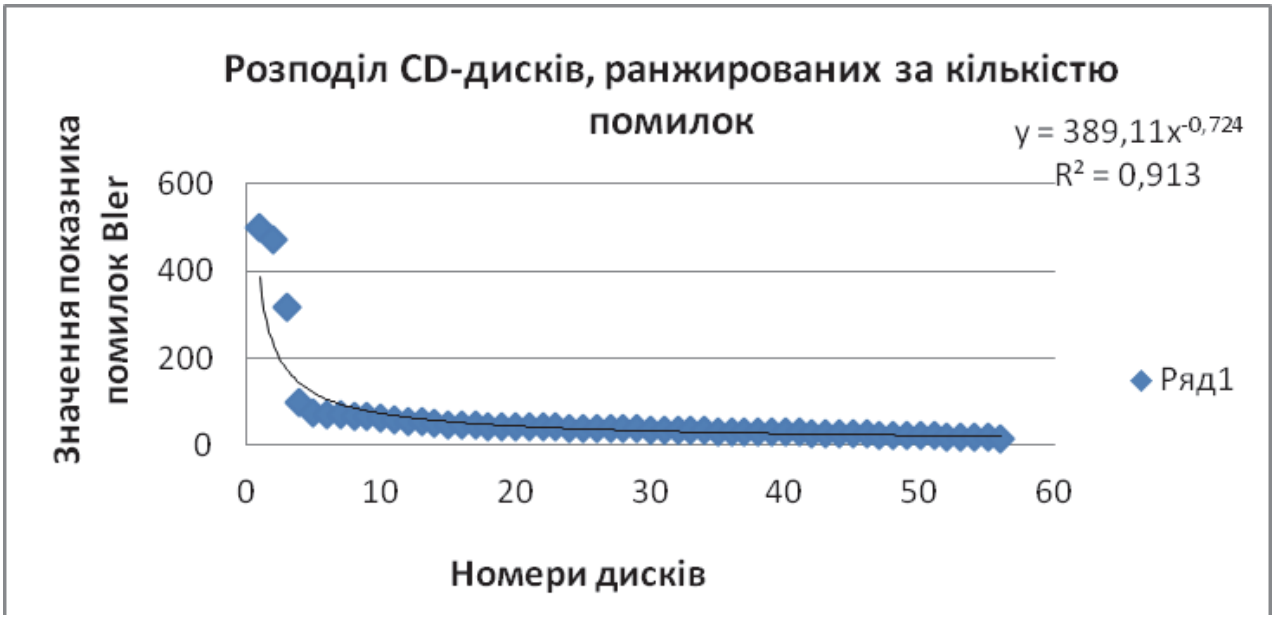

Рис. 2. Ранжировані дані про частину вибірки зі 125 CD-дисків (дані розраховані на основі результатів дослідження [7]) з апроксимацією степеневою функцією

цесу старіння CD- і DVD-дисків і розподілу помилок закономірностям, що визначаються універсальною закономірністю Парето, завдяки якій можна зробити оцінку живучості інформаційних об'єктів, що розміщуються на цих носіях. Відомо, що живучість інформаційного об'єкта оцінюється як імовірність того, що об'єкт залишатиметься непошкодженим протягом визначеного періоду часу t за визначених умов [8]. Якщо інформаційний об'єкт зберігається частинами на n носіях інформації, то ймовірність руйнування цього об'єкта оцінюється як:

$$F_{lost}(t) = \prod_{i=1}^{n} F_i(t).$$

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

Відповідно живучість оцінюється як:

$$S_n(t) = 1 - F_{lost}(t) = 1 - \prod_{i=1}^{n} F_i(t).$$

Ураховуючи те, що ймовірність виникнення помилок на носіях пропорційна часу існування цих носіїв (доведено даними вимірів) і те, що розподіл помилок має степеневий розподіл [5], можна вважати доцільним і обґрунтованим дослідження моделі зі степеневим розподілом помилок, що принципово відрізняється від підходів, в яких використовується пуасонівський потік помилок (теорія систем масового обслуговування) і розподіл помилок за Вейбуллом [8]. 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.

Живучість інформаційних об'єктів під час зберігання великих обсягів даних у мережах

Для довготермінового зберігання великих обсягів даних використовуються розподілені системи. Наприклад, пірінгові, децентралізовані мережі (Peer-to-Peer або P2P). Розподілені системи зберігання підвищують опірність до відмов за допомогою реплікації. Доступність даних [9–12] є складною функцією індивідуальної доступності вузлів,

розподілу корельованих відмов (Correlated Failures) вузлів тощо. Одночасні відмови на великій кількості вузлів можуть зменшити ефективність реплікації. Тому важливо враховувати статистику корельованих відмов [11–13] для визначення доступності даних. Вибух відмов (Failure Burst) [11] визначається відносно вікна розміром w як максимальна послідовність відмов вузлів, кожний з яких відбувається всередині часового вікна *w*.

Оскільки корельовані відмови мають значний вплив на показник доступності даних, їхні характеристики в цій роботі аналізувалися за допомогою вікна спостереження (часового вікна). Тобто була розроблена імітаційна модель множинних відмов – залежності кількості часових вікон з відмовами, що відбуваються всередині вікна від заданого розміру вікон і числа відмов, що спостерігаються у вікнах (рис. 3; 4). З її допомогою моделювались особливості характеристик моментів відмов у разі їхнього експоненціального і степеневого розподілів. Загрози втрати ІО при використанні реплік пов'язані з виникненням кількох одночасних або близьких за часом відмов, коли стає неможливим відновлення. Відновлення пошкоджених даних за рахунок інших реплік може стати неможливим через брак часу на перезапис даних, заміну обладнання, перезапуск системи тощо. Модель множинних відмов надає можливість оцінити живучість ІО, оскільки характеризує негативні впливи, в умовах яких забезпечується доступність ІО.

У процесі моделювання генерувались за допомогою розроблених на R-мові програмних засобів моменти відмов, розподілені за експоненціальним і степеневим законами. Із загальної послідовності генерованих значень часу між сусідніми відмовами визначалися послідовності відмов, які відбувалися всередині часового вікна заданого розміру. Потім було розраховано кількості відмов у цих послідовностях. Після цього визначалося число вікон для кожної кількості відмов у послідовностях. Описана процедура повторювалась для отримання статистичних даних і подальшої побудови відповідних залежностей. Отже, для визначення характеристик впливу відмов серверів на живучість ІО, що зберігаються на них, можна розглядати, зокрема, кількісні значення послідовностей у деякому сенсі близьких за часом моментів відмов серверів.

Саме ці значення – кількості часових вікон, що відповідають кожному числу близьких послідовних відмов, які відбувалися всередині часового вікна, у подальшому позначаються як кількість множинних відмов, і зображаються по осі аплікат (Z) на рис. 3; 4; по осі X зображено задаваний розмір часових вікон, який фактично визначає близькість спостережуваних моментів послідовних відмов у часі. По осі Y зображені кількості близьких послідовних моментів відмов, що спостерігалися в часових вікнах відповідного розміру. Аналіз поверхонь, зображених на цих рисунках, показує, що за експоненціального розподілу відмов більшість часових вікон припадає на вікна з максимальним значенням часу спостереження, а за степеневого розподілу – на вікна з малим значенням часу. Вікна з малим значенням часу спостереження, в які потрапляють близькі в часі відмови, і відповідні їм значення кількості близьких за часом відмов, а також відповідні кількості вікон, є найбільш складними з точки зору забезпечення доступності даних і живучості ІО.

У разі близьких у часі відмов зростає загроза живучості ІО і її забезпечення потребує більших витрат (більшої кількості реплік тощо). Отже, отриманий за допомогою моделювання степеневого розподілу відмов максимум в області значної кількості малих часових вікон з близькими відмовами має враховуватися під час вибирання технічних рішень для довготермінового зберігання. Виявлена для експоненціального розподілу більшість множинних відмов в області великих значень часу спостереження свідчить

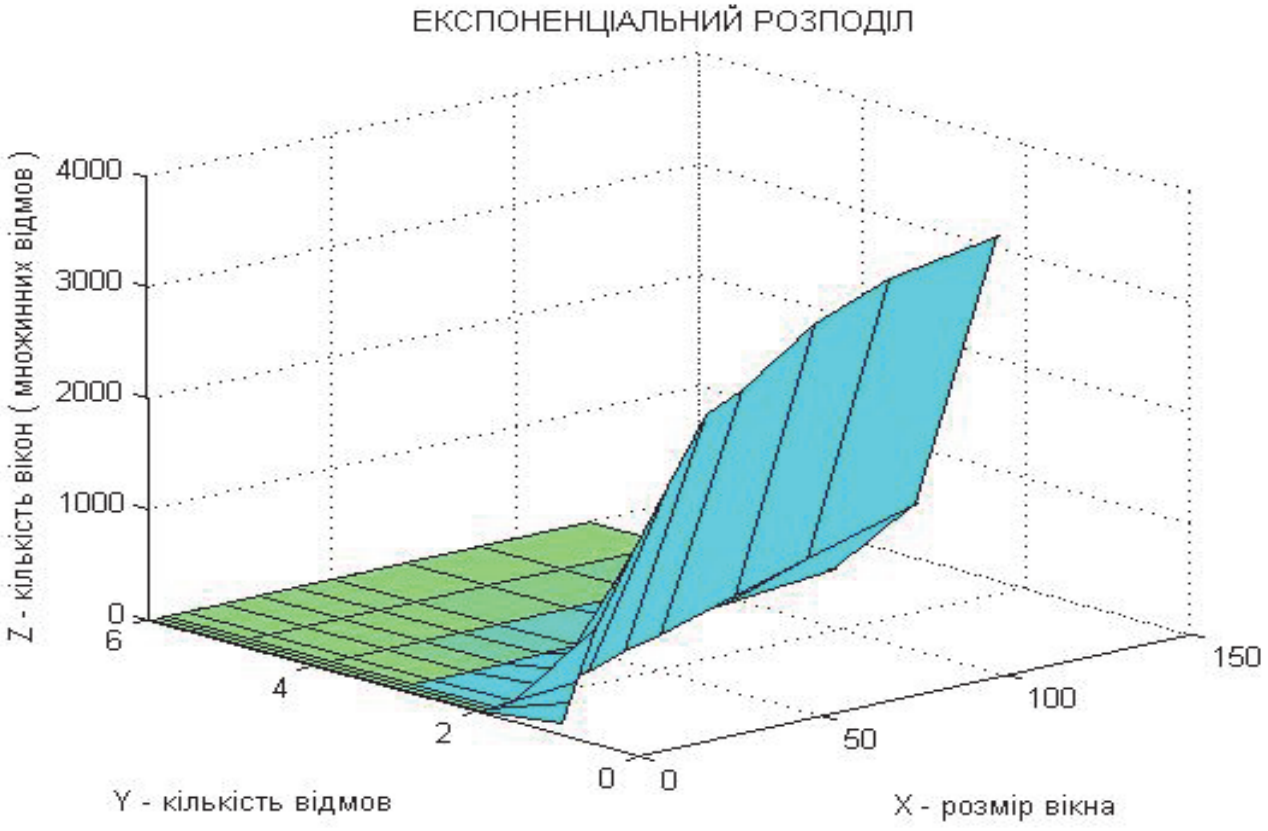

Рис. 3. Оцінка кількості множинних відмов за експоненціального розподілу

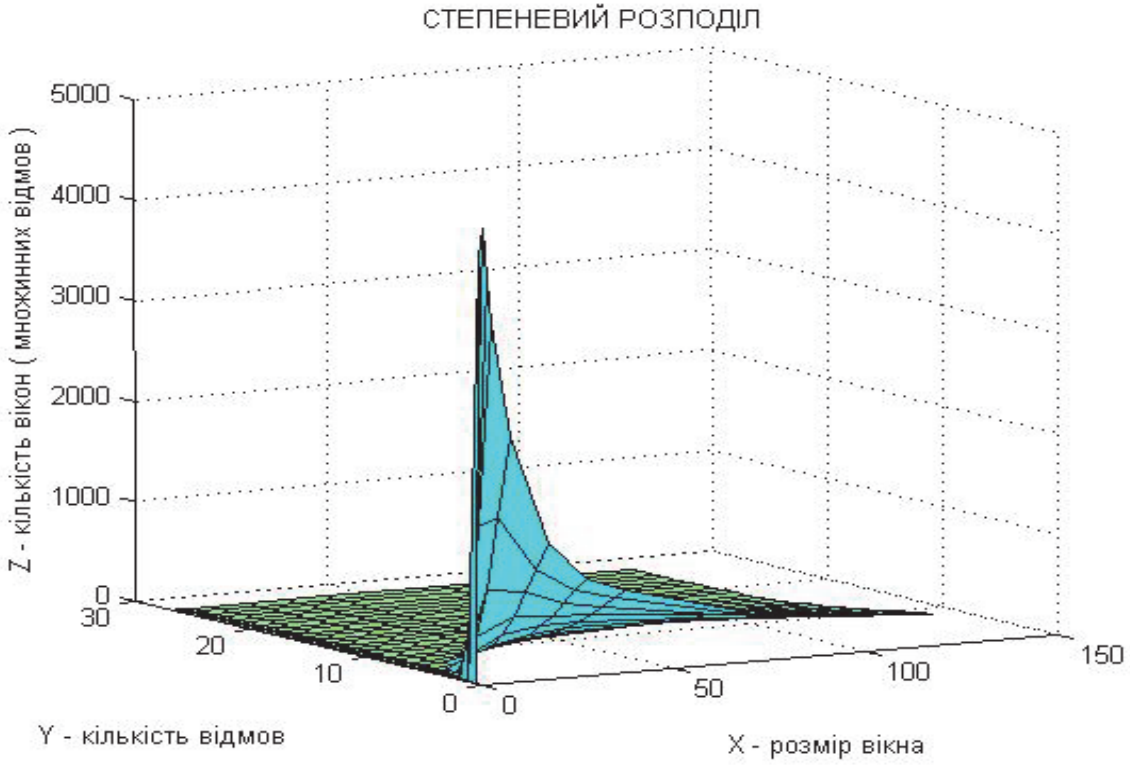

Рис. 4. Оцінка кількості множинних відмов за степеневого розподілу

про менш екстремальний характер потоку відмов при цьому розподілі.

Отримані результати моделювання підтверджуються дослідженнями розподілу латентних помилок секторів накопичувачів у часі [14].

Кожен привід спостерігався протягом року після його першої помилки й відслідковувалося, скільки двотижневих інтервалів за цей період вміщують помилки. Зроблено висновок, що помилки є щільно сконцентрованими в кількох коротких часових інтервалах і їхні сплески моделюються розподілом Парето.

Висновки

У результаті дослідження моделей природного старіння оптичних DVD-дисків на основі вимірювань показників помилок запропоновано використання степеневого розподілу помилок і відповідної аналітичної моделі для оцінки живучості ІО. Степенева модель підтверджена даними, які отримані для колекції CD-дисків у Бібліотеці Конгресу США і рядом інших досліджень.

Ураховуючи, що доступність даних і живучість ІО залежать від близькості моментів відмов, запропоновано модель множинних відмов. Розроблена імітаційна модель дає змогу виявити особливості множинних відмов за експоненціального і степеневого розподілів. Отриманий за допомогою моделювання степеневого розподілу відмов максимум в області значної кількості малих часових вікон з близькими за часом відмовами повинен враховуватися під час вибирання технічних рішень для довготермінового зберігання.

Отримані результати з оцінки живучості ІО для оптичних дисків і для розподілених мереж можуть використовуватися під час розробки методик і інструментальних засобів для управління інфраструктурою довготермінового зберігання великих обсягів даних.

ЛІТЕРАТУРА

1. *ДодоновА.Г., ЛандэД.В.* Живучесть информационных систем. – К.: Наук. думка, 2011. – 256 с.

2. *Li W., Yang Y., Yuan D.* A Novel Cost-effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centres // Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011. – P. 496–502.

3. *Weihs C., Rauber A.* Simulating the effect of preservation actions on repository evolution. In Proc. of iPRES 2011. – P. 62–69.

4. *Luan F., Nygård M., Mestl T.* A Mathematical Framework for Modeling and Analyzing Migration Time // Proceedings of the 10th annual joint conference on Digital libraries, JCDL'10, 2010. – P. 323–332.

5. *Березін Б.О., Ланде Д.В.* Дослідження стану оптичних носіїв при довготерміновому зберіганні цифрової інформації // Електронний документ: актуальні завдання та практичне впровадження (Життєвий цикл електронного документа): матеріали наук.-практ. конф., 11–12 жовтня 2012 р., Київ / – К., 2012. – С. 24–27.

6. *Ланде Д.В., Березін Б.О.* Живучість інформаційних об'єктів при довготерміновому зберіганні великих об'ємів даних / Д.В. Ланде, Б.О. Березін // Інформація та безпека. – 2012. – № 3–4 (11–12). – С. 13–15.

7. *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.

8. *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.

9. *Bhagwan R., Savage S., Voelker G.* Understanding availability // In IPTPS, Int'l Work. on Peer-to-Peer Systems, 2003. – P. 256–267.

10. *Amjad T., Sher M., Daud A.* A survey of dynami c replication strategies for improving data availability in data grids // Future Generation Computer Systems, 28. – Elsevier, 2012. – P. 337–349.

11. *Kermarrec A., Le Merrer E., Straub G., Kempen V.* Availability-Based Methods for Distributed Storage Systems // 2011, INRIA Technical Report v2. – P. 1–12.

12. *Ford D., Labelle F., Popovici F., Stokely M., Truong V., Barroso L., Grimes C., Quinlan S.* Availability in globally distributed storage systems // In Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation, 2010. – P. 1–7.

13. *Heien E., Kondo D., Gainaru A., LaPine D., Kramer B., Cappelo F.* Modeling and Tolerating Heterogeneous Failures on Large Parallel Systems // IEEE/ACM Supercomputing Conference (SC), 2011. – P. 1–11.

14. *Schroeder B., Damouras S., Gill P.* Understanding latent sector errors and how to protect against them. In Proc. of the 8th USENIX FAST, 2010. – P. 1–14.